



**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**  
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ  
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ  
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ  
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

**ΗΧΟΠΟΙΗΣΗ ΕΙΚΟΝΑΣ**  
**(Σχημα, Χρώμα & Υφή)**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**ΚΑΖΑΖΗΣ Ν. ΣΑΒΒΑΣ**

**Επιβλέπων :** ΚΑΜΠΟΥΡΑΚΗΣ ΓΕΩΡΓΙΟΣ

Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, ΑΥΓΟΥΣΤΟΣ 2012





**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**  
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ  
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ  
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ  
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

**ΗΧΟΠΟΙΗΣΗ ΕΙΚΟΝΑΣ**  
**(Σχημα, Χρώμα & Υφή)**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**ΚΑΖΑΖΗΣ Ν. ΣΑΒΒΑΣ**

**Επιβλέπων :** ΚΑΜΠΟΥΡΑΚΗΣ ΓΕΩΡΓΙΟΣ

Επ. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 29 Οκτωβρίου 2012.

.....  
Γεώργιος Καμπουράκης  
Επ. Καθηγητής Ε. Μ. Π

.....  
Βασίλειος Λούμος  
Καθηγητής Ε. Μ. Π

.....  
Ελευθέριος Καγιάφας  
Καθηγητής Ε. Μ. Π

Αθήνα, ΑΥΓΟΥΣΤΟΣ 2012

.....  
Σάββας Ν. Καζάζης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

**Copyright © Σάββας Ν. Καζάζης, 2012**

Με επιφύλαξη παντός δικαιώματος. **All rights reserved.**

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

# Image Sonification (Shape, Color & Texture)



Savvas N. Kazazis

School of Electrical and Computer Engineering

National Technical University of Athens

*Diploma Thesis*

2012 August

---

Supervisor: George Cambourakis

1. Reviewer: Vassilis Loumos

2. Reviewer: Eleftherios Kayafas

Day of the defense:

Signature from head of division:

## **Abstract**

This diploma thesis deals with image sonification and addresses the most common problems found in the field which are dealing with the sonification of shape, color and texture. While most approaches are based on user interaction, we have achieved to establish a one-to-one relationship between the image and the sonified result. By using perceptually meaningful mappings, the properties of an image are directly reflected to the audio domain in a very predictable way. If the task is to convey information, the listener can draw conclusions about the image by decoding the sonified result. Otherwise, using image sonification as a tool to aid sound design, can yield many interesting audio results that are hard to achieve by using only the existing audio based techniques.

Keywords: Image sonification, auditory display, audio-visual, perception, sound design, visually impaired, data mapping, additive synthesis, color, shape, texture.

---



To my parents,  
*Nikos and Maria*  
and... *Nennen.*

*“We can forgive a man for making a useful thing as long as he does not admire it. The only excuse for making a useless thing is that one admires it intensely. All art is quite useless.”*

Oscar Wilde: The picture of Dorian Gray.

# Contents

<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 An Overview of Sonification</b>	<b>3</b>
2.1 Definitions . . . . .	3
2.2 Usage and Application Fields . . . . .	5
2.3 Description and Classification of ‘Classic’ Sonification Techniques	6
2.4 The Sonification Operator . . . . .	10
<b>3 Auditory Perception</b>	<b>13</b>
3.1 The Human Ear . . . . .	13
3.1.1 The outer ear . . . . .	14
3.1.2 The middle ear . . . . .	15
3.1.3 The inner ear . . . . .	15
3.2 Loudness . . . . .	17
3.3 Pitch . . . . .	19
3.4 Timbre . . . . .	22
3.5 Auditory Scene Analysis . . . . .	23
<b>4 Color</b>	<b>27</b>
4.1 The Human Eye and the Sensation of Color . . . . .	28
4.1.1 The Tristimulus theory . . . . .	29
4.1.2 The Opponent Process theory . . . . .	30

## CONTENTS

---

4.2	Perceptual Attributes of Color . . . . .	32
4.3	Color in context . . . . .	33
4.4	Color Spaces . . . . .	35
4.4.1	Trichromacy and Color matching . . . . .	35
4.4.2	The RGB Color Space . . . . .	36
4.4.3	The CMY(K) Color Space . . . . .	36
4.4.4	The YCbCr Color Space . . . . .	37
4.4.5	The HSV, HSL and HSI Color Spaces . . . . .	37
4.5	Gestalt Principles . . . . .	39
<b>5</b>	<b>Sound Synthesis</b>	<b>43</b>
5.1	Additive Synthesis . . . . .	43
5.2	Subtractive Synthesis . . . . .	44
5.3	Modulation Synthesis . . . . .	45
5.4	Physical Modeling . . . . .	48
5.5	Granular Synthesis . . . . .	49
<b>6</b>	<b>Parameter Mapping and Image Sonification</b>	<b>51</b>
6.1	A formalization of Parameter Mapping Sonification . . . . .	52
6.2	Mapping Topologies . . . . .	53
6.3	Drawbacks or Advantages? . . . . .	55
<b>7</b>	<b>Sonification of Shape</b>	<b>57</b>
7.1	Related work . . . . .	57
7.2	Contour Tracing . . . . .	59
7.3	Mapping Spatial Datasets to sound . . . . .	62
7.3.1	Vertical axis to amplitude or pitch, horizontal direction to stereo position . . . . .	65
7.3.2	Defining a new curve for amplitude or frequency, horizontal direction to stereo position . . . . .	65
7.4	Conclusions . . . . .	67

<b>8 Creative Aspects of Image Sonification</b>	<b>69</b>
8.1 Related Work . . . . .	70
8.2 Tracing and Reshaping the Shape of an Object . . . . .	71
8.2.1 Shape as an Amplitude Envelope . . . . .	72
8.2.2 Shape as a Wavetable . . . . .	72
8.3 Gray Level Images . . . . .	72
8.4 Colored Images . . . . .	74
8.4.1 Color Sonification . . . . .	76
8.4.2 Texture Sonification . . . . .	79
8.5 Conclusions . . . . .	81
<b>9 Appendix: Sound Results</b>	<b>83</b>
<b>References</b>	<b>91</b>

## CONTENTS

---

# List of Figures

2.1	The Data Sonification Design Space Map. . . . .	7
2.2	The Analogic - Symbolic continuum. . . . .	10
3.1	Schematic drawing of the outer, middle and inner ear. . . . .	14
3.2	The inner ear with the basilar membrane in the cochlea highlighted in red. . . . .	16
3.3	Schematic drawing of the transformation of frequency into place along the basilar membrane. In (a) three simultaneously presented tones of different frequencies expressed as compound time function produce travelling waves (b), that reach their maximum at three different places corresponding to the characteristic frequencies. . .	17
3.4	At 20 phons the reference level at 1 kHz is by definition 20 dB SPL but at 100 Hz the sound level has to be nearly 40 dB to be perceived as equally loud. . . . .	19
3.5	The helical model of pitch. Musical pitch is depicted as varying along both a linear dimension of height and also a circular dimension of pitch class. The helix completes one full turn per octave, so that tones that stand in octave relation are in close spatial proximity, as shown by D#, D# ', and D# ". . . . .	20
3.6	At 20 phons the reference level at 1 kHz is by definition 20 dB SPL but at 100 Hz the sound level has to be nearly 40 dB to be perceived as equally loud. . . . .	21
3.7	Visual stream segregation. . . . .	24
4.1	Visible spectrum . . . . .	27

## LIST OF FIGURES

---

4.2	The human eye . . . . .	29
4.3	Log of the relative spectral sensitivities of the three kinds of colour receptor in the human eye. . . . .	30
4.4	Types of radially symmetrical cells found in the primate visual system . . . . .	31
4.5	The commonly used luminous efficiency functions of vision defined by the Commission Internationale de LEclairage (CIE). $V(\lambda)$ and $V'(\lambda)$ is the photopic and scotopic luminous functions respectively. . . . .	33
4.6	Simultaneous Contrast: The internal squares all have the same luminance but the changes in luminance in the surrounding areas change the perceived luminance of the internal squares. . . . .	34
4.7	Illustration of context effects. The x-shaped intersections on the two sides of the figure appear quite different. The light reaching the eye from these two regions is the same, however. This can be seen by tracing from one x-shaped region to the other. . . . .	34
4.8	Left: An example of additive mixing based on the RGB colorspace. Right: An example of subtractive mixing based on the CMY colorspace. . . . .	36
4.9	The HSV hexcone (c) after the rotation of the RGB color cube (a, b). Image taken from [36]. . . . .	39
4.10	(a) The Kanisza triangle showing the gestalt principles of good continuation and closure. (b) Similarity. (c) Proximity. (d) Relative proximity. . . . .	40
4.11	Common Fate. . . . .	41
5.1	Classic synthesis techniques classified according to their principles of realization. . . . .	43
5.2	Source-filter synthesis. The transfer function of the time-varying filter $H(z)$ is described by filter coefficients $a(n)$ and $b(n)$ . . . . .	45
5.3	Spectrum of simple amplitude modulation. . . . .	46
5.4	Waveshaping using a quadratic transfer function $f(x) = x^2$ : (a) the input; (b) the transfer function; (c) the result, sounding at twice the original frequency. . . . .	47



## LIST OF FIGURES

---

5.5	The spectrum of frequency modulation. . . . .	47
5.6	A Karplus-Strong model for plucked string tones. . . . .	49
5.7	A granule of 50ms enveloped by a Gaussian window. . . . .	50
6.1	Typical transfer functions for parameter mapping. The black line shows a piecewise linear transfer function. The blue and green dashed lines are respectively sigmoid and exponential transfer functions. . . . .	53
7.1	A square and its projections at 0 and 45 degrees based on the Radon Transform. . . . .	60
7.2	The Moore's Contour Tracing Algorithm. . . . .	61
7.3	A curve (left) and the output of Moore's Contour Tracing Algorithm (right). . . . .	62
7.4	An example curve with labeled breakpoints. The sonification starts from point D, since it is the first point traced by the algorithm. . . . .	64
7.5	The instantaneous frequency envelope, as described in paragraph 7.3.2 . . . . .	66
7.6	A stereo wave file using amplitude and panning as mapping parameters, as described in paragraph 7.3.1 . . . . .	68
8.1	Gray level values mapped to audio sample levels. . . . .	73
8.2	A Hilbert curve of order 4 filling a [16, 16] image. The image could be roughly segmented in 4 regions, each one describing the order of the curve. . . . .	75
8.3	10 saturation levels ranging from 0.1 to 1, mapped to amplitude dBFS units. . . . .	77
8.4	10 value levels ranging from 0.1 to 1, mapped to octave ranges. . . . .	78
8.5	A major scale. All squares have a Value of 0.5 which sets the base frequency at 320 Hz except the last one which has a value of 0.6 setting the base frequency at 640 Hz. The Hue values in degrees are [0 60 120 150 210 270 330 0] which correspond in frequencies of [320.0000 359.2688 403.3564 427.3894 479.8364 538.7195 604.8284 640.0000] Hz . . . . .	79
9.1	Frequency envelope of sound (5). . . . .	84

## LIST OF FIGURES

---

9.2	Amplitude envelope of sound (7).	85
9.3	First few samples of sound (9).	85
9.4	Sample images.	86
9.5	Sample images.	86
9.6	Sample images.	88
9.7	Sample images.	88
9.8	Sample Image.	89

# List of Tables

6.1	One to One Mapping. Table taken from [23]	54
6.2	One to Many Mapping. Table taken from [25]	54
6.3	Many to One Mapping.	54
9.1	Used Mappings	90

## LIST OF TABLES

---

# Chapter 1

## Introduction

This diploma thesis deals with image sonification. More specifically, the main goal is to find ways of translating the image data which describe shape color and texture, to sound. In order the reader to understand our approach to the problem, must be familiar with some concepts that will be used in our final implementations. In Chapter 2 we present an overview of Sonification regarding the definitions, its usage and techniques. Chapter 3 deals with the physiology of the human ear and some topics which are found in the field of psychoacoustics. Since sonification communicates data through sound, a basic knowledge of our auditory perception is of vital importance. Chapter 4 starts with the physiology of the human eye and continues with a presentation of some widely used color spaces. An understanding of how the data are organized in an image, is as useful as the understanding of how the data are organized according to various sound synthesis techniques, which are presented in Chapter 5. In Chapter 6 we emphasize on the Parameter Mapping Sonification technique, since it is the one which we believe that gives the most promising results. Finally, in Chapter 7 and Chapter 8 we present new techniques for sonifying the shape, color and texture of images. The results can be applicable in various domains, ranging from the aid of visually impaired to sound design.

## 1. INTRODUCTION

---

# Chapter 2

## An Overview of Sonification

Sonification is the translation of data into sound. Though it is a relatively new discipline, it has been used for over a century by researchers without them being aware that they were developing a new scientific methodology. The tuning of a musical instrument for example can be considered as a sonification when adjusting the tension of a string or the stiffness of a membrane. One other early example, is the Geiger counter which was invented in 1908 and informs the listener about the levels of radioactivity by emitting sound. Some other early sonification examples can be found at [30]. With the establishment of the International Conference on Auditory Display (ICAD) in 1992, scientists from various disciplines inspire each other towards an effective use and design of Auditory Displays, with sonification being one major subfield of research. Related research disciplines include: Data mining/Statistics, Human Computer Interaction (HCI), Digital Signal Processing, Physiology/Biology, Auditory Perception and Cognition, Psychology, Psychoacoustics, Music Cognition and Musicology.

### 2.1 Definitions

One of the first attempts towards a working definition of sonification is made by Scaletti [30]:

*A mapping of numerically represented relations in some domain under study to relations in an acoustic domain for the purposes of interpreting, understanding,*

## 2. AN OVERVIEW OF SONIFICATION

---

*or communicating relations in the domain under study.*

S. Barrass revised Scaletti's definition [4]:

*A mapping of information to perceptual relations in the acoustic domain to meet the information requirements of an information processing activity.*

Finally, the widely accepted definition of sonification has been given by G. Kramer et al. [31]:

*Sonification is the use of nonspeech audio to convey information. More specifically, sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation.*

However, this definition is criticized as being too general and imprecise, but also because it excludes the use of speech like sounds which can be used in sonifications as well, since speech can be very useful for certain tasks [23]. T. Hermann suggested a more strict definition of sonification [24]:

*A technique that uses data as input, and generates sound signals (eventually in response to optional additional excitation or triggering) may be called sonification, if and only if:*

*Condition 1: The sound reflects objective properties or relations in the input data.*

*Condition 2: The transformation is systematic. This means that there is a precise definition provided of how the data (and optional interactions) cause the sound to change.*

*Condition 3: The sonification is reproducible: given the same data and identical interactions (or triggers) the resulting sound has to be structurally identical.*

*Condition 4: The system can intentionally be used with different data, and also be used in repetition with the same data.*



This definition has also received criticism [58], especially the 4th condition, since sonifications aim towards a specific user task and the mappings are defined in a way that reflect properties among specific datasets and data types, making it hard to be well suited for different data and tasks. The sonification of a music score can only be achieved if the notes (the data) belong to the instrument's pitch range and the relationships between the data, do not violate the physics of the instrument. For example, a glissandi can be achieved with stringed instruments but it is not possible for all the percussive ones.

Sonification is a subtype of Auditory Display, which could be broadly defined as any display that uses sound to communicate information [25]. S. Barrass, again based on the definitions of Scaletti, gave the following definition for Auditory Information Design [4]:

*Auditory Information Design is the design of sounds to support an information processing activity, focusing on the specific task like interpreting, understanding or communicating relations in the data.*

## 2.2 Usage and Application Fields

The human auditory system is very sensitive to rhythm, amplitude and pitch changes, therefore complex patterns and temporal changes in the data can be easily and rapidly detected when projected in an Auditory Display. Furthermore the ear has a much more broader bandwidth than our visual system, therefore patterns that that are hard to be perceived visually or user tasks that would be hard to achieve or would be time consuming when using vision, may require less effort when an Auditory Display is used. There are also some situations where the eyes are occupied, monitoring a different process than the ears. Consider as an example driving a car and changing the gears. The pitch and intensity of the engine, are acting as indicators of when the driver should change a gear. Another major difference between our visual system and our auditory system, is that the second is able to monitor (listen to) many streams at the same time, something which is impossible for our eyes. *Background listening* [30], is very useful in monitoring tasks since one can pay attention when only a specific sound

## 2. AN OVERVIEW OF SONIFICATION

---

occurs, acting for example as an alert function. This sound “awareness” sources from our ability to easily remember, memorize and recognize melodic, rhythmic and timbral qualities of various sonic structures. Auditory Displays can be used for:

- Alarms, notifications, alerts and warnings.
- Status, process, and monitoring messages.
- Scientific data exploration.
- As an aid tool for visually impaired people.
- Educational/didactic purposes.
- Multimodal displays.
- Artistic purposes.

For a thorough description of the above mentioned fields, refer to [7].

### 2.3 Description and Classification of ‘Classic’ Sonification Techniques

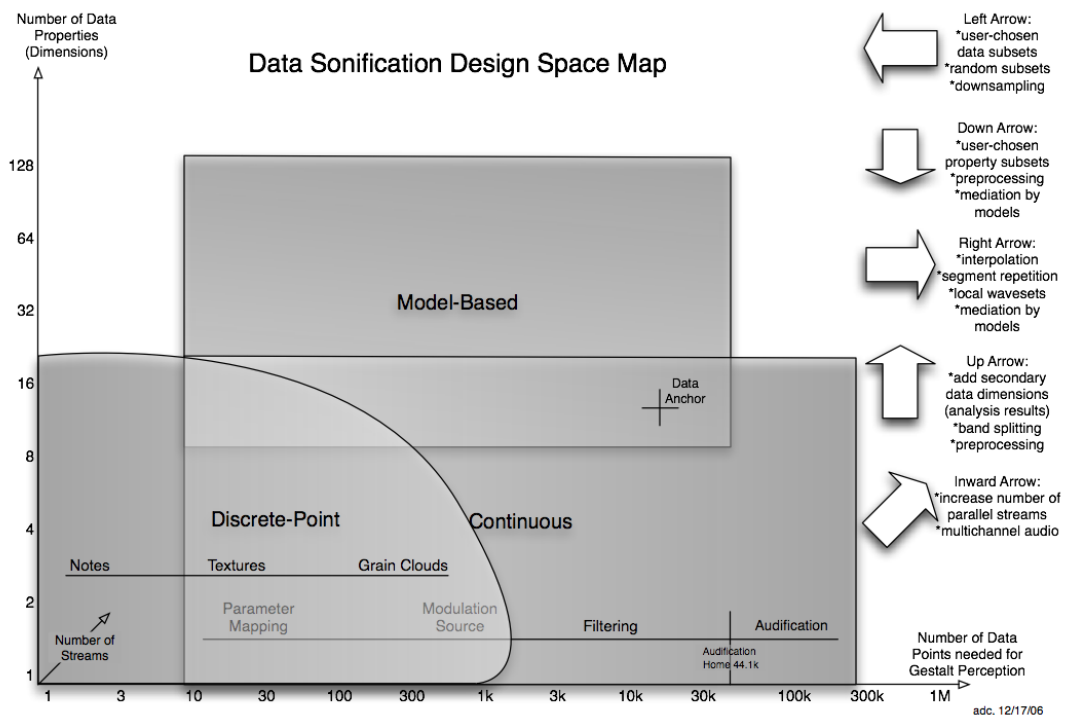
G. Kramer et al. [31] presented four main design issues that should be taken into consideration:

1. Is there a psychologically-based or application-supported natural taxonomy of sonification techniques?
2. What types of data or tasks lend themselves naturally to effective sonification?
3. Which acoustic cues and data mappings are intuitive and facilitate the presentation of complex, multidimensional displays?
4. What factors limit how well information can be extracted from a sonification?

## 2.3 Description and Classification of ‘Classic’ Sonification Techniques

De Campo [16] classifies sonification approaches in three broad categories:

1. Discrete Point Data Representation
2. Model Based Data Representation
3. Continuous Data Representation



**Figure 2.1:** The Data Sonification Design Space Map. Image taken from [16].

Sonification techniques that belong to the *Discrete Point Data Representation* category include:

### Auditory Icons

Auditory Icons are communicative sounds, usually modeled after real world (physical) sounds, which provide feedback in a user interface. Their meaning can intuitively be connected with a specific action such as the sound which occurs when someone deletes a file from a computer, or empties the trash bin.

## 2. AN OVERVIEW OF SONIFICATION

---

### **Earcons**

Earcons are usually brief melodies or abstract sounds and since they are symbolic representations of an action, have to be learned by the user before a meaning can be attached to a sound. As an example, consider the words 'Save' and 'File' each of them being represented by a different earcon. If the earcons are played back in a sequence, they represent the action 'Save File'.

### **Spearcons**

Spearcons are time compressed spoken phrases often to a point where they become incomprehensible and are used to facilitate menu navigation, usually through text to speech software.

Sonification techniques that belong to the *Continuous Data representation* category include:

### **Parameter Mapping**

Parameter Mapping is the most widely used technique for data exploration, employing most of the times a passive mode of interaction. Data parameters are mapped to sound parameters such as frequency or pitch, duration, amplitude or loudness, spatial cues, brightness of the sound, rhythm, etc... Changes in a data dimension cause changes in the acoustic dimension, but most sound parameters are not perceptual independent as for example pitch with intensity, therefore there is a limit in data dimensionality that can be effectively projected in an Auditory Display.

### **Audification**

Audification is the direct translation of data into sound and is particularly useful when dealing with very large data sets. Usually the waveforms that result from the data have to be frequency shifted and time compressed or expanded, to the audio range.

**Model-Based Sonification** (MBS) as the name suggests, belongs to the *Model*

## 2.3 Description and Classification of ‘Classic’ Sonification Techniques

*Based Data Representation.* Model-Based Sonification was introduced by Hermann [23] as an alternative to the Parameter Mapping Sonification. It is an inherently interactive technique, in which the user explores the data relations by exciting a system and listening back its acoustic response. The system is a virtual acoustic object, whose structure is not only dependent on the data but also on their interaction, often defined by theoretical acoustics or virtual physics. Hermann describes MBS as, “Thus the data more or less directly becomes the sounding instrument, which is examined, excited or played by the listener.” One major advantage over the Parameter Mapping technique is that it allows for much more higher data dimensionality.

Besides de Campo’s classification scheme, there are also available some other approaches for classifying sonifications. These include:

### **The Semiotic Categorization**

Semiotic theory involves the study of signs and their meaning. Since sonification aims to highlight some properties of the data by making use of sound, there is space for a semiotic perspective in Auditory Display design, by treating sound signals as signs. Semiotics can be divided in three categories:

- Semantics, which describe the relationship between a sign and the signified. Auditory icons, belong to this category, since meaning is attached to sound through an iconic association.
- Syntactics, which describe the relations between signs in a formal structure, such as the structuring elements that make up a language. Earcons belong to this category.
- Pragmatics, which study the way in which meaning arises through a certain context. In an Auditory Display the qualities of a sound can be described by a lexical approach, as in Parameter Mapping Sonification in which the signs (sounds) are created from the data [4].

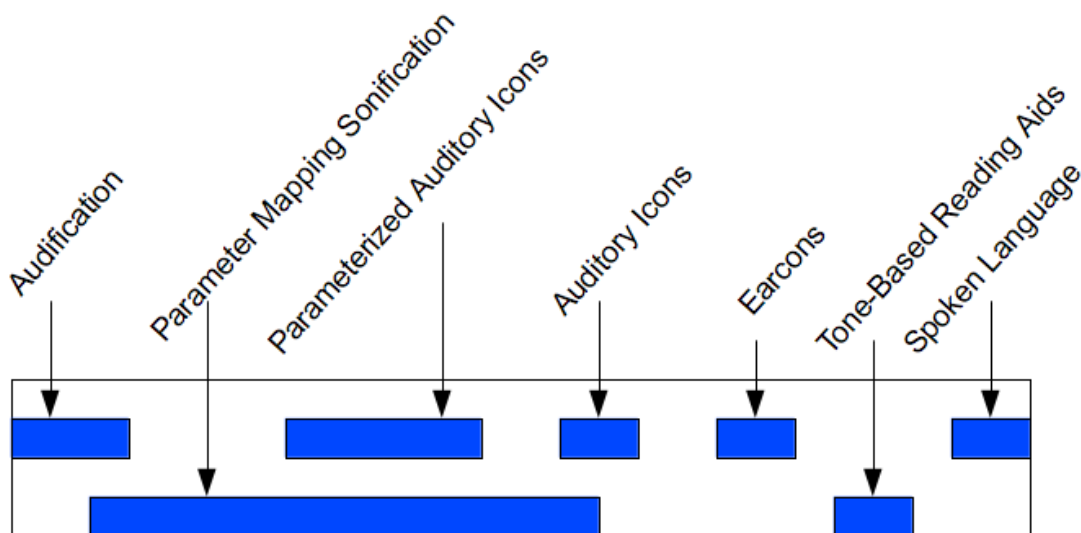
### **The Analogic Symbolic continuum**

A symbolic representation of information, uses categorical signs with a high level

## 2. AN OVERVIEW OF SONIFICATION

---

of abstraction and the relationships between the representations, do not reflect intrinsic relationships between the elements being represented. On the other hand, in analogic representations there is an immediate and intrinsic correspondence between the the signifier and the signified. Hermann [23] gives a nice example for representing temperature. The symbolic representation would make use of words such as ‘cold - warm - hot’, but other words can also be used once an agreement is made on what these represent, for example ‘blue - orange - red’. The analogic representation would make use of a thermometer, in which the height of the quicksilver analogically represents temperature.



**Figure 2.2:** The Analogic - Symbolic continuum. After Hermann [23]

### 2.4 The Sonification Operator

Sonifications translate the data from a domain science to sound through a sound generation technique aiming to a perceptually meaningful result, but the formulation in the domain science is usually different from the formulation describing the sound generation technique. To overcome this ambiguity in formulation with respect to the sonification methods, J. Rohrhuber [49] recently introduced the

## 2.4 The Sonification Operator

---

sonification operator  $\mathring{S}$ :

$$\mathring{S} : A(d) \rightarrow \mathring{y}(\mathring{t}, d, p) \quad (2.1)$$

The variables involved in the sonification process are denoted with a ring.  $A$  is a function, relation or in general the domain science,  $d$  are the domain variables,  $p$  are the sound parameters set in the sonification.  $\mathring{t}$ , stands for the sonification time and is used in order to differentiate from a change in time that could occur in the data domain. This formalization describes a transformation from the data domain into the signal domain but it has been criticized because it does not take into account any perceptual qualities. A nice example is given in [58] which describes sampling time versus sonification time.

$$t_s = \frac{n}{f_s} \quad (2.2)$$

$f_s$  is the sampling rate and  $n$  is the number of samples. The sonification time  $\mathring{t}$ , according to this formulation corresponds to the number of samples  $n$ :

$$n \triangleq \mathring{t} \quad (2.3)$$

Although sonification (listening) time is perceived as continuous, eventually involving a digital-to-analog conversion, the sonification algorithm is executed in discrete time.

## 2. AN OVERVIEW OF SONIFICATION

---



## Chapter 3

# Auditory Perception

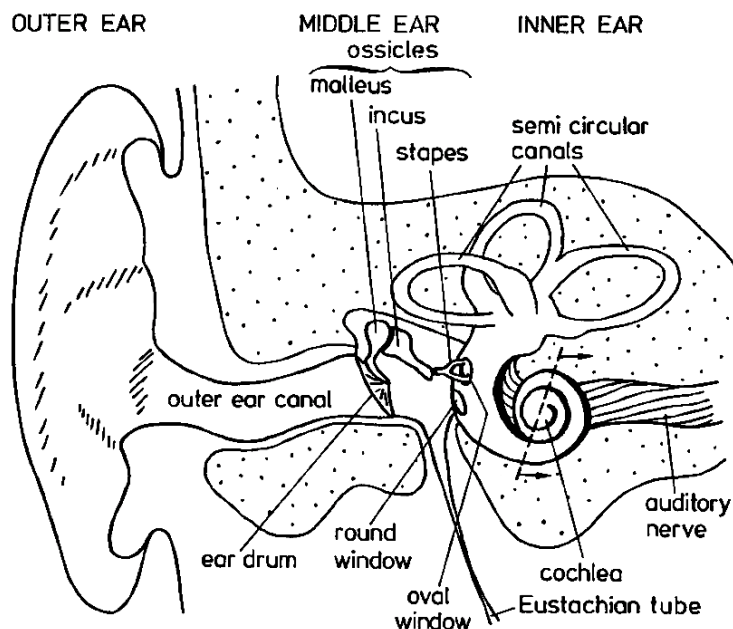
An understanding of how our auditory system works, is an essential precondition for designing effective Auditory Displays and meaningful sonifications. We start by presenting the physiology and function of the human ear and we continue with some aspects of psychoacoustics. Pitch and loudness are the perceptual correlates of frequency and intensity. Dealing with sound from a psychophysical point of view, can lead to functional mappings. Finally we present an introduction to Auditory Scene Analysis [12], which describes the strategies that our brain uses to fuse or to segregate sounds into auditory streams.

### 3.1 The Human Ear

The human ear consists of three main parts: The outer ear, the middle ear and the inner ear. The sound from the external environment is collected and filtered by the outer ear. Later on it propagates to the middle ear through the external auditory channel, causing vibrations to the eardrum. The middle ear converts the air pressure waves into liquid waves. The liquid waves will be transformed to nerve impulses in the inner ear and will be transmitted to the brain through the auditory nerve.

### 3. AUDITORY PERCEPTION

---



**Figure 3.1:** Schematic drawing of the outer, middle and inner ear. Image taken from [19].

#### 3.1.1 The outer ear

The outer ear consists of the pinna and the concha which collect and filter the sound before delivering it to the middle ear through the external auditory canal. It is of high importance for sound source localization, for example many mammals have the ability to move their pinna in order to focus their hearing. As the sound reaches our body, reflections take place between the outer ear, the head and the torso which have an effect to the overall sound pressure that reaches the eardrum. The differences between level, frequency content and timing (interaural time differences) of the acoustic cues that arise at each ear, are used by our auditory system to construct the auditory space.

The external auditory canal is a tube bounded at the end by the tympanic membrane. It transmits the sound pressure wave from the pinna to the middle ear. As the sound signal funnels from the large aperture of the pinna towards the smaller aperture of the auditory canal, it gets amplified in a region around 4 KHz which is why humans are more sensitive to this frequency region. This

spectral filtering is directly related to the direction of the incoming sounds and enables us to perceive sounds as being outside the head. Listening music with headphones which emit sound directly to the ear canal, is a different experience from listening with speakers, where the sound is filtered by the outer ear. The filtering is also an individualized process since the filter's characteristics depend on the precise shape of the outer ear.

### 3.1.2 The middle ear

The outer ear is filled with air while the inner ear is filled with fluid which surrounds the sensory cells. In order to excite the cells, a transfer function has to take place which converts the air pressure waves to fluid waves. This is achieved through the mechanism of the middle ear which is comprised of the ear drum, the three middle bones (malleus which is attached to the ear drum, incus and stapes) and the stapes footplate which together with a ring shaped membrane (oval window), induces fluid movement in the cochlea of the inner ear. Transmission of sound through the middle ear is most efficient at middle frequencies ranging from 500 Hz to 4 KHz.

As the impedance is much higher in fluids than air, the pressure must be amplified through an impedance matching mechanism. This is achieved by two ways; Firstly, the size of the eardrum is about 17 times larger than the size of the oval window and secondly the three middle bones act as "levers". Through this mechanism the vibrations are amplified by a factor of 20.

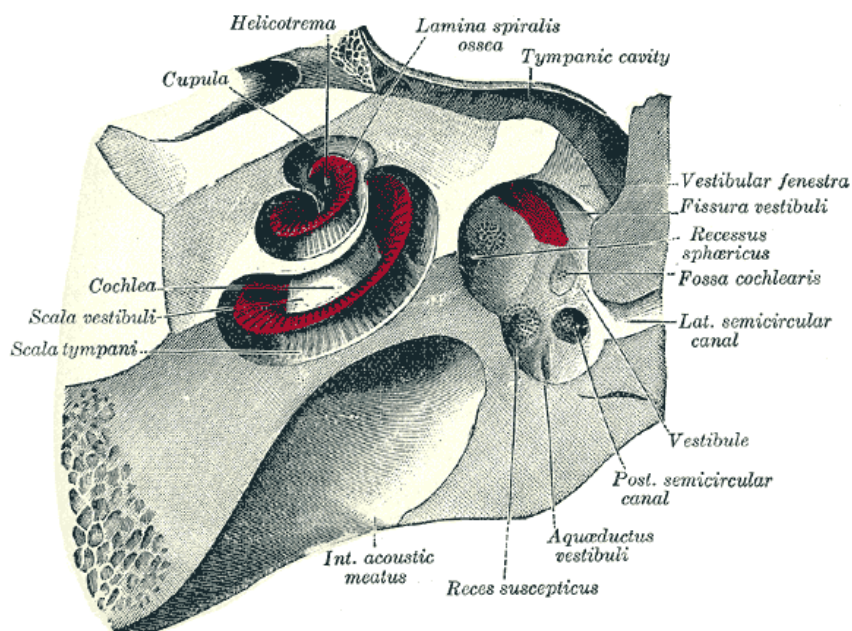
### 3.1.3 The inner ear

The cochlea has a coiled structure and converts the sound energy into electrochemical impulses which will be transmitted to the brain through the auditory nerve. It forms 2.5 turns allowing a basilar membrane length of 32mm (highlighted in red, Figure 3.2).

The basilar membrane is moved up and down by the pressure changes in the cochlea induced by the movement of the stapes footplate on the oval window. When the ear is exposed to a pure tone, a traveling wave propagates along the basilar membrane, whose envelope shows a maximum at a frequency dependent

### 3. AUDITORY PERCEPTION

---

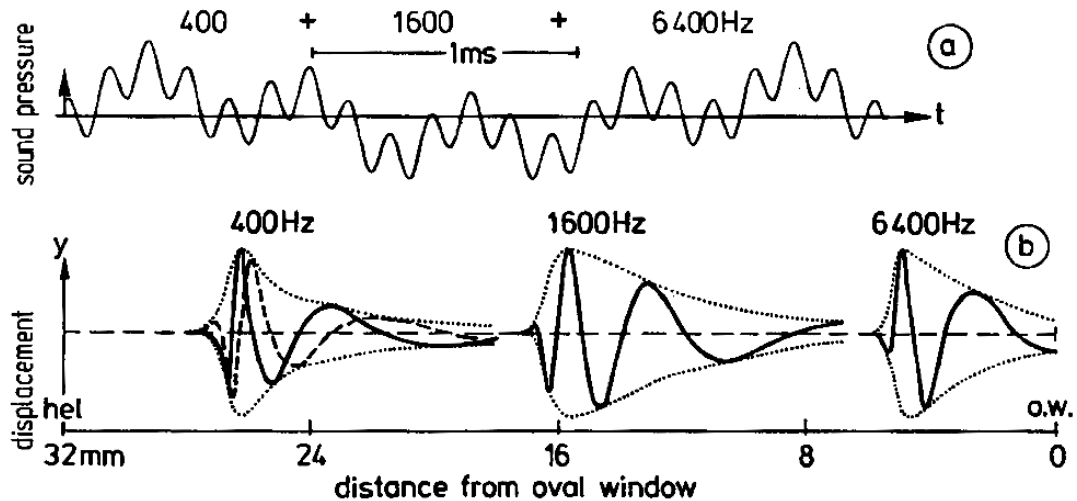


**Figure 3.2:** The inner ear with the basilar membrane in the cochlea highlighted in red. Adapted from <http://en.wikipedia.org/wiki/File:Gray923.png>

position. The stiffness and mass of the basilar membrane varies along its length, so a high frequency tone will cause a resonance closer to the basal end, which is located next to the oval window and the middle ear, while a low frequency tone will cause a resonance at the apical end. The separation by location on the basilar membrane is known as the place principle. As the maxima of the envelopes of the basilar membrane are closer together at high frequencies, frequency resolution is expected to decrease with increasing frequency.

The haircells, which are found within the organ of Corti and placed along the basilar membrane, will transmit to the auditory nerve the excitation locations of the basilar membrane. As a consequence, haircells respond selectively to specific frequencies. The frequency that gives maximum response at a particular point on the basilar membrane is known as the characteristic frequency. These cells have hairs sticking out of the top, called stereocilia which transduce the sound energy to electrical energy as their potential changes, depending on their deviation from the equilibrium position.

There are two kinds of haircells: (a) The inner haircells, which are arranged



**Figure 3.3:** Schematic drawing of the transformation of frequency into place along the basilar membrane. In (a) three simultaneously presented tones of different frequencies expressed as compound time function produce travelling waves (b), that reach their maximum at three different places corresponding to the characteristic frequencies. Image taken from [19].

in a row on the inner side of the organ corti. (b) The outer haircells are arranged in three rows, near the middle of the corti. Although there are about 12000 outer haircells and only 3500 inner haircells, more than 90% of the auditory nerves receive signals from the inner haircells. The outer haircells provide a form of feedback and operate as amplifiers. At high sound levels, they get saturated and the adequate stimulus operates almost exclusively on the inner haircells. On the other hand, at low levels the inner haircells are slightly stimulated in a direct way. Therefore an interaction between the inner and outer haircells, is assumed to be responsible for the large dynamic range and the sharp frequency selectivity at low levels.

## 3.2 Loudness

Loudness is the apparent or subjective intensity of a sound. A sound is described by its time varying pressure  $p(t)$  measured in PASCAL (Pa). Sound pressure

### 3. AUDITORY PERCEPTION

---

with intensity  $I$  are related by:

$$I = \frac{p(\tilde{t})^2}{Z_o} \quad (3.1)$$

where  $Z_o = 415 \text{ Pa s m}^{-1}$  is the impedance of air.  $p(\tilde{t})$  is the effective sound pressure. Our auditory system is sensitive to sound pressure levels ranging from  $10^{-5} \text{ Pa}$  (absolute threshold) to  $10^2 \text{ Pa}$  (threshold of pain). To cope with this broad range, loudness is often measured by the Sound Pressure Level (SPL). SPL is a logarithmic measure of the effective sound pressure of a sound, relative to a reference value. Sound pressure level  $L$  and sound intensity level are related by the equation:

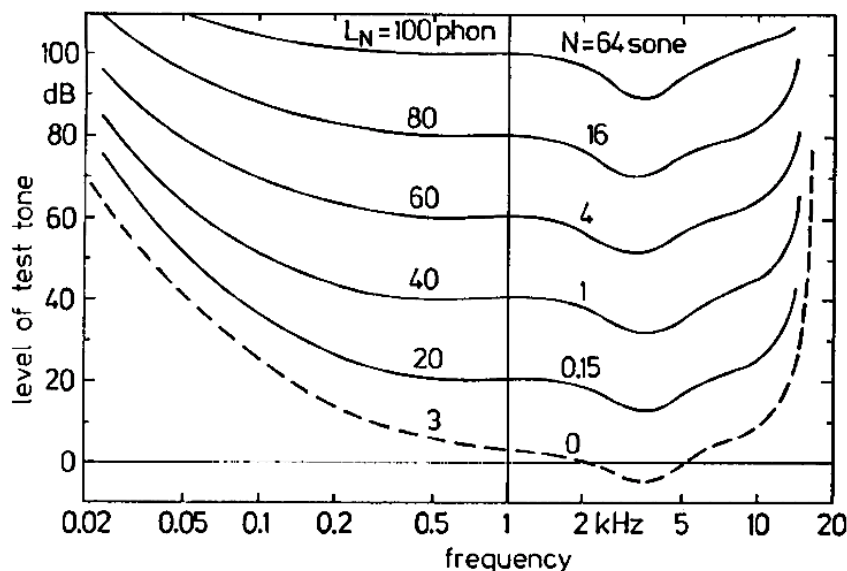
$$L = 20 \log_{10} \frac{p}{p_o} \text{ dB} = 10 \log_{10} \frac{I}{I_o} \text{ dB} \quad (3.2)$$

The reference value of the sound pressure  $p_o$  is standardized to  $20 \mu \text{ Pa}$  and corresponds to the lowest intensity sound that we are able to discriminate. The reference value  $I_o$  is defined as  $10^{-12} \text{ W/m}^2$ . A sound wave of frequency 1 kHz and intensity  $I_o$  just exceeds the threshold in quiet.

The perception of loudness strongly depends on frequency. This dependency is presented in Fig. which shows the *equal loudness contours*, also known as *Fletcher Munson curves* or *isophones*. Equal loudness contours were derived experimentally by asking the listeners to adjust the intensity of pure tones, so that they become equally loud with a 1 KHz reference tone. By examining these curves we see that our ear is more sensitive to the region of 2 KHz to 5 KHz.

The subjective loudness level when compared with a 1 kHz sine tone is measured in *phon*. In other words, a phon is the perceived loudness at any frequency that is judged to be equivalent to a reference sound pressure level at 1 kHz. By definition, 1 phon is equal to 1 dB SPL at a frequency of 1 KHz. The dashed line in Figure 3.4 shows the threshold in quiet which corresponds to 3 dB at 1 KHz and not 0 dB, therefore is indicated by 3 phon.

A qualitative comparison between the loudness of two tones is achieved via the unit *sone*, which is also derived experimentally. According to the definition, the loudness of a 1KHz tone at 40 dB SPL is equivalent to 1 sone. Therefore



**Figure 3.4:** At 20 phons the reference level at 1 kHz is by definition 20 dB SPL but at 100 Hz the sound level has to be nearly 40 dB to be perceived as equally loud. Image taken from [19].

a tone of 2 sones is perceived twice as loud, a tone of 4 sones is perceived four times as loud, etc... Our ear is capable of perceiving level changes as small as 1 dB throughout the whole dynamic range which is about 120 dB.

Loudness also depends on the duration of the sound. The sounds used to estimate the equal loudness contours had a duration around 500 ms. Loudness stays constant when the duration of the sound is greater than 100 ms and for shorter durations loudness decreases. On the other hand, if the ear is exposed for a long period of time to sounds ranging from moderate to high levels, there is a reduction of the perceived loudness. This is referred to as adaptation or fatigue and can cause permanent hearing damage if the sound levels are above 110 dB SPL.

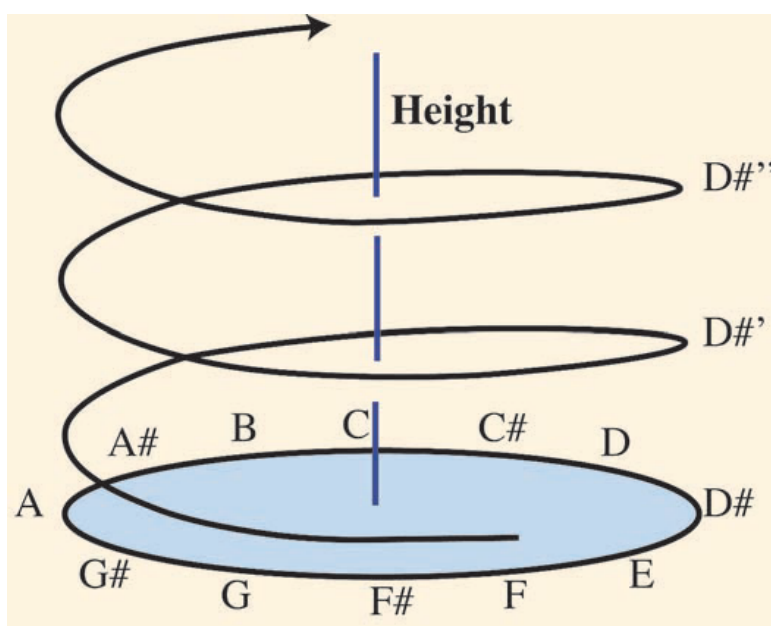
### 3.3 Pitch

Pitch is a positional perception of frequency. A musical scale represents changes in pitch and thus one can define pitch relations in terms of *pitch height* and *pitch*

### 3. AUDITORY PERCEPTION

---

*circularity*. Frequencies which are spaced at octaves apart, exhibit a perceptual similarity and have the same names. For example, when comparing C2 with C3, C3 is higher by an octave (frequency doubling) but both belong to the same pitch class (as all Cs) and “close” a circle which consists of the intermediate notes D2, D2#, E2, F2, F2#, G2, G2#, A2, A2# and B2.

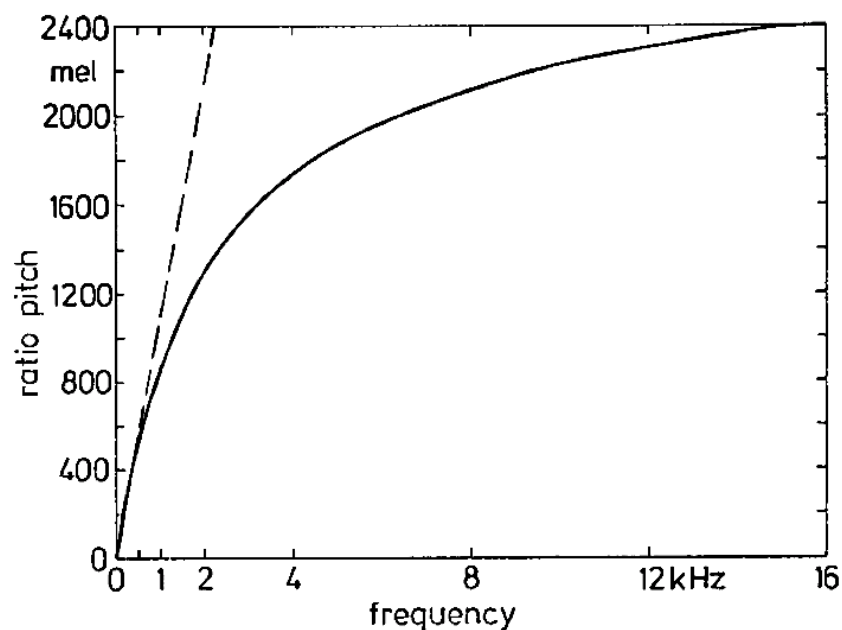


**Figure 3.5:** The helical model of pitch. Musical pitch is depicted as varying along both a linear dimension of height and also a circular dimension of pitch class. The helix completes one full turn per octave, so that tones that stand in octave relation are in close spatial proximity, as shown by D#, D# ', and D# ''. Image taken from [17].

Though we hear frequencies in the range of 20 Hz to 20 KHz, the sensation of pitch starts around 50 Hz and extends up to 5 KHz. Beyond that frequency, pitch discrimination is weakened. This is confirmed through psychoacoustic experiments which make use of *pitch ratios*. The subjects listen to a tone of a specific frequency and try to adjust the frequency of a second one, so that it has double or half the pitch of the first one. At low frequencies the matching is quite linear (especially for trained musicians) and a halving in pitch corresponds to a halving in frequency. However at higher frequencies, the linearity between pitch and frequency is lost. Halving the pitch of a tone at 8 KHz corresponds to a



frequency of about 1300 Hz not 4 KHz. The unit for measuring ratio pitch is *mel*, an abbreviation of the word melody. Mel is defined experimentally by choosing a reference frequency in a region where pitch ratios are proportional to frequency ratios and assuming a proportionality factor of 1.



**Figure 3.6:** At 20 phons the reference level at 1 kHz is by definition 20 dB SPL but at 100 Hz the sound level has to be nearly 40 dB to be perceived as equally loud. Image taken from [19].

The graph shown in Figure 3.6 was constructed by taking as a reference point the frequency of 125 Hz. A reference tone of 125 Hz is equal to 125 mel but a tone of 1300 Hz produces 1050 mel which is half of the mel produced by a 8 KHz tone. A formula which is widely used for converting frequency  $f$  to mel is the following:

$$mel = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (3.3)$$

Most of the sounds are not pure sinusoids but rather consist of many partials with each partial having its own frequency and energy. A harmonic complex

### 3. AUDITORY PERCEPTION

---

tone consists of partials whose frequencies are integer multiples of a common frequency, called the fundamental frequency. These partials are called overtones or harmonics of the fundamental frequency. A complex tone consisting of 1600, 1800 and 2000 Hz will be perceived as having a pitch of 200 Hz i.e the pitch of its fundamental frequency, even if this frequency is not physically present. This phenomenon is referred to as *residue pitch*, *periodicity pitch*, *virtual pitch* or *the missing fundamental*.

Just noticeable differences (JNDs) in frequency modulation (variation or vibrato in musical terms) are different from the JNDs in frequency steps. At low frequencies the JND in frequency modulation is almost constant, around 3.6 % and for frequencies over 500 Hz is approximately 0.7%. Our ear is far more sensitive in frequency changes rather than modulations. Surprisingly, our sensitivity is increased if there is a pause between the presented tones. At frequencies below 500 Hz the JND is 1 Hz or even smaller and above 500 Hz increases approximately to 0.2 %.

Sound pressure levels have rather a weak effect on pitch perception. In general, for high pressure levels, frequencies below 2 KHz seem to decrease with intensity while frequencies above 4 KHz seem to increase. On the other hand JNDs in frequency are level dependent only below 25 dB. Below this value the JND increases with decreasing level.

#### 3.4 Timbre

Timbre is defined by the American Standards Association (ANSI) in terms of what is not, rather of what it is:

*That attribute of sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.*

Timbre is multidimensional and is related to various sensations, such as roughness and brightness. In most cases it is used to describe the perceived quality of a sound. With respect to the ANSI definition, is worth noting that most research dissociates timbre from the dimensions of pitch, loudness and duration.

Timbre dimensions are usually specified through dissimilarity tests in which the subjects rate the dissimilarity of different sounds. A multidimensional scaling is then performed on the results to specify the dimensions and their importance. Sounds are placed along these dimensions and are investigated to find suited timbre attributes. For example one can describe the timbre verbally, along the dimension of “*Dull - Sharp*”. A sound which is described as sharp has its overall energy concentrated in upper region of the spectrum.

Psychoacoustics associate timbre with models of sharpness (or brightness), roughness and fluctuation strength [29]. *Fluctuation strength* describes the sensation caused by slow amplitude modulation (4 Hz) within auditory filters. *Sharpness* describes the distribution of frequencies in the spectrum (spectral envelope). *Roughness* is the sensation caused by rapid amplitude modulation (less than 70 Hz) within auditory filters and is the result of *beating* between the frequency components. It is often used as a measure of tonal dissonance or consonance [40]. Sounds that introduce fast beating are considered to be rough or dissonant.

Some dimensions describe the steady portion of the sound in terms of spectral content and other dimensions describe its temporal evolution. Important timbre attributes calculated directly from the spectrum of a sound are the *spectral centroid*, *log of the rise time (attack time)* and the *spectral flux*. Spectral flux, shows how fast the power spectrum changes from frame to frame.

## 3.5 Auditory Scene Analysis

Auditory Scene Analysis (ASA) was introduced by Albert Bregman [12]. ASA is the process of the listeners cognition that groups or segregates sounds into auditory streams, in correspondence with real world phenomena. An *auditory stream* is formed by a perceptual grouping of different parts of the sound that seem to belong together. A nice visual example demonstrating segregated sequences is given by Bregman (shown in Figure 3.7). The sequence of letters is visually segregated in two different streams sentences. The one sentence is “A cat sits”, and the other is “And I sit too”. Another example of stream segregation can be illustrated through the “cocktail party” effect. One can follow a conversation despite the ‘noise’ caused by music or other people talking.

### 3. AUDITORY PERCEPTION

---

AI CSAITT STIOTOS

AI CSAITt StIoToS

**Figure 3.7:** Visual stream segregation. After Bregman [12].

Streams are segregated by processing the sound sequentially (in successive time frames) but also simultaneously. The level of listening attention is also very important. When listening to an orchestra playing, instruments of the same timbre can be grouped together but is also possible for one to follow the individual melodic lines of different instruments.

Segregation is a precondition for *auditory grouping* and can be seen as a top-down or bottom up process. As a top-down process *schema based segregation* makes use of attention and learning. For example one can easily hear his/her name being mentioned in a conversation in which he/she is not actively participating. As a bottom up process *primitive stream segregation* makes use of acoustic cues and is considered to be innate. Major acoustic cues for *primitive grouping* are:

- Proximity in frequency and time
- Periodicity
- Continuity
- Onset and offset
- Amplitude and Frequency modulation
- Rhythm
- Spatial location

Auditory Scene Analysis borrows some concepts found in Gestalt psychology which were originally developed for examining visual perception. According to gestalt psychology, our visual perception is based on considering figures as a whole rather on the examination of their component parts. Some important gestalt principles used in the auditory domain are:

### 3.5 Auditory Scene Analysis

---

- **Similarity:** The sounds share the same attributes. For example if they belong to the same timbre family, they are perceived to be similar.
- **Proximity:** Sound components are spaced close together (e.x in frequency). The closer the spacing the more probable the grouping.
- **Good continuation:** A sine wave interrupted (briefly) by white noise will be perceived as a single sound.
- **Common Fate:** Sounds that share a common kind of change. For example if they change at the same rate (e.x onset/offset, modulation) or towards the same direction (e.x in frequency).
- **Closure:** Our brain tends to complete forms even if they are presented incomplete. A good example is the missing fundamental. The fundamental frequency is perceived in a complex tone, though it may not be physically present.

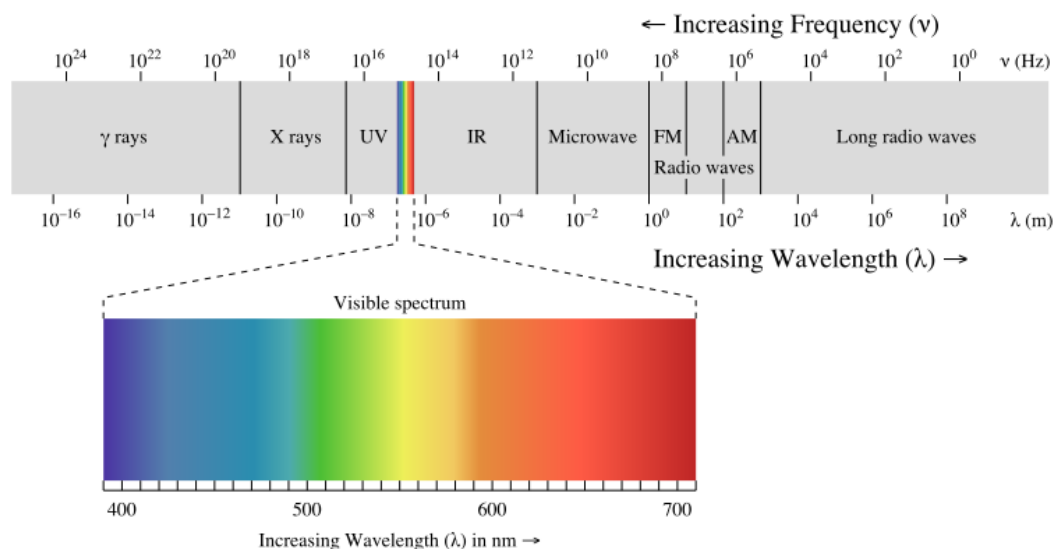
### 3. AUDITORY PERCEPTION

---

# Chapter 4

## Color

The visible part of the electromagnetic spectrum spans a region from 400 nm to 700 nm. The sun-light contains almost an equal amount of different wavelengths and is often referred to as ‘white’. Early color experiments started with Newton, who showed that white light can be decomposed through a prism into a continuum of different colors (wavelengths). Monochromatic light (i.e consisting of a single



**Figure 4.1:** Visible spectrum Adapted from <http://en.wikipedia.org/wiki/File:EMspectrum.svg>

wavelength) is very rarely found in nature. What we usually experience as color,

## 4. COLOR

---

is a mixture of many different wavelengths in different proportions. As it will later become evident, the perception of color is a pure psychological phenomenon, depending both on the spectrum of the emitted light and the physiology of our visual system.

### 4.1 The Human Eye and the Sensation of Color

The human eye is a sphere, typically 12mm in radius, enclosed by a protective membrane, the sclera. The main structures are the iris, lens, pupil, cornea, retina, vitreous humor, optic disk and optic nerve. The light passes through the cornea and aqueous humor and reaches the pupil, which regulates the amount of light admitted to the lens. The ciliary muscle changes the shape of the lens providing variable focus and distance adaptation. The lens focuses the light on the sense cells of the retina, which is located to the rear wall of the eye. Between the lens and the retina is a semi-colorless, viscous material called the vitreous humor, which absorbs some frequencies of the incoming light.

The retina provides the first layer of “image processing” of our visual system. There are two types of light - sensitive cells in the retina: *rods* and *cones*. Each retina contains about 120 million rods and 8 million cones. When they respond to the light beam, they transduce the input into nerve impulses which are transmitted up the optic nerve, through several substructures, to the visual cortex of the brain.

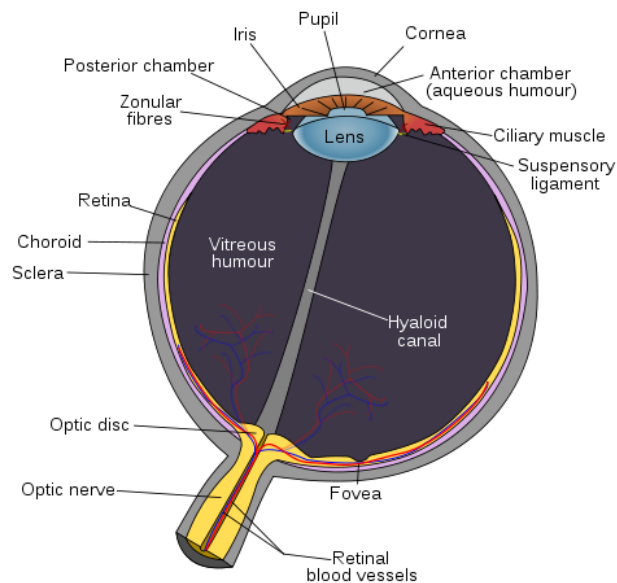
*Rods* are mainly concentrated in the periphery of the retina and operate best at very low light levels. They are extremely sensitive to light and insensitive to color, thus they provide achromatic vision at low levels of illumination (scotopic vision, night vision).

*Cones* are mainly concentrated in the central vision center of the retina, in particular in the fovea. They are less sensitive to light than rods but they provide both luminance and color vision in daylight (photopic vision).



## 4.1 The Human Eye and the Sensation of Color

---



**Figure 4.2:** The human eye Adapted from [http://en.wikipedia.org/wiki/File:Schematic diagram of the human eye en.svg](http://en.wikipedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg)

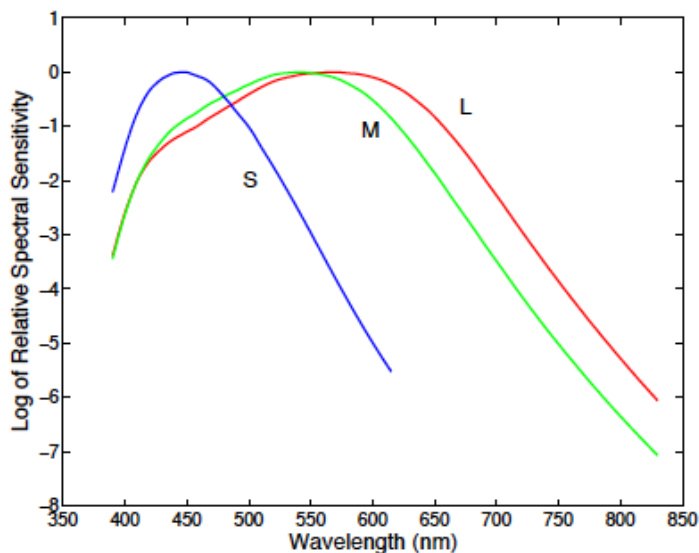
### 4.1.1 The Tristimulus theory

There are three types of cones, which act as band pass filters to the incoming light:

- **S-type** (or B) cones get excited by small wavelengths and show a peak sensitivity around 420nm. Therefore, they make a stronger contribution to the perception of blue.
- **M-type** (or G) cones get excited by medium wavelengths and show a peak sensitivity around 550nm. Therefore, they make a stronger contribution to the perception of green.
- **L-type** (or R) cones get excited by long wavelengths and show a peak sensitivity around 570nm. Therefore, they make a stronger contribution to the perception of red.

## 4. COLOR

---



**Figure 4.3:** Log of the relative spectral sensitivities of the three kinds of colour receptor in the human eye. Figures plotted from data available at <http://www.cvrl.org>

### 4.1.2 The Opponent Process theory

The opponent process theory was first proposed by Ewald Hering and it was later verified by neurobiologists, who discovered the existence of cells responsible for this process. According to this theory, the photoreceptor outputs interact to produce the following three color opponent channels:

- **Red - Green:** This channel is referred to as “Red-or-Green” or “Red minus Green”. It demonstrates the fact that color is not experienced as “reddish green” or “greenish red”.
- **Yellow - Blue:** This channel is referred to as “Yellow-or-Blue” or “Yellow minus Blue”. It demonstrates the fact that color is not experienced as “yellowish blue” or “blueish yellow”.
- **White - Black:** Responsible for the perception of lightness.

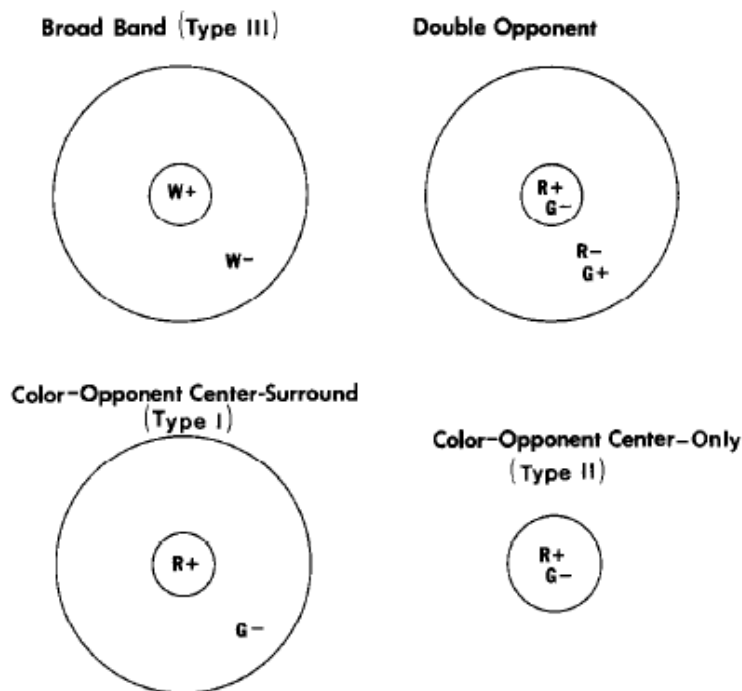
Color opponent cells can be *single-opponent* or *double-opponent*. Single opponent cells are divided in:

## 4.1 The Human Eye and the Sensation of Color

---

- *Concentric single-opponent cells*: Receive input from R or G cones either in the center or surround and have opponent actions. Respond to brightness but also to large spots of monochromatic light.
- *Concentric broad-band cells*: G and R cones act together in either the center or antagonistic surround. Mostly respond to brightness.
- *Co-extensive single opponent cells*: have a uniform receptive field. The B cones are antagonized by the G and R cones acting together.

*Double opponent cells* integrate input from single opponent cells and receive input from both R and G cones (or B with R and G) in the center and the surround of the receptive field. Therefore, respond to red-green and yellow-blue contrasts.



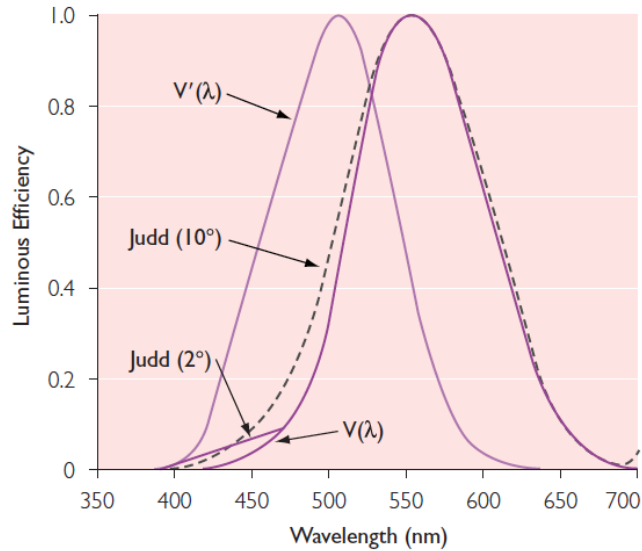
**Figure 4.4:** Types of radially symmetrical cells found in the primate visual system  
Image taken from [34].

### 4.2 Perceptual Attributes of Color

The perception of color is a subjective experience. A stimulus can be defined in terms of physical variables such as *dominant wavelength*, *intensity* and *purity*. Purity is the proportion of white-light and a pure-light needed to define a specific color.

- **Hue:** Psychophysically responds to the dominant wavelength of light and is the basic component of color.
- **Brightness:** This term is used for light sources and psychophysically responds to the perceived intensity of the emitted light. In color terms it can be thought of how much ‘black’ is mixed with a color. Brightness varies as a function of wavelength and therefore some Hues are perceived as brighter than others (Figure 4.5).
- **Lightness:** Relative to brightness and usually refers to objects and the reflected light. More precisely is “The brightness of a stimulus relative to the brightness of a stimulus that appears white under similar viewing situations.”
- **Colorfulness:** “The perceived quantity of hue content (difference from gray) in a stimulus.” Colorfulness increases with luminance.
- **Chroma:** Colorfulness compared to white: “The colorfulness of a stimulus relative to the brightness of a stimulus that appears white under similar viewing conditions.”
- **Saturation:** Psychophysically responds to the purity of a color, in terms of mixture with white or the vividness of Hue. “The colorfulness of a stimulus relative to its own brightness.”

The perceptual dimensions of Hue, Saturation and Brightness, suffice to describe a light which is viewed in isolation. The percept of a single isolated light is called an *unrelated color*. Some colors though such as ‘brown’ or ‘grey’, only exist when the light is viewed within the context of at least one another light.



**Figure 4.5:** The commonly used luminous efficiency functions of vision defined by the Commission Internationale de LEclairage (CIE).  $V(\lambda)$  and  $V'(\lambda)$  is the photopic and scotopic luminous functions respectively. Image taken from [52].

This percept is called a *related color*. Hue, Brightness, Lightness, Colorfulness and Chroma suffice to describe any color.

## 4.3 Color in context

*Color constancy*, refers to our ability to perceive the same colors of objects, despite changes in illumination. However this is only an ability (based mainly on color memory and chromatic adaptation) and not a fact. Late studies [?] show that changes in illumination can result in a perceived difference in the color of objects.

*Chromatic induction*, occurs when the perception of light is modified by the presence of a second surrounding light and may result in:

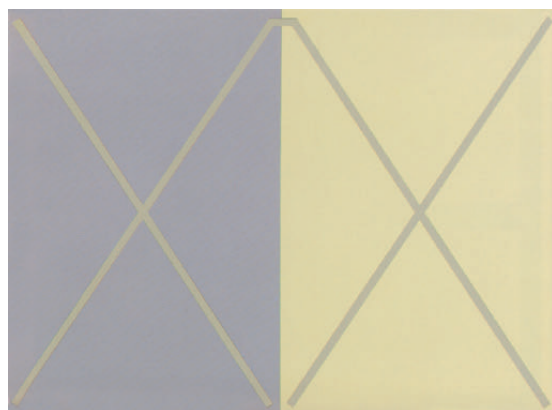
- *Chromatic contrast*: A chromatic inducing field shifts the color of a patch away from the color of the inducing light.
- *Chromatic assimilation*: The appearance of a light shifts toward rather than away from the color of an inducing field.

## 4. COLOR

---



**Figure 4.6:** Simultaneous Contrast: The internal squares all have the same luminance but the changes in luminance in the surrounding areas change the perceived luminance of the internal squares.



**Figure 4.7:** Illustration of context effects. The x-shaped intersections on the two sides of the figure appear quite different. The light reaching the eye from these two regions is the same, however. This can be seen by tracing from one x-shaped region to the other. Adopted from [52]

Fovea is the central portion of the retina of the eye and is responsible for central vision and *visual acuity*. It has a resolution of  $2^\circ$  which corresponds closely to a circular area in the field of view whose diameter is approximately equal to 3.5 % of the distance. *Texture* is the sensation of areas that correspond to this minimum resolution rather than the sensation of single (countable) points that make up the whole. Some attributes of texture are; uniformity, density, coarseness, roughness, regularity, linearity, directionality, direction, frequency, and phase.

## 4.4 Color Spaces

A color space (also color model) is usually a 3-dimensional space used to describe a gamut of colors according to some attributes. Color spaces can be broadly categorized as device-dependent and device-independent. For the purposes of this thesis we will present only some device-dependent color spaces.

### 4.4.1 Trichromacy and Color matching

A particular color  $T(\lambda)$  can be described uniquely by a linear combination of only three primary colors  $P_1(\lambda)$ ,  $P_2(\lambda)$  and  $P_3(\lambda)$ , as long as negative matching is allowed and the three primaries are independent (i.e none of the three is a match for the mixture of the other two).

$$T(\lambda) \equiv a_1 P_1(\lambda) + a_2 P_2(\lambda) + a_3 P_3(\lambda) \quad (4.1)$$

This formula is known as Grassmann's Law and the symbol ' $\equiv$ ' indicates a visual match. The coefficients  $a$  indicate the amount of color required for the match and if they are positive we have an *additive matching*. As we mentioned, the matching may require the addition of another primary or white  $W = P_1 + P_2 + P_3$  in the test color. If we denote  $d$  the amount of white we have:

$$T(\lambda) + dW(\lambda) \equiv a_1 P_1(\lambda) + a_2 P_2(\lambda) + a_3 P_3(\lambda) \quad (4.2)$$

$$T(\lambda) \equiv \underbrace{(a_1 - d)}_{b_1} P_1(\lambda) + \underbrace{(a_2 - d)}_{b_2} P_2(\lambda) + \underbrace{(a_3 - d)}_{b_3} P_3(\lambda) \quad (4.3)$$

The coefficients  $b$  may have negative values but at least one must be positive. In this case we have a *subtractive matching*.

The spectral distribution of light and the perceived color is a many-to-one mapping. The essence of trichromacy, is that the perceived hue depends on the three dimensional vector of signals detected by the three cone mechanisms in combination. Two spectrally different colors that point to the same vector can not be distinguished by the human eye. For example, equal excitations of the G and R cones by a bichromatic light at 530 nm and 630 nm will produce

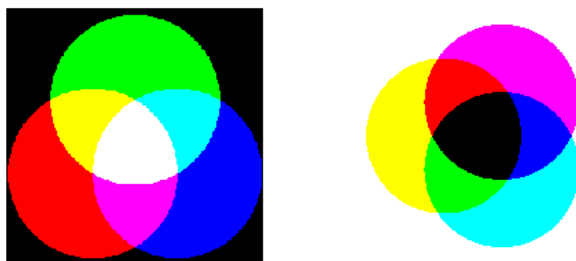
## 4. COLOR

---

the sensation of “yellow”, the exact same sensation that would result from an excitation by a monochromatic light at 550 nm. This phenomenon is referred to as *color metamerism*.

### 4.4.2 The RGB Color Space

RGB is a linear color space based on additive mixing. It uses as primary colors the Red, Green and Blue which are typically assigned to wavelengths of 645.16nm, 526.32nm and 444.44nm respectively. Practically though, RGB uses whatever phosphors a monitor has as primaries. The gamut of available colors is presented within a cube (often called the RGB color cube) whose edges represent the R, G and B weights (Figure 4.9)



**Figure 4.8:** Left: An example of additive mixing based on the RGB colorspace. Right: An example of subtractive mixing based on the CMY colorspace.

### 4.4.3 The CMY(K) Color Space

CMY is also a linear color space based on subtractive mixing. It uses as primaries the Cyan ( $C$ ), Magenta ( $M$ ) and Yellow ( $Y$ ) which are the complement colors of



the Red, Green and Blue respectively.

$$(C, M, Y) = (1 - R, 1 - G, 1 - B) \quad (4.4)$$

CMY is used in color-printing devices since Cyan, Magenta and Yellow are the primary colors of pigments. The addition of Black (K) as an extra primary is very important, since mixing colored inks results to Brown.

#### 4.4.4 The YCbCr Color Space

YCbCr is an orthogonal color space and separates brightness (Y), from a red-difference chroma component (Cr) and a blue-difference chroma component (Cb). Since our eye is more sensitive in brightness rather than color, this colorspace is widely used for image and video compression (e.x JPEG, MPEG). Analogous to YCbCr is the YPbPr which is used for analog video and is derived from a linear combination of the R, G, and B analog values (ranging from 0 to 1) using two defined constants ( $K_B$ ) and ( $K_R$ ):

$$Y' = k_R R + (1 - k_R - k_B)G + k_B B \quad (4.5)$$

$$P_b = \frac{(B - Y)}{2(1 - k_B)}, \quad P_r = \frac{(R - Y)}{2(1 - k_R)} \quad (4.6)$$

The values of ( $K_B$ ) and ( $K_R$ ) are typically set to 0.30 and 0.11 respectively. The digital colorspace YCbCr (with 8 bits per channel per pixel) derives by scaling and offset the YPbPr :

$$(Y, C_b, C_r) = (16, 128, 128) + (219Y', 224P_b, 224P_r) \quad (4.7)$$

#### 4.4.5 The HSV, HSL and HSI Color Spaces

These color spaces result from non-linear transformations of the RGB colorspace and are defined in polar coordinates; Hue (H) is a function of the angle and Saturation (S) is proportional to radial distance. Value (V, meaning Brightness),

## 4. COLOR

---

Lightness (L) and Intensity (I) are the distances along the axis perpendicular to the polar coordinate plane.

Since we think about color in terms of Hue, Saturation and Brightness, such color spaces have the advantage to represent color intuitively, as opposed to the other color models presented above. For example it is very difficult for one to specify the required weight of primaries (or the coordinate values) in order to match a target color. If we rotate the RGB color cube so that its neutral (grey) axis  $(0,0,0) - (1,1,1)$  becomes the vertical axis to the new 3-D space (Figure 4.9), we can define brightness  $L(\mathbf{c})$  as a function of a color  $\mathbf{c} = (R, G, B)$ :

$$L_{HSV}(\mathbf{c}) = V = \max(R, G, B) \quad (4.8)$$

$$L_{HSL}(\mathbf{c}) = L = \frac{\max(R, G, B) + \min(R, G, B)}{2} \quad (4.9)$$

Through the above transforms, the HSV colorspace is geometrically a hexcone and HSL is a double-hexcone. The primary colors lay on the surface level which corresponds to white ( $V=1$ ) in the HSV model and middle gray ( $L=0.5$ ) in the HSL model. *Chroma* ( $C$ ) is used for defining Saturation ( $S$ ):

$$C = \max(R, G, B) - \min(R, G, B) \quad (4.10)$$

$$S_{HSV} = \frac{C}{V}, \quad S_{HSL} = \frac{C}{1 - |2L - 1|} \quad (4.11)$$

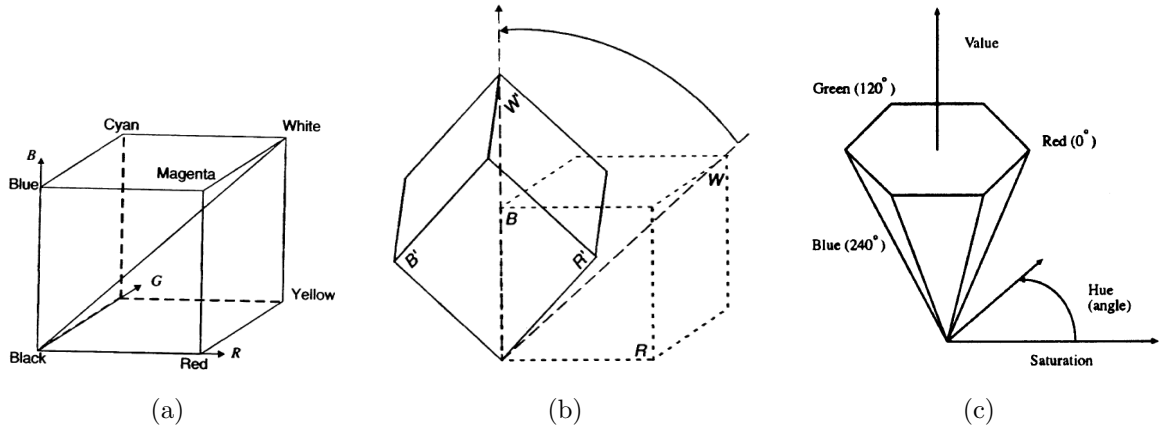
Hue ( $H$ ) is usually expressed in degrees  $[0, 360)$ . A transformation from hexagon to circle is achieved via:

$$C_2 = \sqrt{R^2 + G^2 + B^2 - RG - RB - GB} \quad (4.12)$$

$$\theta = \cos^{-1}[0.5(2R - G - B)/C_2] \quad (4.13)$$

$$H = \begin{cases} \theta/360, & G \geq B \\ 1 - \theta/360 & G \leq B \end{cases} \quad (4.14)$$

The HSI colorspace is widely used in computer vision:



**Figure 4.9:** The HSV hexcone (c) after the rotation of the RGB color cube (a, b). Image taken from [36].

$$H_{HSI} = H, \quad L_{HSI}(\mathbf{c}) = I = \frac{R + G + B}{3}, \quad S_{HSI} = 1 - \frac{\min(R, G, B)}{I} \quad (4.15)$$

Conclusively, the definitions of Saturation in the HSI and HSV models are closer to the psychometric definition given in paragraph 4.2

## 4.5 Gestalt Principles

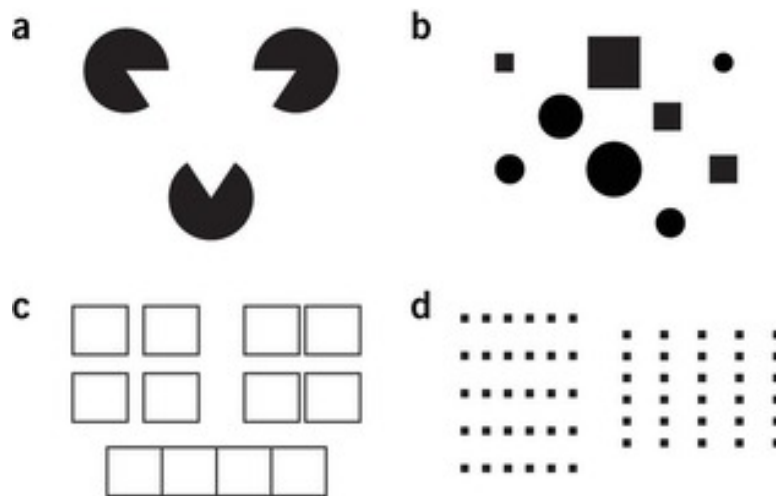
As mentioned in Chapter 3, our visual perception relies on some principles found in Gestalt psychology. The same principles that were described in the auditory domain will also be described here.

- **Similarity:** Elements with similar characteristics will be seen as though they are grouped together.
- **Proximity:** Elements located near one another will tend to be seen as a group or unit.
- **Common Fate:** Elements engaged in the same pattern motion or common occupation will be seen as though they are grouped together.

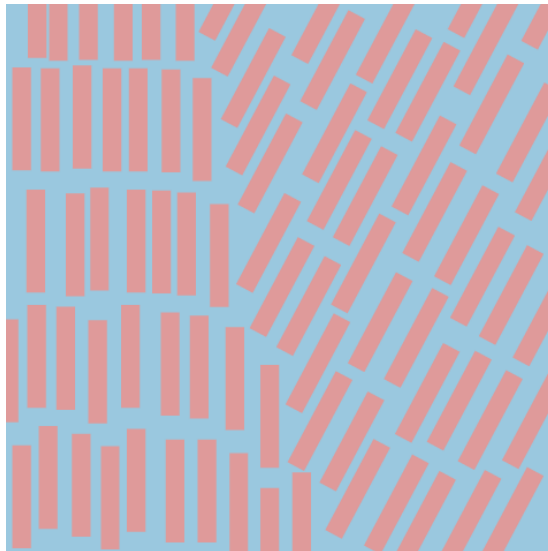
## 4. COLOR

---

- **Good Continuation:** If possible, the perceptual field will be organized along lines of continuous contour or flow.
- **Closure:** A field containing broken figure parts, is usually organized to be seen as a number of closed figures.



**Figure 4.10:** (a) The Kanisza triangle showing the gestalt principles of good continuation and closure. (b) Similarity. (c) Proximity. (d) Relative proximity. Image taken from: <http://www.nature.com/nmeth/journal/v7/n11/carousel/nmeth1110-863-F1.jpg>



**Figure 4.11:** Common Fate. Image taken from: [http://www.teaching.louisabufardec.net/111/files/weblog/designTips/worksheets/gestalt\\_common-fate.gif](http://www.teaching.louisabufardec.net/111/files/weblog/designTips/worksheets/gestalt_common-fate.gif)

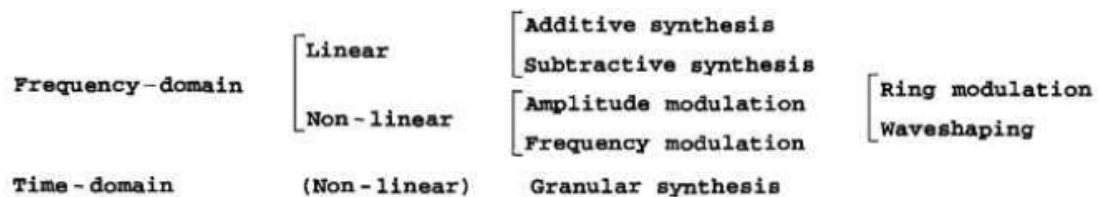
## 4. COLOR

---

# Chapter 5

## Sound Synthesis

In this chapter we present some classic sound synthesis techniques. In the case of sonification, some techniques may be more amenable than others, with respect to the dimensionality and type of the data.



**Figure 5.1:** Classic synthesis techniques classified according to their principles of realization. Image taken from [10].

### 5.1 Additive Synthesis

*Additive Synthesis* is one of the earliest synthesis techniques. According to Fourier theory, any signal can be composed by a sum of sinusoids, each one having its own amplitude, phase and frequency.

$$s(t) = \sum_i a_i(t) \sin(2\pi f_i t + \phi_i) \quad (5.1)$$

This is digitally implemented as a bank of sine or cosine oscillators with specified amplitude values, frequencies and phase offsets. The above equation (5.1)

## 5. SOUND SYNTHESIS

---

represents static spectra. A time varying timbre can be achieved by using time dependent frequency and amplitude values for each oscillator:

$$s(t) = \sum_i a_i(t) \sin(f_i(t)t + \phi_i) \quad (5.2)$$

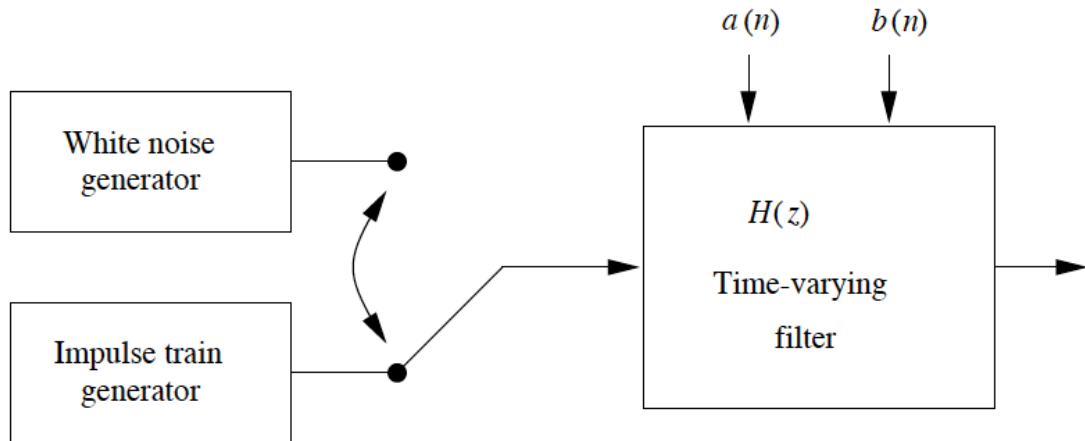
Additive synthesis has a highly predictable output, therefore can be effectively used in *Parameter Mapping sonification*. For example, data may represent the frequency ratios of the partials.

In the early years of computer music, additive synthesis was abandoned in favor of *subtractive synthesis*, due to the required computational load. Noisy parts of the signal, such as transients require a huge amount of oscillators for an adequate reproduction. The computational load can be severely reduced by using a *table look-up* oscillator along with *interpolation* between the values, instead of sine or cosine functions; A wavetable (whose length is usually a power of 2) is filled with the values of a sine or cosine function and is read periodically at the desired speed.

### 5.2 Subtractive Synthesis

*Subtractive Synthesis* is the counterpart of additive synthesis and is based on a *source-filter* model. The spectral envelope of the source-sound is altered through a filter, forming a new one. The source can be any waveform, usually rich in harmonics such as white noise, square waveform or any other arbitrary waveform, which is afterwards filtered. Filter characteristics such as type, bandwidth, cut-off or center frequency and peak amplitude, control the overall quality of the sound and can be effectively used as control parameters in Parameter-Mapping sonification. Dynamic spectra can be achieved if the parameters of the filter are time-varying and ‘noisy’ signals can be easily generated, as opposed to additive synthesis.





**Figure 5.2:** Source-filter synthesis. The transfer function of the time-varying filter  $H(z)$  is described by filter coefficients  $a(n)$  and  $b(n)$ . Image taken from [54]

## 5.3 Modulation Synthesis

The multiplication of a carrier signal  $f_c$  with a modulation signal  $f_m$  will result in; *Amplitude modulation* if the modulator is a uni-polar signal, or in ring modulation if the modulator is a bipolar signal. When the modulator is in the inaudible frequency range (i.e frequencies below 20Hz) the perceived effect has a tremolo quality. Higher modulation frequencies will cause sidebands. In simple amplitude modulation, the frequency of the carrier signal will be present and the sidebands  $(f_c + f_m)$ ,  $(f_c - f_m)$  will have half the amplitude of the carrier frequency. Ring modulation can be expressed as:

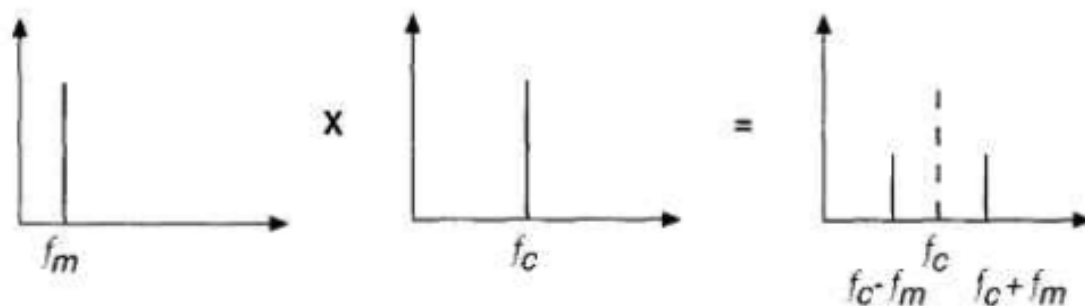
$$s(t) = \sin(2\pi f_m t) \sin(2\pi f_c t) \quad (5.3)$$

In this case the carrier frequency is left out from the final spectrum and the sidebands may be harmonic, if the frequencies of the modulator and the carrier are an integer ratio of one another, or otherwise inharmonic.

Another type of modulation is the case of applying a transfer function to an input signal. The shape of the output-waveform depends on the shape of the input-waveform, the type of the transfer function but mostly on the amplitude of the incoming signal. This type of modulation is called *waveshaping*. Figure 5.4

## 5. SOUND SYNTHESIS

---



**Figure 5.3:** Spectrum of simple amplitude modulation. Image taken from [10].

shows an example of waveshaping: The transfer function

$$f(x) = x^2 \quad (5.4)$$

is used to square the input signal

$$x(n) = a \cos(\omega n + \phi) \quad (5.5)$$

The result will be:

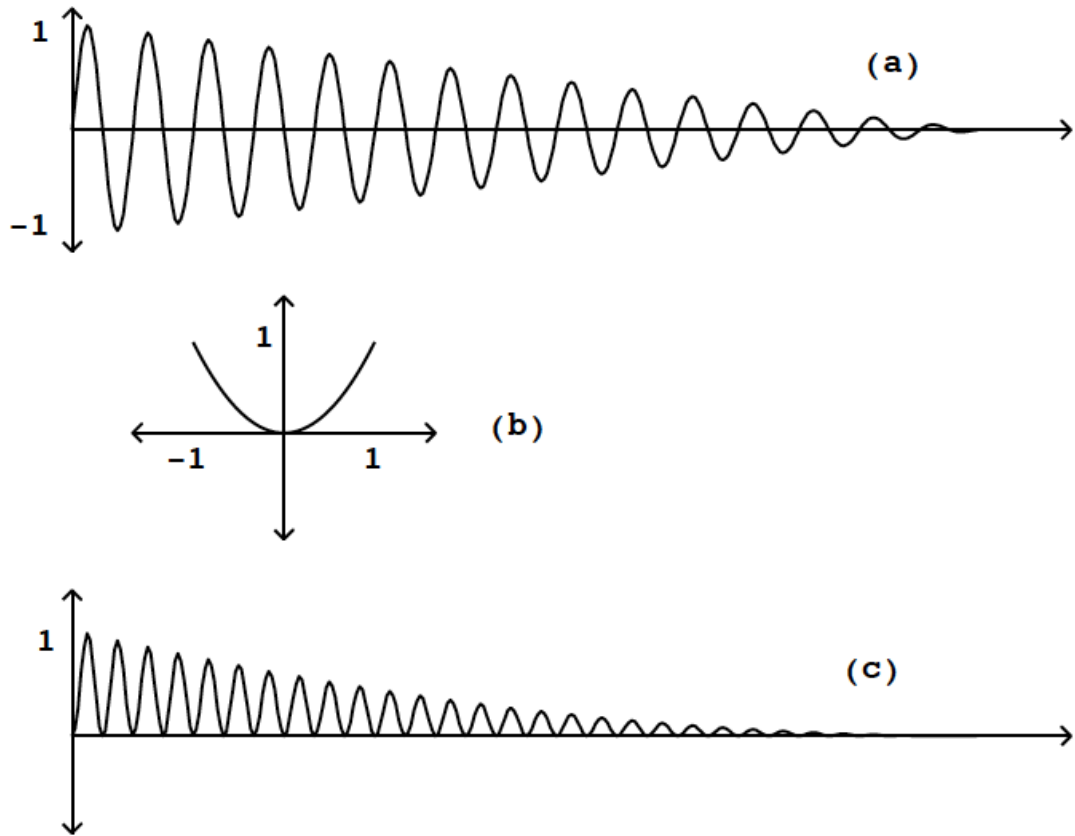
$$f(x[n]) = \frac{a^2}{2} (1 + \cos(2\omega n + 2\phi)) \quad (5.6)$$

*Frequency Modulation* is a special case of waveshaping in which a modulator with angular frequency  $\omega_m$  modifies the frequency of a carrier  $\omega_c$ . The resulting signal is richer in harmonic content than in ring modulation. It can be expressed as:

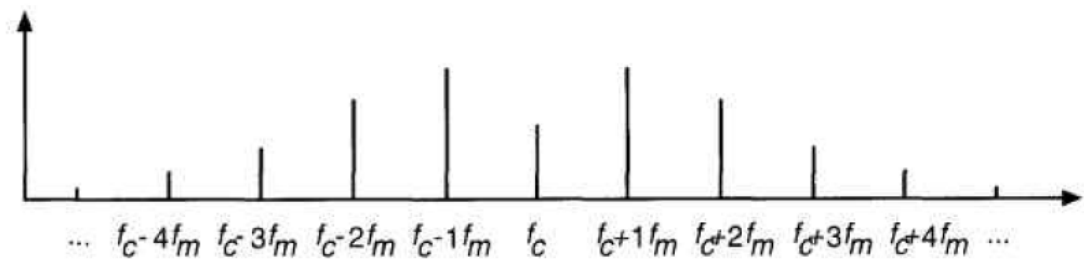
$$s(t) = \sin(a \sin(\omega_m t) + \omega_c t) \quad (5.7)$$

The parameter  $a$  is called the *index of modulation* and controls the amount of distortion applied to the carrier. The resulting spectra can be seen in Figure 5.5. If the modulator varies slowly with time, the perceived effect has the quality of vibrato, otherwise strong sidebands arise.

As in ring modulation, if the ratio of the carrier  $f_c$  to modulator  $f_m$  is not a rational number the spectrum will be inharmonic. Otherwise, if  $f_c/f_m$  is a ratio of integers, the spectrum will be harmonic and the fundamental frequency  $f$  is determined by equation 5.8:



**Figure 5.4:** Waveshaping using a quadratic transfer function  $f(x) = x^2$ : (a) the input; (b) the transfer function; (c) the result, sounding at twice the original frequency. Image taken from [46].



**Figure 5.5:** The spectrum of frequency modulation. Image taken from [10]

## 5. SOUND SYNTHESIS

---

$$\frac{f_c}{f_m} = \frac{N_c}{N_m}, \quad f = \frac{f_c}{N_c} = \frac{f_m}{N_m} \quad (5.8)$$

where  $f_c$  and  $f_m$  are the  $N_c$ th and  $N_m$ th harmonics respectively.

The sound-output of these modulation techniques is often unpredictable, therefore are seldom used in sonification. The mapping of data to specific sound qualities is less direct and intuitive when compared to other synthesis techniques. Furthermore, there is only a limited number of control parameters which makes it unsuitable for the sonification of high-dimensional data.

### 5.4 Physical Modeling

In general, *physical modeling* uses computational models which describe the physics of natural sound sources such as musical instruments or the human voice. Some examples of physical modeling are:

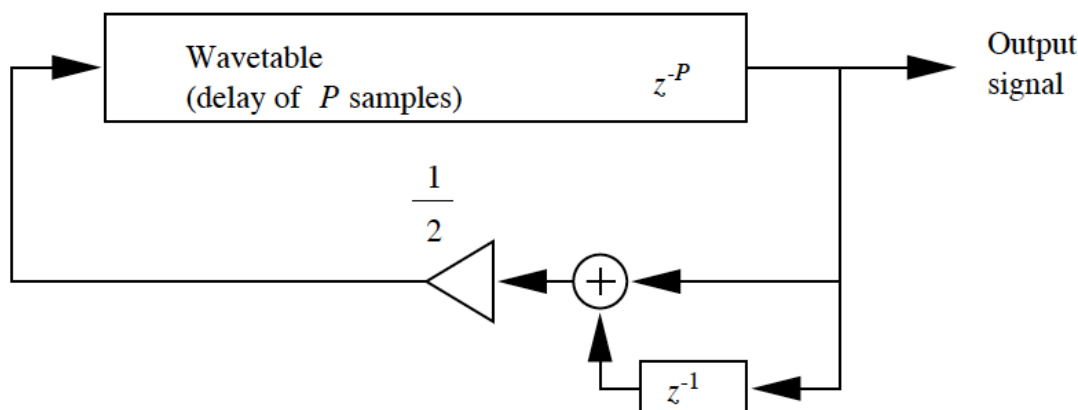
- *Mass-spring networks*, are implemented by using blocks of digital filters.
- *Delay lines* (usually called digital waveguides) in combination with *digital filters* and *non linear elements* such as scattering junctions which transmit and reflect part of the wave.
- *Modal Synthesis*, in which a sound source is analyzed according to its normal modes, mode frequencies and decay times. Additive synthesis can afterwards be used for the sound simulation, according to the analysis data.

A classic, simple and computationally efficient example of physical modeling synthesis, is the *Karplus-Strong* algorithm which can be used for the simulation of plucked-string and drum sounds. Figure (5.6) shows the Karplus-Strong plucked-string model. A wavetable filled with random numbers, is read periodically and the output sample  $y(n)$  is the average of two consecutive samples of the wavetable:

$$y(n) = \frac{1}{2}[y(n - P) + y(n - P - 1)], \text{ where } P \text{ is the delay line length.} \quad (5.9)$$

The transfer function of this system represents a lowpass filter responsible for the decay of the tone:

$$H(z) = \frac{1}{2}(1 + z^{-1}) \quad (5.10)$$



**Figure 5.6:** A Karplus-Strong model for plucked string tones. Image taken from [54]

## 5.5 Granular Synthesis

Granular synthesis was first introduced by Denis Gabor in his classic paper “Theory of Communication” [20]. According to this theory, any sound can be described in terms of short grains with typical durations between 10ms to 100ms. Each grain is consisted of an amplitude envelope and a waveform. Figure (5.7) shows a sine wave enveloped by a gaussian window.

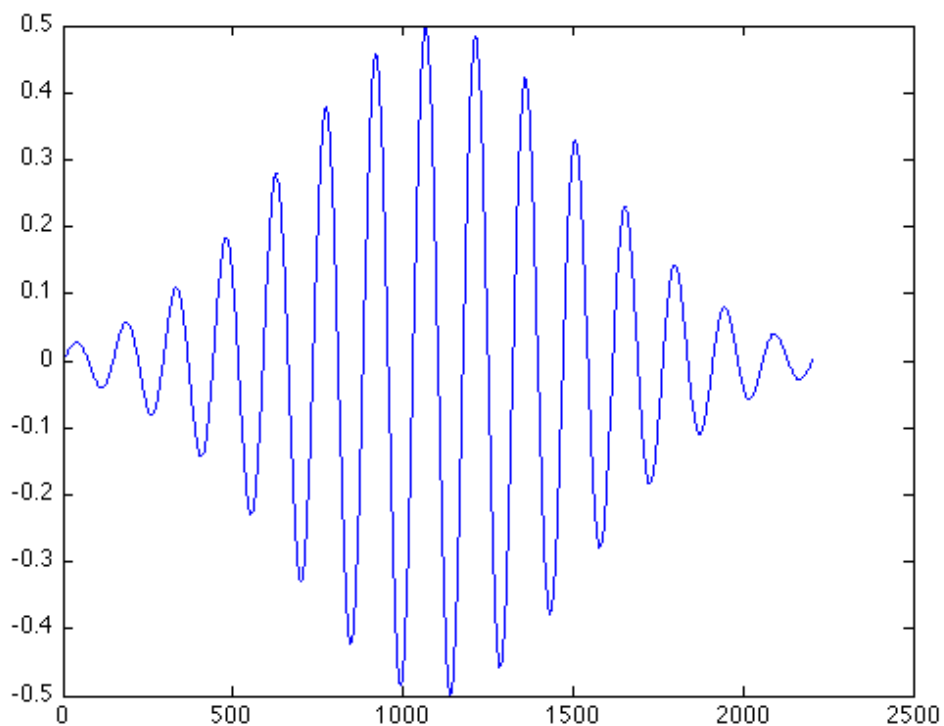
The envelope can be of any shape and has a strong impact on the final sound. Since grains last only a few milliseconds, one has to build a stream of grains (often called *clouds*) to get a usable sound output. The generation of such streams is facilitated when a higher level of organization exists. Sonification can be used to control parameters such as:

- Grain waveform type
- Grain envelope

## 5. SOUND SYNTHESIS

---

- Grain duration
- Cloud density, which expresses the number of grains per unit time.
- Cloud amplitude envelope



**Figure 5.7:** A granule of 50ms enveloped by a Gaussian window.

The interested reader can find more information on Granular synthesis and similar techniques in [48].

## Chapter 6

# Parameter Mapping and Image Sonification

There are at least four different encoding schemes which our brain uses, for the internal representation of information derived from audible frequencies [42]:

- *Verbal*
- *Visuospatial*, where the information is encoded in a ‘picture-like image’.
- *Motor*, which is associated with rhythm.
- *Sensory-Musical* representations.

Our approach should exclude the Verbal representation and we will try to sonify visual representations (images) in a way which the other three encoding schemes are internally preserved.

S. Barrass in [5] briefly mentions some sonification methods and offers valuable design tips. Since we found ourselves “Designing a display to support exploration and discovery of patterns in complex multi-attribute, multi-dimensional and/or time-varying data”, we will use the *Parameter Mapping Sonification* (PMSon) method, in order to get “a perceptually structured information soundscape”.

An effective PMSon should be:

## 6. PARAMETER MAPPING AND IMAGE SONIFICATION

---

- *Intuitive*: Intuition is directly related with *perceptually valid* mappings and *polarity issues*. For example, when sonifying the temperature, a positive polarity would use an increasing frequency for a rising temperature, and a negative polarity would use a decreasing frequency for a rising temperature.
- *Pleasant*: The user of the system should not feel uncomfortable or annoyed even after a long period of listening.
- *Precise*: The data should be prepared prior to sonification according to the available sound synthesis parameters but also according to their structure. In some cases, when data are regarded as noisy, a data reduction step is required (for example through Principal Component Analysis) otherwise the sound result will be noisy as well, especially in the case of high dimensional data.

Data may represent a continuous variable, a statistic process or there are cases in which the data points are not equidistant. According to the data form, a suitable sound synthesis technique has to be chosen. For example when sonifying statistic processes, granular synthesis may be one of the most suitable options, as done in [11, 59, 47].

### 6.1 A formalization of Parameter Mapping Sonification

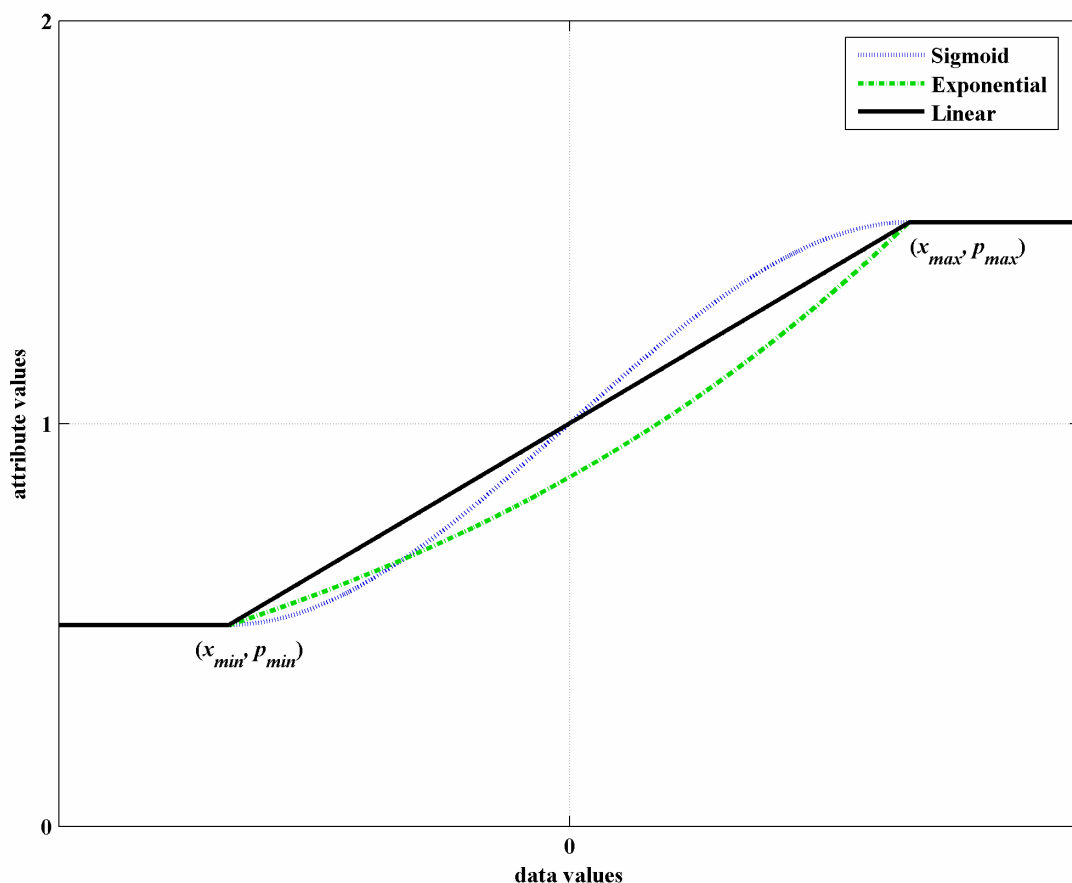
Hermann in [23] introduced the following formalization of PMSon: Given a  $d$ -dimensional dataset  $[\vec{x}_1, \dots, \vec{x}_N]$  and an  $m$ -dimensional vector  $\vec{p}$  of acoustic attributes (which are parameters of the signal generator), an acoustic event in the sonification can be described by a signal generation function  $f : \mathbb{R}^{m+1} \rightarrow \mathbb{R}^q$  which computes a  $q$ -channel sound signal  $s(t) = f(\vec{p}, t)$ . A Parameter Mapping Sonification is then computed by:

$$s(t) = \sum_{i=1}^N f(g(\vec{x}_i), t) \quad (6.1)$$

where  $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$  is the parameter mapping function.



The shape of the mapping function may have various shapes as seen in Figure 6.1. For example it can be linear (an ideal and rare situation), exponential, step-function, sigmoid etc...



**Figure 6.1:** Typical transfer functions for parameter mapping. The black line shows a piecewise linear transfer function. The blue and green dashed lines are respectively sigmoid and exponential transfer functions. Image taken from [18].

## 6.2 Mapping Topologies

Hermann in [23] proposes also a more simple and readable textual representation of the mapping function. The variable names of the dataset, point to attribute vectors which are been given a meaningful name as it can be seen in the following

## 6. PARAMETER MAPPING AND IMAGE SONIFICATION

---

assignment tables. The mapping topology may be :

- **One to One**

Data Feature		Sound Synthesis Parameters
petal length [-, -]	→	onset [0, 2]
petal width[3, 5]	→	pitch [7, 10]
sepal length [-, -]	→	amplitude [60, 90]
0.5	→	duration

**Table 6.1:** One to One Mapping. Table taken from [23]

- **One to Many**

Data Feature		Sound Synthesis Parameters
datafeature1[0, 30]	→	$\Delta$ gain [-90 dBV, 0 dBV]
datafeature1[20, 50]	→	between unvoiced and voiced
datafeature1[40, 70]	→	blends between the vowels [a:] and [i:]
datafeature1[60, 90 ]	→	fundamental freq 82 to 116 Hz
datafeature1[80, 100]	→	brightening of the vowel

**Table 6.2:** One to Many Mapping. Table taken from [25]

- **Many to One**

Data Feature		Sound Synthesis Parameters
saturation [-, -]	→	$\Delta$ gain [-50 dBFS, 0 dBFS]
total number of pixels in a region [-, -]	→	$\Delta$ gain [-50 dBFS, 0 dBFS]

**Table 6.3:** Many to One Mapping.

The symbol ‘\_’ means that the limits are extracted from the minimum and maximum values directly from the data and that are not predefined. This has important consequences on the sound output (as it will be shown in later chapters) and also reduces the reproducibility of the system.

### 6.3 Drawbacks or Advantages?

Though PMSon offers enough flexibility in high-dimensional data, still has some drawbacks [28]:

1. *No unique mapping*: Increasing the dimensionality of the data leads to a larger number of parameter mapping possibilities.
2. *Limited Dimensionality*: The sonification is limited in dimensions according to the parameters of the sound engine.
3. *Variance*: It's not invariant to rotations or translations of the data.
4. *Independence/Perceptual uniformity*: The pitch is perceptually nonlinear. Pitch and duration are perceptually dependent.
5. *Relationship*: There is no possibility to exploit the relationship between different data points, for example examining the local density between two different data points.

In fact, some of these drawbacks can act as advantages in image sonification.

- Comments to 1: In image sonification, (if the image is colored) we have a total of 2 by 3 dimensions. 2 for the pixel orientation and the other 3 describe the color (for example RGB triplets, Hue/Saturation/Value...), meaning that the possibilities are quite limited, always with respect to the sound synthesis method. Furthermore, the options of different mappings enhance the artistic applications of sonification.
- Comments to 2: We want the output of the sound engine to be as predictable as possible.
- Comments to 3: Imagine the sonification of a Square. If the image is rotated by 45 degrees around its center the resulting image is called Rhombus. If it looks different, we want it to sound different as well.
- Comments to 4: In most cases, the whole concept of perception is non-linear. Dimensional correlations can also be found in vision.

## 6. PARAMETER MAPPING AND IMAGE SONIFICATION

---

- Comments to 5: In image sonification, this difficulty can be overridden in the data preparation step. If the sonification is not a simple superposition of different events, but is based on rules with respect to the organization of the data, data relationships can be audible.

Besides, the same author uses PMSon to sonify a stack of images [27].

The final implementations depend on whether we want to sonify an image or sonify the process of observing an image. As nicely put in [53]: “This metaphor is thus no longer a sonification of a particle system, but the sonification of the observation of a particle system”. When we observe an object through multiple views in 2 dimensions, we can discover its structure in the 3 dimensional space. Likewise, when we ‘move’ through acoustic scenes, we can discover the relationship that exists between multi-dimensional data [26].

# Chapter 7

## Sonification of Shape

In this chapter we present a technique to sonify the shape of an object in an image, with the restriction that the curve is piecewise smooth, simple and closed. For tracing the curve we use the Moore's Contour Tracing Algorithm with Jacob's stopping criteria, and then each pixel is sonified by the order that was traced. We construct a new curve, based on the spatial datasets, which describes an amplitude envelope or the instantaneous frequency of a signal, and in which each tracked pixel has a specific duration, and a spatial position in the stereo field. The listener can draw conclusions about the shape of the curve by decoding the sonified result.

### 7.1 Related work

While many image sonification approaches have been presented since the beginning of ICAD (*International Community for Auditory Display*), little research has been done in sonifying the shape of an object within an image. Furthermore, the latter attempts are based on user interaction, so a direct relationship between the image and the sonified result is missing. This approach aims to establish a direct, one to one relationship between the image and the sonified result.

SoundView [56] is an experimental vision substitution system for the blind. The users of this system explore an image by using a pointer device which acts like a virtual gramophone needle. The sonified result depends both on the color of the area been explored, as well as on the velocity of the pointer. In a later

## 7. SONIFICATION OF SHAPE

---

experiment which tested the usability of SoundView [57], users were asked to identify binary shapes. Though the results were encouraging, curved shapes were harder to recognize, due to the linear movement of the pointing device, which is the most usual exploration motion used. Two quite similar approaches are presented in [50] and [62]. In [50] the workspace is divided between sound areas and no sound areas. As the user points over a region close to the curve, the sound amplitude increases, reaching the maximum value when the pointer lies on the curve. Furthermore, the pitch varies according to the shape of the curve, for example a straight line would produce a steady pitch whose intensity would vary according to the proximity of the pointer to the curve. In [62], another similar shape sonification scheme is used, to aid the visually impaired in basic shape recognition tasks. The image should be binary, and the shape is extracted using the Canny edge detector. When the pointer is placed on a pixel that belongs to the curve, a pitched sound occurs whose frequency depends on the vertical position of the pointer (local area sonification). The shape recognition task is further enhanced by also sonifying the pointer's location to edge distance, using the Felzenszwalb algorithm, where distance is mapped to a pulse train frequency, reaching its maximum whenever the distance to edge is minimum.

The authors of [13] sonify line graphs which contain two data series. The users try to decode the sound and sketch back the two data series, with their possible intersection points, the shape of the curves and the minimal and maximum points. The authors propose the  $x$  horizontal coordinate to represent time, while the  $y$  vertical coordinate is mapped to MIDI notes.

When sonifying more complex graphs or curves, the mapping of time to a dimension of the image is not profound. A challenging task is how can one achieve a dimensional reduction, in our case from 2D to 1D. One transform that allows that, and has been used in the past for texture sonification [38], is the Radon transform. By taking projections in several angles the original image can be reproduced by using the inverse transform. Sonifying these projections is more likely to distract the listener from a shape recognition task. Consider for example two projections of a square shape the first being in 0 degrees and the second at 45 degrees as shown in Figure 7.1. While the 0 degree projection would result in a steady situation, if pitch is used for example as a mapping parameter, the

second one would evolve as a triangle i.e with an increasing and then decreasing pitch.

Another tempting approach would be to sonify the curvature of a shape. Sonifying the magnitude of curvature is presented in [51], in order to detect curvature discontinuities which are important cues for 3D perception [32]. The users of this system explore the shape of a virtual object haptically and hear the sonified result at the same time. Sonification is an important feature of the system because curvature discontinuities are difficult to detect by touching the object or by using the visual feedback of the system. The system calculates the minimum and the maximum absolute values of the curvature and then maps them to a frequency range of 3 octaves. The sign of the curvature (positive for concave and negative for convex curves) is mapped to stereo panning. Discontinuities in curvature resolve to a frequency modulated sound and are the only points in the curve that the sound is not stable over time. Shapes that look smooth may have curvature discontinuities and furthermore the magnitude of curvature does not always evolve. For example, the curvature of a circle is a constant i.e the reciprocal of the radius.

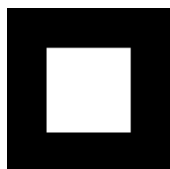
## 7.2 Contour Tracing

The listener should be able to perceive the sonified result, using a visuospatial encoding strategy [42], therefore a perceptually meaningful mapping from the spatial coordinates to time is of vital importance.

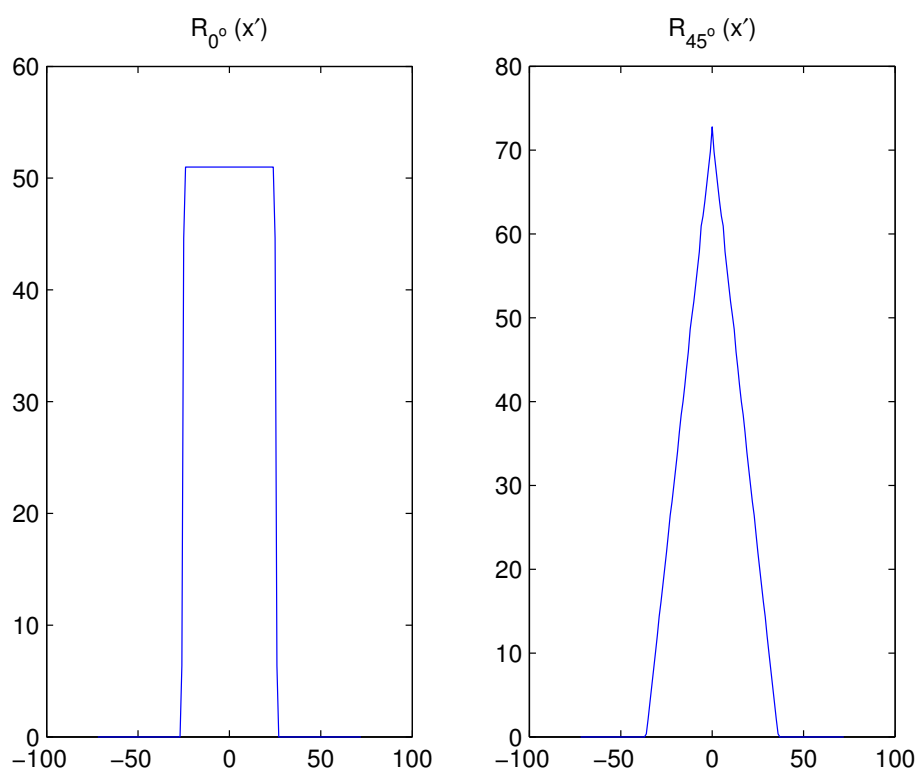
Given a curve,  $f : \mathbb{R} \rightarrow \mathbb{R}^2$  and its parametric representation  $f(t) = (x(t), y(t))$ , its derivatives  $dx/dt$  and  $dy/dt$ , provide information about the evolution of the curve. In order to take them into account in our proposed sonification scheme, we use the Moore's Contour Tracing Algorithm with Jacob's stopping criteria [45]. The algorithm searches for neighboring pixels in a clockwise direction, starting from bottom to top and from left to right, though in this implementation we start from top to bottom. In order to avoid tracing the same pixels as seen in Figure 7.2, we limit our approach for sonifying piecewise smooth, simple, closed curves and positively oriented because of the clockwise search direction. Another restriction is that the Moore's Contour Tracing algorithm works only on lines that

## 7. SONIFICATION OF SHAPE

---



(a) A square shape



(b) Projections at 0 and 45 degrees

**Figure 7.1:** A square and its projections at 0 and 45 degrees based on the Radon Transform.



have single pixel width, so the object has to undergo thinning. The output of the algorithm is a series of  $x[n]$  and  $y[n]$  values representing the  $x$  and  $y$  coordinates respectively, of the tracked pixels. By calculating the derivatives  $x[n] - x[n - 1]$  and  $y[n] - y[n - 1]$  we have a description of the evolution of the curve at a given moment.

$$x[n] - x[n - 1] = \begin{cases} -1 & \text{the curve evolves from right to left,} \\ 0 & \text{no movement in the x axis,} \\ 1 & \text{the curve evolves from left to right.} \end{cases} \quad (7.1)$$

$$y[n] - y[n - 1] = \begin{cases} -1 & \text{the curve evolves from bottom to top,} \\ 0 & \text{no movement in the y axis,} \\ 1 & \text{the curve evolves from top to bottom.} \end{cases} \quad (7.2)$$

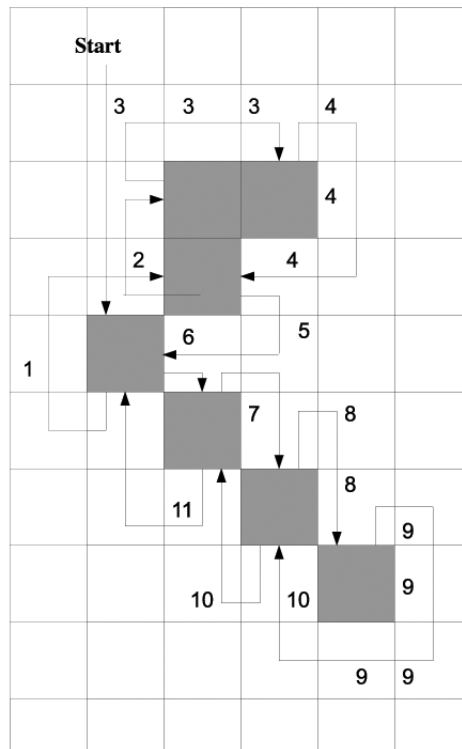
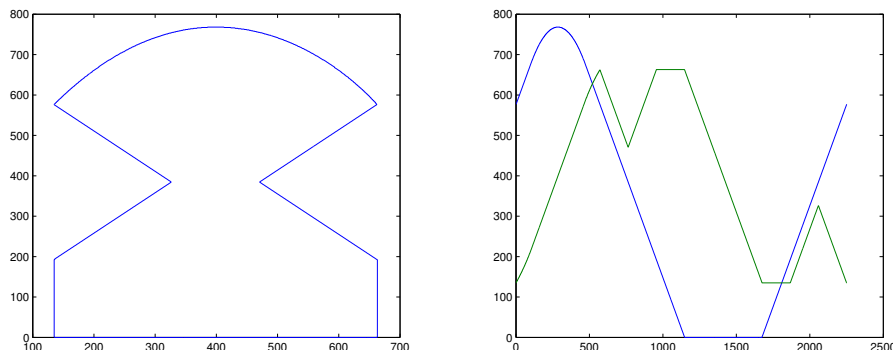


Figure 7.2: The Moore's Contour Tracing Algorithm.

## 7. SONIFICATION OF SHAPE

---



**Figure 7.3:** A curve (left) and the output of Moore's Contour Tracing Algorithm (right).

### 7.3 Mapping Spatial Datasets to sound

From now on, the sonification scheme is very straightforward, and when the coordinate values or their derivatives, are mapped to a different sound quality, the sonified result should be easily encoded in a visuospatial way. Mapping pitch to one coordinate versus amplitude to the other could be problematic, since pitch and amplitude are correlated in human perception. In general, tones below 2 Khz are perceived to decrease with increasing amplitude while tones above 4 Khz increase with increasing amplitude [40]. Furthermore, pitch affects the perception of direction but so does the intensity of a sound, as long as there is no conflict between them. If both are changing, then the perception of pitch dominates the perception of intensity [43]. We continue by presenting two different sonification techniques based on two different mappings: amplitude and pitch versus stereo position. Since we have not taken into account the overall size of the image, the scalings are arbitrary, still meaningful.

Using amplitude as a mapping parameter has the advantage of offering frequency and time independence but on the other hand the sonified data could be heavily compressed, since there is a limited bandwidth in which a 3 decibel difference is perceived as being only one step louder. It might also be undesirable to make use of the full bandwidth because there is always a noise floor in the listening space, so moving towards the lower end of the decibel scale would not

### 7.3 Mapping Spatial Datasets to sound

---

be practically useful, but also because high audio levels could be uncomfortable, or even dangerous for the listener. Because of these limitations we scale the coordinate values according to a power law, before normalizing them into a suitable range. For preliminary testing and demonstration purposes we found the function  $f(x) = x^e$ , to behave well before normalizing the values in the range 0.001 (-60dBFS) to 1 (0dBFS). Pixels are sonified one by one, in the order that were detected by the Moore's Contour Tracing Algorithm. If each pixel is mapped to a constant duration, over its mapped amplitude level, we construct an amplitude envelope  $a(t)$  which is then applied to a 1 Khz sine wave. The listener should draw conclusions about the shape of the curve by decoding the amplitude envelope  $a(t)$  of the final signal (equation 7.3), with respect to panning strategies that will be described in the next sections. Of course, the smaller the pixel grid (the workspace) the easier the shape recognition would be, because the values which are mapped to each pixel would vary greatly and consequently their difference would be easier to perceive.

$$f(t) = a(t)\sin(2\pi 1000t) \quad (7.3)$$

If the size of the image is big enough to cause data compression when amplitude mapping is used, then a mapping to pitch is a better solution. The tracked pixels are divided into regions which are mapped to a base frequency value and then each pixel represents an increment in *cents* according to the region it belongs.

The system of cents is a logarithmic scale used for measuring frequency ratios. There are 1200 cents in an octave, 200 in a whole tone and 100 in a semitone. An interval between two notes with frequencies  $f_1, f_2$  can be measured in  $n$  cents with the following formula:

$$n = 1200 \cdot \log_2\left(\frac{f_2}{f_1}\right) \quad (7.4)$$

For preliminary testing and demonstration purposes, each region consists of 120 pixels, starting from a base frequency of 120 Hz, and each pixel represents a 10 cent increment, as seen in (4) where  $i = 0\dots(\text{total pixels})$  and  $r = 1\dots(\text{total regions})$ :

## 7. SONIFICATION OF SHAPE

---

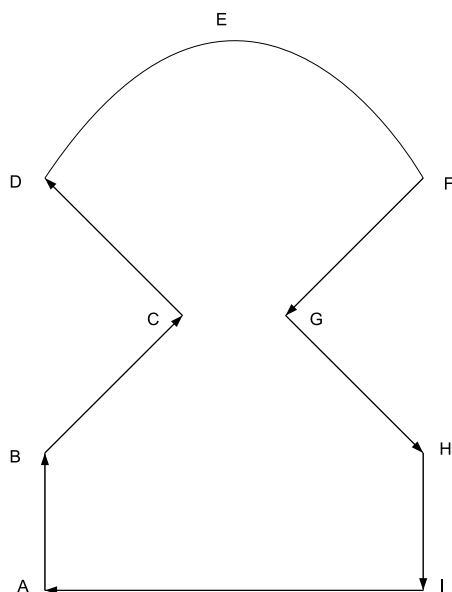
$$p(i) = (120 \cdot 2^{r-1}) \cdot 2^{(i-(r-1) \cdot 120) \cdot 10/1200} \quad (7.5)$$

In this case the listener should draw conclusions about the shape of the curve, by decoding the instantaneous frequency envelope of the signal,

$$f(t) = \sin(2\pi p(i)t) \quad (7.6)$$

along with panning.

Each pixel has a duration mapping of 50 ms, because durations between 50 ms to 70 ms are found to be effective in graph sonification [14], though we found it also practical and effective to use even smaller durations as the image gets larger (relative to the axes).



**Figure 7.4:** An example curve with labeled breakpoints. The sonification starts from point D, since it is the first point traced by the algorithm.

### 7.3.1 Vertical axis to amplitude or pitch, horizontal direction to stereo position

In this sonification scheme, the derivative of the x coordinate is mapped to panning while the y coordinate is mapped to amplitude or frequency, after being normalized to the above mentioned ranges. We apply the following rules:

$$x[n] - x[n - 1] = \begin{cases} -1 & \text{panning is set to the left channel,} \\ 0 & \text{panning is set to the centre,} \\ 1 & \text{panning is set to the right channel .} \end{cases} \quad (7.7)$$

Translating quantitatively the Figure 7.4 in terms of amplitude and stereo position we have: From A to E the amplitude increases starting from its minimum value and reaching its maximum at E. From E to I the amplitude decreases reaching its minimum value again, and from I to A remains at its previous minimum value. Describing the first few breakpoints for the stereo position we have, from A to B the sound is panned in the centre, from B to C the sound is panned to the right channel and from C to D the sound is panned to the left channel.

### 7.3.2 Defining a new curve for amplitude or frequency, horizontal direction to stereo position

In this sonification scheme we construct a new curve based on the derivatives of the x and y coordinates. If we denote:  $c(k) = x(n) - x(n - 1)$  and  $d(k) = y(n) - y(n - 1)$ , then we define a new curve  $T(k)$  by applying the following rules with  $T(1) = 0$  and  $k = 2 \dots (n - 1)$  :

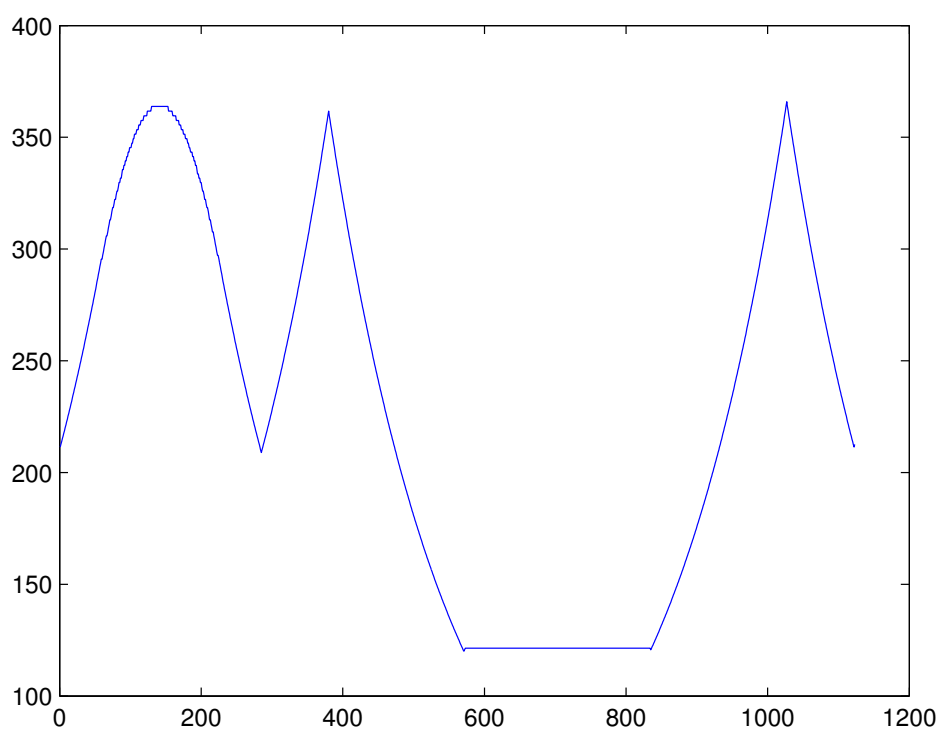
$$T(k) = c(k)d(k) + T(k - 1), \quad \text{if } c(k) \neq 0 \quad (7.8)$$

$$T(k) = d(k) + T(k - 1), \quad \text{if } c(k) = 0 \text{ and } d(k) \neq 0 \quad (7.9)$$

The new curve can then be sonified after being scaled to start from zero and normalized to the values described above using amplitude or frequency, as mapping parameters. The difference between two curves one having  $c(k) > 0$  ,  $d(k) > 0$  and the other having  $c(k) < 0$  ,  $d(k) < 0$  is their orientation. The first is read in a clockwise direction while the other is read anti-clockwise. In order

## 7. SONIFICATION OF SHAPE

---



**Figure 7.5:** The instantaneous frequency envelope, as described in paragraph 7.3.2

the listener to be able to distinguish between these two modes, we also sonify the derivative of the x coordinate separately and mapping it to stereo position as we did in the previous approach described by (6).

Translating again quantitatively the Figure 7.4 in terms of amplitude and stereo position we have: From A to B the amplitude increases and the sound is panned to the centre. From B to C the amplitude increases further with the sound panned to the right channel. From C to D the amplitude decreases by the same amount that had increased from B to C, and the the sound is panned to the left channel. From D to F the sound is panned to the right channel and the amplitude increases from D to E and decreases from E to F by the same amount. From F to G the the amplitude increases and the sound is panned to the left channel and from G to H decreases by the same amount with the sound panned to the right channel. From H to I the amplitude decreases with the sound panned to the centre, and from I to A the amplitude remains the same with the sound panned to the left channel.

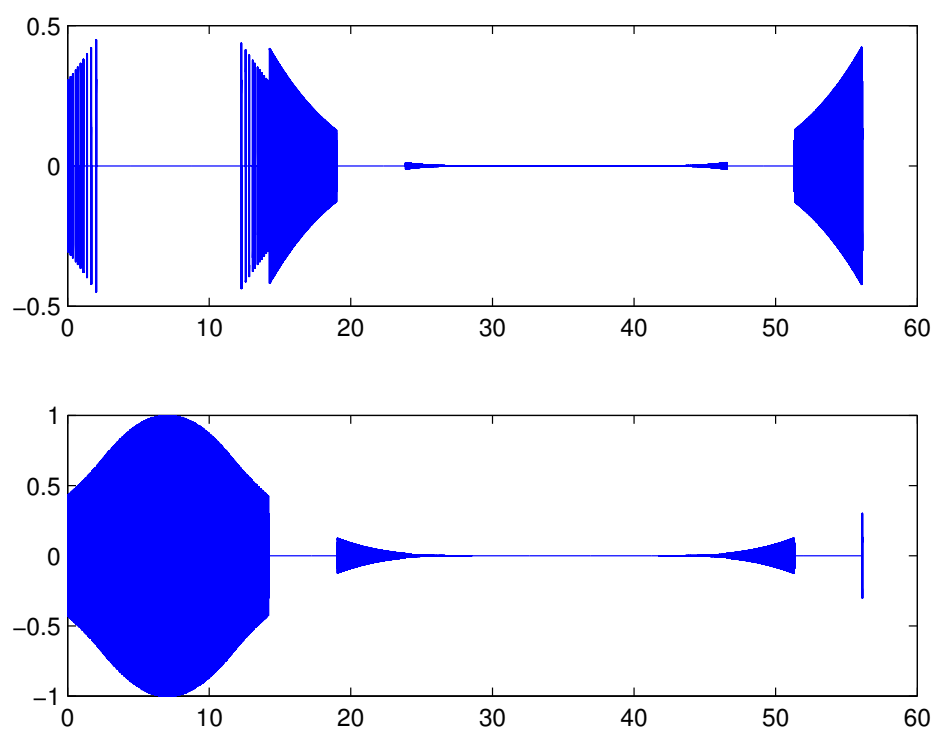
Intuitively speaking, this scheme will represent better in the audio domain symmetries that exist or not in the shape.

## 7.4 Conclusions

In this chapter we have presented a new technique for sonifying the shape of an object in an image and a one to one relationship is established between the amplitude envelope or the instantaneous frequency of a signal and its spatial position in the stereo field versus the visual shape of the object. Instead of using only pitch and panning as mapping parameters, the intensity of a sound is perceptually correlated with the motion along a curve in an image as well, so extracting an amplitude envelope from datasets that belong to the curve is meaningful for a shape recognition task. The datasets not only consist of the x and y coordinate values but also from their inter relationship as described by the value of their derivatives. Experiments should be done in order to test the usability of our proposed scheme and lead to more perceptually meaningful scalings in terms of pitch and amplitude.

## 7. SONIFICATION OF SHAPE

---



**Figure 7.6:** A stereo wave file using amplitude and panning as mapping parameters, as described in paragraph 7.3.1



## Chapter 8

# Creative Aspects of Image Sonification

In this chapter we emphasize on the creative aspect of image sonification and how it can be effectively used as a sound design tool. The most commonly encountered problems in the field, are addressed by using perceptually meaningful mappings, so that the properties of an image are directly reflected to the audio domain. We start by expanding the method presented in chapter 7 which is used to sonify the shape of an object, and later on we address the problems of dealing with color. Gray level images are sonified in a way which could be classified as a *non-standard synthesis*, paying also attention to the importance of the *scan path*. For color sonification, the HSV (Hue, Saturation, Value) colorspace is used and finally, we associate texture-sonification of colored images, with additive synthesis.

An image is far more than a simple collection of pixel values, as sound is not just a collection of sample amplitude levels. Mapping the parameters from one domain to the other is only limited by our imagination, but in this approach we are after perceptually meaningful associations that would lead to predictable sound results, therefore simplicity in the design plays a key role. We do not attempt to set the right boundaries in terms of scaling the data, but rather describing effective ways to organize them, emphasizing on the creative aspect of image sonification.

### 8.1 Related Work

In [39] P. Meijer sonifies a grey scale [64, 64] image. Each pixel is represented by a sinusoidal oscillator whose amplitude is proportional to the gray level, while the frequency depends on the vertical position of the pixel. The image is scanned from left to right, column by column and the sound generated by each column, is the sum of its oscillator outputs. A sonification of stacked images is presented in [27] using a color coding strategy. The intensities of an RGB triplet represent the pitch of an instrument on a diatonic scale and the rhythm is based on a set of markers placed on the stacked image. The resulting sound can be a rhythmic pattern, a harmonic representation or by combining both, a melody. In [55] the author tries to make a qualitative sonification of portrait pictures by mapping distance (for example the distance between the eyes) to MIDI frequency and the size of the objects to tempo.

Considerations regarding the scan-path versus the spot of a glance can be found in [21], where the author pays attention to the streaming of audio data, with respect to what we are supposed to see. Though scan-path theory suggests a “top-down internal cognitive model”, in [35] is suggested that the scan should be done from left to right, since we are used to read this way, and from bottom to top, because bottom presents the foreground and top the background of a picture. By this way “the user gets the impression that he is located at the edge of the scan-line”. Another viewpoint on time versus scan-path is presented in [60]. A ‘pointer’ is defined as “a data element, or a set of data elements, which is mapped to auditory domain at the same time.” ‘Paths’ include a number of ‘pointers’ over time. A ‘path’ can be a straight line, a set of distributed points, or arbitrary curves which span the image. If time is fixed during the sonification process, it is called ‘scanning’, else if the user alters the time stamp during the sonification, it is called ‘probing’. Two different time mapping schemes are also presented. In the first one, the horizontal and vertical position of the pixel define the pitch, while the brightness of the pixel is associated with its duration. In the second approach, a pointer has a rectangular shape with a time-varying size, increasing or decreasing while moving along a perpendicular path. By this way, the screen provides a new dimension to which “any sonic properties, including

pointer paths, can be assigned”. In [61], *Raster mapping* is proposed for the sonification of texture. The image is raster scanned and each pixel corresponds to one audio sample by linearly mapping the range of gray level Values, to audio sample levels  $[-1, 1]$ . The authors claim that “Sonified sounds preserve the feeling of the original images quite similarly in the auditory domain”.

Many color sonification techniques make use of the HSV or the HSL color model. Most mappings associate Hue with Pitch and Lightness or Value with Amplitude. In [22] Lightness is mapped to Pitch and Colorfulness (i.e Saturation) to Loudness. A mapping based on Subtractive Synthesis can be found in [56]. White noise passes through a lowpass filter whose cut-off frequency depends on Brightness, representing the grey levels, while color is added by passing the same signal through 12 parallel resonant filters spaced at octaves apart, whose frequencies depend on Hue. Some other, more arbitrary mapping choices, can be found in [37, 9].

## 8.2 Tracing and Reshaping the Shape of an Object

While most image to sound approaches map the spatial coordinates that describe the shape of an object to time (horizontal axis) and frequency (vertical axis), we will use the method described in chapter 7, because it offers a unique way to “read” the shape and easily manipulate the spatial datasets, even if the curve is closed. Since we do not use panning as a mapping parameter in the current approach, we need to expand the algorithm which describes a new curve in the time domain based on the traced one, in order to have a distinction between curves that evolve vertically, from those that evolve diagonally. As an horizontal evolving curve implies a steady state, a vertical one should imply an abrupt change. The construction of the new curve is based on the following rules:

$$T[1] = c, \quad \text{where } c \text{ is a constant} \quad (8.1)$$

$$T[k] = T[k - 1], \quad \text{if } y'[n] = 0 \quad (8.2)$$

$$T[k] = T[k - 1] + \sum_{x'[n]=0} y'[n], \quad \text{if } x'[n] = 0 \quad (8.3)$$

$$T[k] = T[k - 1] + x'[n]y'[n], \quad \text{if } x'[n] \neq 0 \text{ and } y'[n] \neq 0 \quad (8.4)$$

### 8.2.1 Shape as an Amplitude Envelope

The new shape defined above, can be used as an amplitude envelope after being linearly scaled or according to a power law to the range of  $[0, 1]$ , though in some cases it might be undesirable to reach zero. In general, the resulting audio envelope corresponds better to its visual interpretation if the values are scaled according to a power law, such as the  $x^e$ , before being normalized. The duration of the envelope can be of any length, by assigning a constant duration to the points that make up the curve.

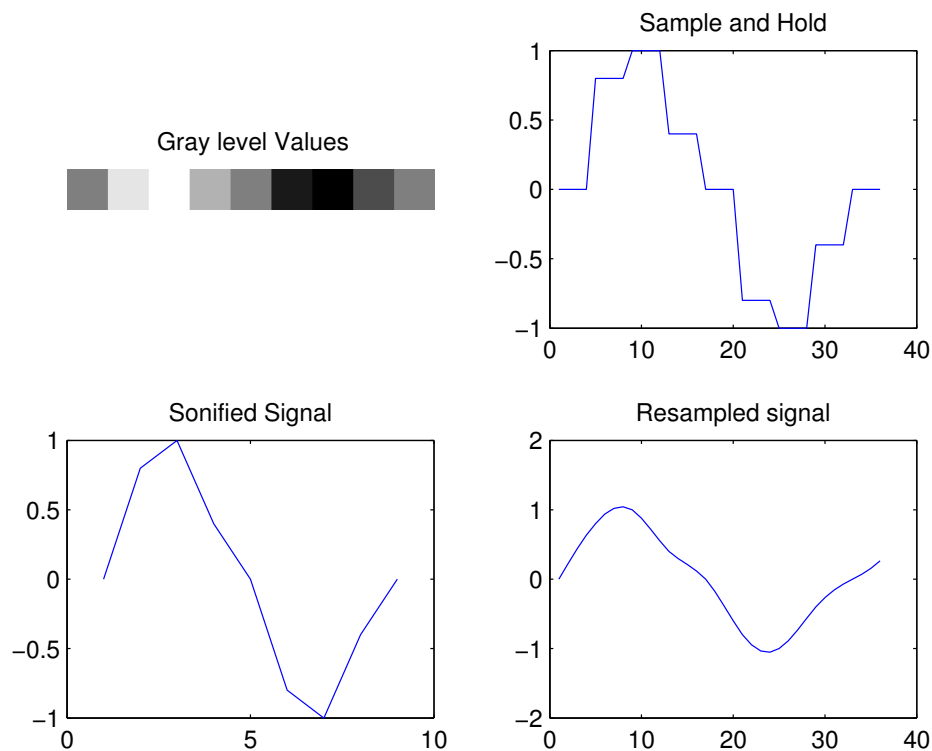
### 8.2.2 Shape as a Wavetable

Though there are many software platforms that allow the user to draw by hand a waveshape, it might be desirable to audify shapes that are part of a scientific result or are designed very accurately, with the aid of mathematical functions, or specialized drawing/cad programs. In this case, the spatial datasets of the reshaped curve are linearly scaled to the range of  $[-1, 1]$ , with each tracked pixel representing one audio sample. The result is a wavetable, whose length depends on the tracked pixels but if desirable, resampling the data at higher rates with interpolation is an option, in order to increase the wavetable's length, making it possible to scan it in different frequencies while maintaining the original quality of the pixel-form.

## 8.3 Gray Level Images

Each pixel can be represented as a single sample in the audio domain, by linearly scaling the pixel Values which range from  $[0, 1]$  to the audio sample levels ranging from  $[-1, 1]$  as done in [?]. This way we have a description of the oscillating

points around middle gray, but a normalization process should also be taken into consideration, by finding the minimum and maximum pixel Values present in the image, before scaling them to audio levels. Stretching the original waveform is also an option, either as described in 8.2.2 or by ‘Sample and Hold’ the Value of a pixel by a desired factor as seen in Figure 8.1.



**Figure 8.1:** Gray level values mapped to audio sample levels.

The process of sonifying gray level images could be described as a *non-standard synthesis* method [8], in which the “amplitude values” are directly derived from the pixel gray level Values, and “time values” are derived from the patterns that arise by using a particular scan path. Since the scanning process is directly related with the sound output, we are after a unique mapping that relates the 2-Dimensional image with the 1-Dimensional audio result and is able to transmit the original feeling of the image in terms of texture to sound, by preserving

## 8. CREATIVE ASPECTS OF IMAGE SONIFICATION

---

neighboring pixel Value relations. One way to achieve that, is by scanning the image with the *Hilbert Space Filling Curve* (Figure 8.2), as it fills an area in a self similar way without repeating itself. It can be encoded with the initial string  $L$  and the following string rewriting rules:

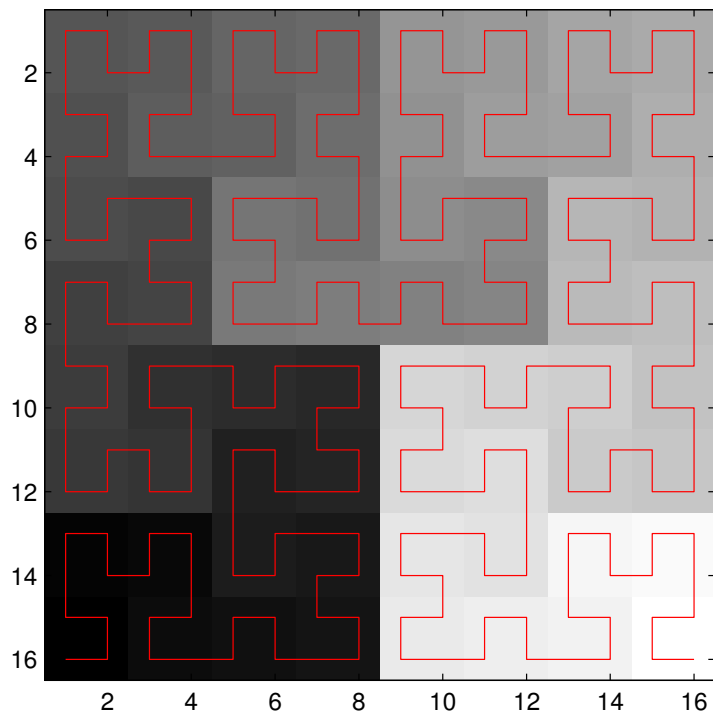
$$\begin{aligned}L &\rightarrow +RF - LFL - FR+ \\R &\rightarrow -LF + RFR + FL- \\F &\rightarrow \text{go one pixel forward} \\+ &\rightarrow \text{turn right} \\- &\rightarrow \text{turn left}\end{aligned}$$

A Hilbert curve of order  $n$  is capable of scanning an image with a maximum size of  $[2^n, 2^n]$ , so it can be applied to images of arbitrary sizes as well, with points of the curve that lie outside of the image being simply ignored.

### 8.4 Colored Images

Color sonification is directly dependent on the color space that is used to describe a particular color. Since painters think more in terms of the HSV color space when building up their color palette, maybe this could be the most ideal color space for image sonification, because it offers a very predictable way to construct a sound palette. HSV is widely used in computer vision techniques, mainly because it can be easily segmented and creatively manipulated as in [?] for example, where is used for automatic color harmonization. Furthermore, its relationship with sound apart from being constructive can also be descriptive, considering as an example S. Barrass' Timbre Brightness Pitch model (TBP) [6] in which the geometry of the model is derived from the HSL color space.

However, one major limitation of using this color space should be mentioned, which will also be inherited in the sound design process. The Euclidean distance between two points in the Hue plane is not perceptually uniform, considering as an example the green color which spans a wide range of degrees, when compared to yellow which spans only a few.



**Figure 8.2:** A Hilbert curve of order 4 filling a  $[16, 16]$  image. The image could be roughly segmented in 4 regions, each one describing the order of the curve.

### 8.4.1 Color Sonification

Since we are using the HSV color space we need to find the perceptually meaningful counterparts of Hue Saturation and Value in the audio domain. Saturation is a measure of purity of a given color, and could be associated with "how much" a given color, stands out of a given zero point. Therefore it can be mapped to audio amplitude levels by using the decibel scale, because this way the linearity of Saturation levels could be reflected on linearly perceived amplitude changes. The range of  $[0, 1]$  is scaled linearly to dBFS values ranging from a user specified minimum to a maximum. The use of thresholding functions can be of great importance, because a direct mapping from the inaudible -90 dBFS to the maximum of 0 dBFS could be problematic for sound design purposes. Instead, a thresholding function can be used which reduces Saturation values below a limit to the inaudible audio level, and for values above this limit, a scaling can be applied. For demonstration purposes we used a threshold of 0.05 and used a range of  $[-50, 0]$ dBFS.

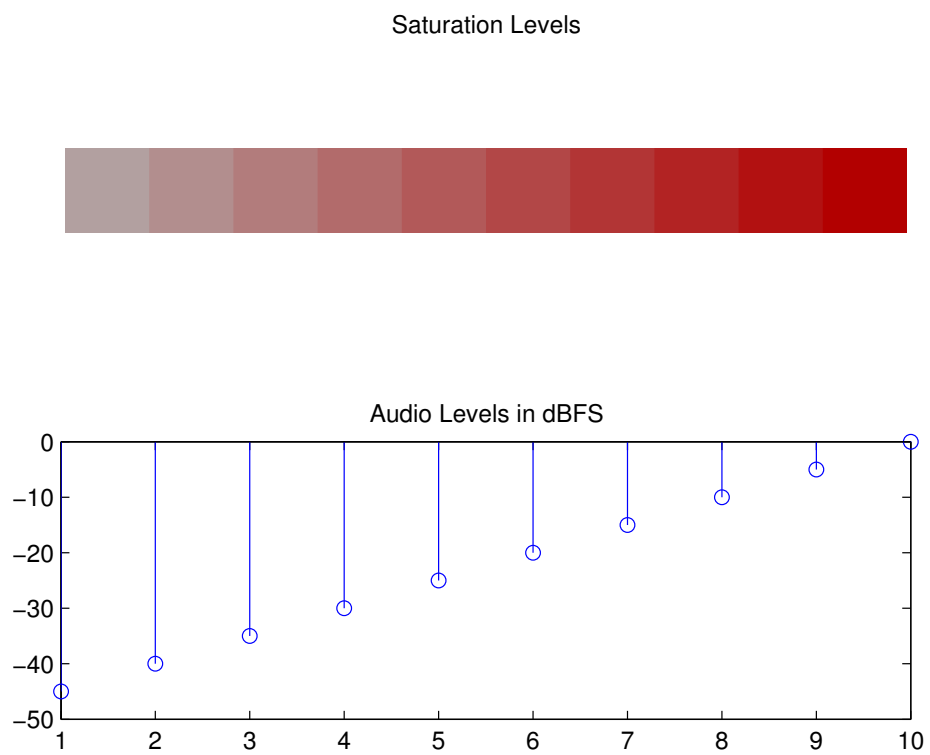
Associating pitch circularity with Hue circularity leads to the possibility that a pixel's Value could be regarded as a transposition factor, acting upon the Hue values. In the presented scheme, the range of Value is divided in 10 equal regions each one setting a base frequency in octave bands, starting from 20 Hz up to 10.240 Khz, therefore a change in region represents octave transpositions in frequency, as demonstrated in Fig. 4.

$$b_n = 20 \cdot 2^n, \quad n = 0 \dots 9 \quad (8.5)$$

Values above region 9 (5.12 Khz) are not perceptually meaningful, because the sense of pitch is weakened as we move towards higher frequencies, but it may be practically useful in sound design, since the upper region of the frequency spectrum contributes largely to the perceived quality of a sound.

The frequency values which span each octave, are defined by Hue which is mapped to cents. The 360 degrees Hue plane, is linearly scaled to the range of  $[0, 1199]$  which represent cents. This way we get a perceptually meaningful mapping from Hue to frequency, since the distance between 30 degrees is represented by a 100 cents increment (i.e a semitone), and 1 degree increment represents

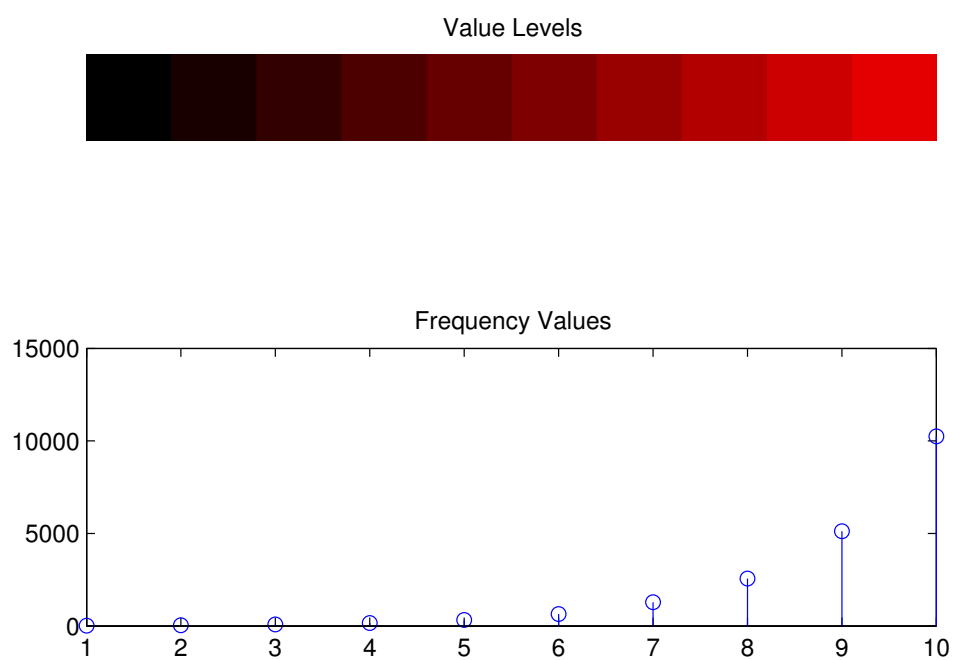




**Figure 8.3:** 10 saturation levels ranging from 0.1 to 1, mapped to amplitude dBFS units.

## 8. CREATIVE ASPECTS OF IMAGE SONIFICATION

---



**Figure 8.4:** 10 value levels ranging from 0.1 to 1, mapped to octave ranges.

about a 3.3 cents increment in frequency, well within the Just Noticeable Difference (JND) of pitch, but on the other hand the resolution of a pixel's Value is heavily compressed. By using this mapping the corresponding frequency of a pixel is given by equation (6) where  $c$  is the value in cents and  $b_n$  is described by equation (5).

$$f = b_n \cdot 2^{c/1200} \quad (8.6)$$



**Figure 8.5:** A major scale. All squares have a Value of 0.5 which sets the base frequency at 320 Hz except the last one which has a value of 0.6 setting the base frequency at 640 Hz. The Hue values in degrees are [0 60 120 150 210 270 330 0] which correspond in frequencies of [320.0000 359.2688 403.3564 427.3894 479.8364 538.7195 604.8284 640.0000] Hz

Of course, the scalings can be altered according to the image's quality parameters that we want to emphasize. If for example the variation is greater in Value rather than Hue, pixel Values can be mapped to smaller regions than octaves and the Hue values would fill the frequency space in between. This way a better frequency resolution is achieved but the perception of the sonified color in terms of Hue, is suppressed by the perception of Value.

### 8.4.2 Texture Sonification

The above presented color sonification scheme can be used for sonifying the texture of an image by using additive synthesis, which offers the possibility to move from pure tones, to highly complex sounds. The scanning process is from top to bottom as the eye scan path theory suggests, and all pixels which belong in the same line are sonified at the same time, a process which results in additive synthesis. A bank of sine oscillators is used, all having the same phase offset,

## 8. CREATIVE ASPECTS OF IMAGE SONIFICATION

---

whose number is defined by the width (rows) of the image and their frequencies depend on Value and Hue of the corresponding pixels. All pixels have the same duration, which offers frequency versus time independence, but could lead to problems unless special care is taken for the phase advance in each wavetable, when moving from one line to the next. If the pointer's position that reads the wavetable is reset (i.e starting from the beginning of the wavetable) when moving from line to line, audible clicks arise resulting in a metronome like effect, which could be very useful if the goal is to convey information, but unwanted in sound design. One possible solution to this problem could be achieved, by reading the current pointer's position in the wavetable and transmit the phase information as we progressively scan the lines, so that we continue reading the wavetable from the point it stopped, instead of resetting it. By using this approach, the audible amplitude discontinuities will only be a result of varying Saturation between successive pixels. What we essentially end up with, is a bank of oscillators, each one having its own instantaneous frequency and amplitude envelope. For an image of size  $[M, N]$  we have:

$$s(t) = \sum_{j=1}^N \frac{a_{ij}}{N} \sin(\phi_j(t)) \quad (8.7)$$

Where  $a_{ij}$  is the instantaneous amplitude of the  $j^{th}$  oscillator derived from the Saturation of each pixel, as described in section 5.1 and converted from the decibel scale to amplitude values.  $\phi_j$  is the instantaneous phase of the  $j^{th}$  oscillator:

$$\phi_j(t) = 2\pi \int_0^t f_j(\tau) d\tau \quad (8.8)$$

Where  $f_j$  is the instantaneous frequency of the  $j^{th}$  oscillator as described by equation (6). The scanning process is from top to bottom therefore  $t$  satisfies:

$$t + (i - 1)T \leq t < t + iT, \quad i = 1 \dots M \quad (8.9)$$

$T$  is a time constant which controls the speed of the scanning process. By experimenting with different scanning speeds, if the spectral content of the image is suitable, we can move between harmonic or inharmonic sounds (longer durations) to noisy ones (shorter durations).

### 8.5 Conclusions

We have addressed the most common problems found in image sonification which are dealing with sonification of shape, color and texture, and by using perceptually meaningful mappings, the properties of an image are directly reflected to the audio domain in a very predictable way. Using image sonification as a tool to aid sound design, can yield many interesting audio results that are hard to achieve by only using the existing audio based techniques. It gives rise to new sound manipulation approaches, since effects that were only applicable in the visual domain, may now have their audio counterpart. All the sounds which accompany this thesis were created using MATLAB [3], mainly for image feature extraction and mapping definitions, Csound [1] as the main sound engine and AC Toolbox [2] for generating the Csound score based on the data.

## 8. CREATIVE ASPECTS OF IMAGE SONIFICATION

---

# 9

## Appendix: Sound Results

Sounds 1 to 5 are generated from figure 7.4 with the methods presented in paragraph 7.3.1.

1: White Noise. No scaling prior to sonification. Number of pixels: 1123. pixel duration: 0.05 sec. Amplitude Levels in dBFS: [0.005, 1].

2: White Noise. Scaling prior to sonification with  $f(x) = x^e$ . Number of pixels: 1123. pixel duration: 0.05 sec. Amplitude Levels in dBFS: [0.005, 1].

3: Same as above but each pixel has a duration of 0.01sec.

4: 1 KHz sine wave. Scaling prior to sonification with  $f(x) = x^e$ . Number of pixels: 1123. pixel duration: 0.01 sec. Amplitude Levels in dBFS: [0.005, 1].

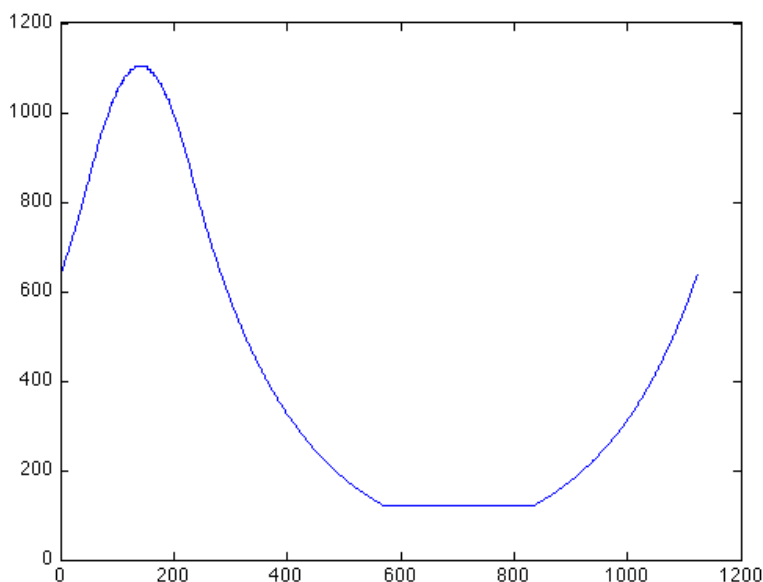
5: Sonification with music cents (Figure 9.1).

6: Sonification with music cents as described in paragraph 7.3.2.

7: Shape as an amplitude envelope (Figure 9.2) with the algorithm presented in paragraph 8.2. Sonification of figure 7.4. White Noise. Scaling prior to sonification with  $f(x) = x^e$ . Number of pixels: 1123. pixel duration: 0.01sec. Amplitude Levels in dBFS: [0.005, 1].

## 9. APPENDIX: SOUND RESULTS

---



**Figure 9.1:** Frequency envelope of sound (5).

8: Generating a wavetable from figure 7.4. Playback frequency at 80 Hz.

9: Same as above. Playback frequency at 160 Hz. (Figure 9.3)

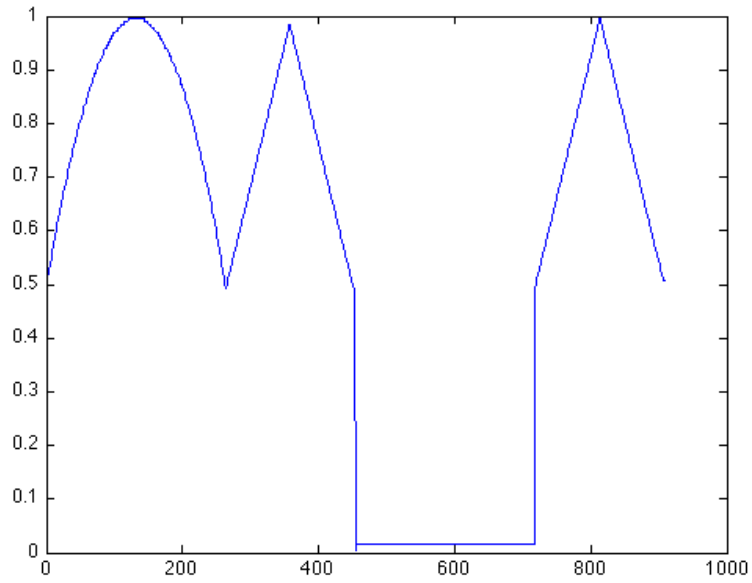
10: Scanning Figure 9.4 (a) with the Hilbert Space Filling Curve as described in paragraph 8.3. Sonified image dimensions: 512x512.

11: Scanning Figure 9.4 (b) with the Hilbert Space Filling Curve as described in paragraph 8.3. Sonified image dimensions: 512x512.

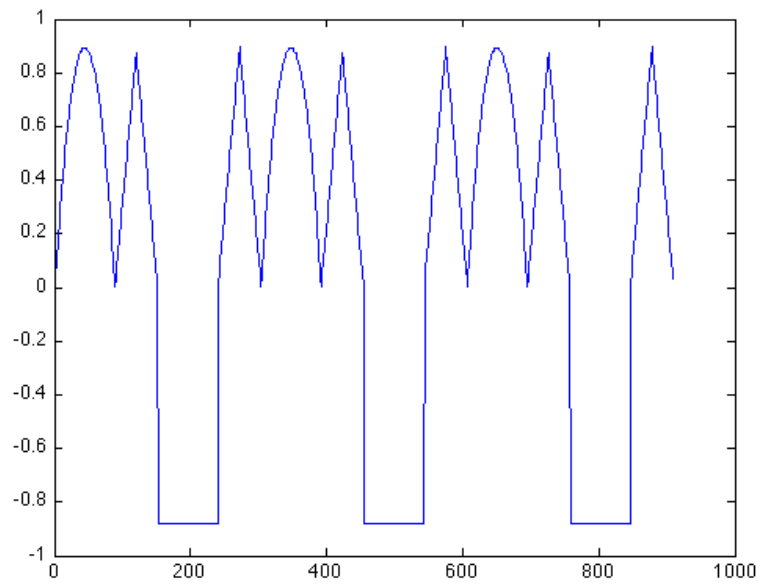
12: Scanning Figure 9.4 (c) with the Hilbert Space Filling Curve as described in paragraph 8.3. Sonified image dimensions: 512x512.

13: Scanning Figure 9.4 (c) with the Hilbert Space Filling Curve as described in paragraph 8.3. Sonified image dimensions: 2024x2024.





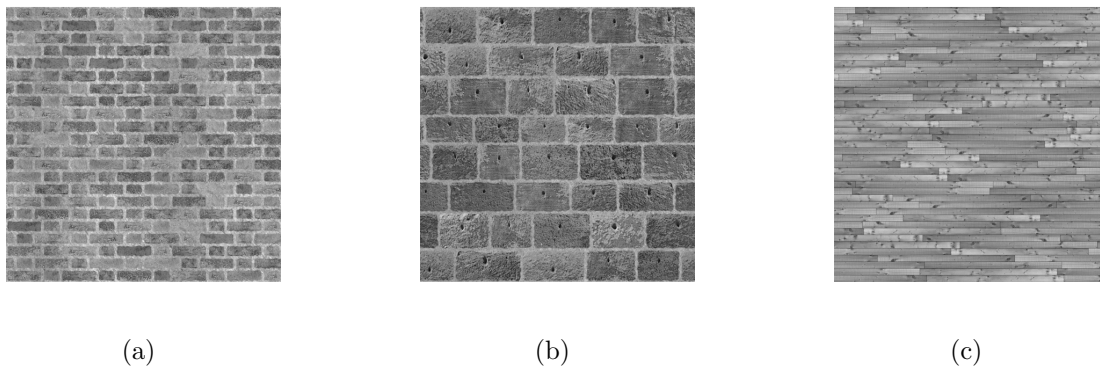
**Figure 9.2:** Amplitude envelope of sound (7).



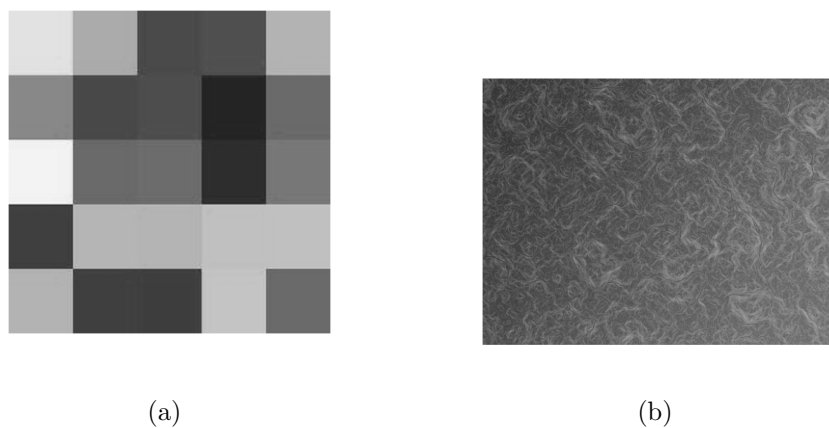
**Figure 9.3:** First few samples of sound (9).

## 9. APPENDIX: SOUND RESULTS

---



**Figure 9.4:** Sample images.



**Figure 9.5:** Sample images.

14: Scanning Figure 9.5 (a) with the Hilbert Space Filling Curve as described in paragraph 8.3. Sonified image dimensions: 450x450.

Sounds 15 to 19 are sonified with the same procedure but the images are scanned row by row instead of using the Hilbert Space Filling curves.

15: From figure 9.4 (a).

16: From figure 9.4 (b).

---

17: From figure 9.4 (c).

18: From figure 9.5 (b).

19: From figure 9.5 (a).

20: Saturation Levels. From figure 8.3

21: Octaves. From figure 8.4

22: Major Octave. From figure 8.5

The following sounds were created according to paragraph 8.4 and by using the mappings shown in table 9.1.

23: From figure 9.6 (a).

24: From figure 9.6 (b).

25: From figure 9.6 (c).

26: From figure 9.7 (a).

27: From figure 9.7 (b).

28: From figure 9.7 (c).

29: From figure 9.8

The following sounds were created by using the mappings shown in table 9.1 but now all pixels are sonified at the same time. Since the sound can have an infinite duration, we present only the first few seconds.

## 9. APPENDIX: SOUND RESULTS

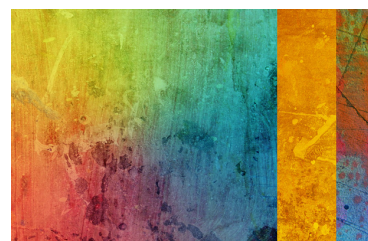
---



(a)



(b)

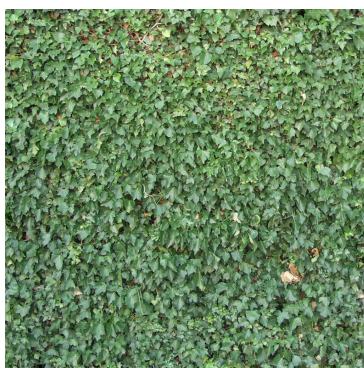


(c)

**Figure 9.6:** Sample images.



(a)



(b)



(c)

**Figure 9.7:** Sample images.



**Figure 9.8:** Sample Image.

30: From figure 9.6 (a).

31: From figure 9.6 (b).

32: From figure 9.6 (c).

33: From figure 9.7 (a).

34: From figure 9.7 (b).

35: From figure 9.7 (c).

## 9. APPENDIX: SOUND RESULTS

---

Data Feature		Sound Synthesis Parameters
Saturation [0, 1]	→	Amplitude [-1, 1]
Total Saturation of pixels in a row [-, -]	→	Amplitude [-1, 1]
Value [0, 0.3)	→	Base Frequency [10]
Value [0.3, 0.36)	→	Base Frequency [20]
Value [0.36, 0.42)	→	Base Frequency [40]
Value [0.42, 0.48)	→	Base Frequency [80]
Value [0.48, 0.54)	→	Base Frequency [160]
Value [0.54, 0.66)	→	Base Frequency [320]
Value [0.66, 0.72)	→	Base Frequency [640]
Value [0.72, 0.78)	→	Base Frequency [1280]
Value [0.78, 0.84)	→	Base Frequency [2560]
Value [0.84, 0.9)	→	Base Frequency [5120]
Value [0.9, 1]	→	Base Frequency [10240]
Hue [0, 30)	→	Cents [0, 100]
Hue [30, 60)	→	Cents [100, 200]
Hue [60, 90)	→	Cents [200, 300]
Hue [90, 120)	→	Cents [300, 400]
Hue [120, 150)	→	Cents [400, 500]
Hue [150, 180)	→	Cents [500, 600]
Hue [180, 210)	→	Cents [600, 700]
Hue [210, 240)	→	Cents [700, 800]
Hue [240, 270)	→	Cents [800, 900]
Hue [270, 300)	→	Cents [900, 1000]
Hue [300, 330)	→	Cents [1000, 1100]
Hue [330, 360)	→	Cents [1100, 1200]

**Table 9.1:** Used Mappings

# References

- [1] <http://www.csounds.com/>.
- [2] <http://www.koncon.nl/ACToolbox/>.
- [3] <http://www.mathworks.com/products/matlab/>.
- [4] S. BARRASS. **Auditory Information Design**. *PhD thesis*, Australian National University, 1997.
- [5] S. BARRASS. **Sonification design patterns**. *in Proc. of the 9th International Conference on Auditory Display (ICAD)*, Boston, MA, USA, July 6-9, 2003.
- [6] S. BARRASS. **A perceptual framework for the auditory display of scientific data**. *Transactions on Applied Perception*, vol.2, No. 4, pp. 389–402, October 2005.
- [7] D. BENSON. **Music: A Mathematical Offering**. available online:<http://www.maths.abdn.ac.uk/~bensondj/html/maths-music.html>, Department of Mathematical Sciences, University of Aberdeen, Scotland, UK, 2006.
- [8] P. BERG. **Composing sound structures with rules**. *Contemporary Music Review*, Vol. 28, No. 1, pp. 75–87, February 2009.
- [9] G. BOLOGNA, B. DEVILLE, AND T. PUN. **Pairing colored socks and following a red serpentine with sounds of musical instruments**. *in Proc. of the 14th International Conference on Auditory Display (ICAD)*, Paris, France, June 24-27, 2008.
- [10] R. BOULANGER, editor. **The Csound book. Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming**. The MIT press, Cambridge, Massachusetts, London, England, 2000.

## REFERENCES

---

- [11] T. BOVERMANN, T. HERMANN, AND H. RITTER. **The local heat exploration model for interactive sonification.** *in Proc. of the 11th International Conference on Auditory Display (ICAD)*, Limerick, Ireland, July 6-9, 2005.
- [12] A. S. BREGMAN. **Auditory Scene Analysis.** MIT Press, Cambridge, MA, 1990.
- [13] L. M. BROWN AND S. A. BREWSTER. **Drawing by Ear: Interpreting sonified line graphs.** *in Proc. of the 9th International Conference on Auditory Display (ICAD)*, Boston, MA, USA, July 6-9, 2003.
- [14] L. M. BROWN, S. A. BREWSTER, R. RAMLOLL, M. BURTON, AND B. RIEDEL. **Design guidelines for audio presentation of graphs and tables.** *in Proc. of the 9th International Conference on Auditory Display (ICAD)*, Boston, MA, USA, July 6-9, 2003.
- [15] D. COHEN-OR, O. SORKINE, R. GAL, T. LEYVAND, AND YING-QING XU. **Color harmonization.** *in Proc. of ACM SIGGRAPH 2006*, Volume 25 Issue 3, pp. 624–630, July 2006.
- [16] A. DE CAMPO. **Toward a data sonification design space map.** *in Proc. of the 13th International Conference on Auditory Display (ICAD)*, Montreal, Canada, 2007.
- [17] D. DEUTSCH. **The paradox of pitch circularity.** *Acoustics Today*, July, 8–15, 2010.
- [18] F. EL-AZM. **Sonification and augmented data sets in binary classification.** *Master thesis*, Institute of Informatics and Mathematical Modeling, Technical University of Denmark, 2005.
- [19] H. FASTL AND E. ZWICKER. **Psychoacoustics. Facts and Models.** Springer-Verlag Berlin Heidelberg, 2007.
- [20] D. GABOR. **Theory of communication.** *J. Inst. Elect. Eng*, vol. 93, no. 111, pp. 429-457, London, 1946.
- [21] G. EVREINOV. **Spotty: Imaging sonification based on spot-mapping and tonal volume.** *in Proc. of the 7th International Conference on Auditory Display (ICAD)*, Espoo, Finland, July 29-August 1, 2001.



## REFERENCES

---

- [22] K. GIANNAKIS AND M. SMITH. **Towards a theoretical framework for sound synthesis based on auditory-visual associations.** *in Proc. of the AISB 2000 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 2000.
- [23] T. HERMANN. **Sonification for Exploratory Data Analysis.** *PhD thesis*, Bielefeld University, Germany, 2002.
- [24] T. HERMANN. **Taxonomy and definitions for sonification and Auditory Display.** *in Proc. of the 14th International Conference on Auditory Display (ICAD)*, Paris, France, June 24-27, 2008.
- [25] T. HERMANN, A. HUNT, AND J. G NEUHOFF, editors. **The Sonification Handbook.** Addison Wesley Publishing Company, COST Office and Logos Verlag Berlin GmbH, 2011.
- [26] T. HERMANN, P. MEINICKE, AND H. RITTER. **Principal Curve Sonification.** *in Proc. of the 6th International Conference on Auditory Display (ICAD)*, Georgia Institute of Technology, Atlanta, Georgia, USA, ?April 2-5, 2000.
- [27] T. HERMANN, T. NATTKEMPER, W. SCHUBERT, AND H. RITTER. **Sonification of multi-channel image data.** *in Proc. of the Mathematical and Engineering Techniques in Medical and Biological Sciences (METMBS 2000)*, pp. 745–750, CSREA Press.
- [28] T. HERMANN AND H. RITTER. **Listen to your Data: Model-based sonification for data analysis.** *Advances in intelligent computing and multimedia systems*, p. 189194, Int. Inst. for Advanced Studies in System research and cybernetics, Baden-Baden, Germany, 1999.
- [29] K. JENSEN. **Timbre Models of Musical Sounds. From the model of one sound to the model of one instrument.** *PhD thesis*, Department of Computer Science, University of Copenhagen, 1999.
- [30] G. KRAMER, editor. **Auditory display: Sonification, Audification, and Auditory Interfaces.** Addison Wesley Publishing Company, Santa Fe Institute Studies in the Sciences of Complexity, Proceedings, Volume XVIII, 1994.
- [31] G. KRAMER. **Sonification report.** *NSF, Tech. Rep, available: <http://www.icad.org/node/400>*, March 1999.

## REFERENCES

---

- [32] A. KRISTJANSSON AND P. U. TSE. **Curvature discontinuities are cues for rapid shape analysis.** *Perception & Psychophysics*, vol. 63, no. 3, pp. 390–403, 2001.
- [33] H. LEVKOWITZ. **Color theory and modeling for computer graphics, visualization, and multimedia applications.** Kluwer Academic Publishers, 1997.
- [34] M. S. LIVINGSTONE AND D. H. HUBEL. **Anatomy and physiology of a color system in the primate visual cortex.** *The journal of Neuroscience, Vol. 4, No. 1, pp. 309–356*, Jan. 1984.
- [35] T. MAHLER, P. BAYERL, H. NEUMANN, AND M. WEBER. **Visual attention in auditory display.** in *Proc. of the International tutorial and research conference on Perception and Interactive Technologies*, 2006.
- [36] P. MARAGOS. **Computer Vision.** Lecture Notes, NTUA, Athens, 2002.
- [37] D. MARGOUNAKIS AND D. POLITIS. **Converting Images to Music using their Colour Properties.** in *Proc. of the 12th International Conference on Auditory Display (ICAD)*, London, UK, June 20-23, 2006.
- [38] A. C. G. MARTINS AND R. M. RANGAYYAN. **Experimental evaluation of auditory display and sonification of textured Images.** in *Proc. of the 4th International Conference on Auditory Display (ICAD)*, Palo Alto, CA, November 2-5, 1997.
- [39] P. B. L. MEIJER. **An Experimental System for Auditory Image Representations.** *IEEE Transactions on Biomedical Engineering*, vol. 29, No. 2, pp. 112–121, February 1992.
- [40] B. C. J. MOORE. **An Introduction to the Psychology of Hearing.** *Elsevier Academic Press, London, UK*, 2004.
- [41] G. M. MURCH. **Visual and Auditory perception.** The Bobbs Merrill Company, Inc. USA, 1973.
- [42] M. A. NEES AND B. N. WALKER. **Encoding and representation of information in auditory graphs: Descriptive reports of listener strategies for understanding data.** in *Proc. of the 14th International Conference on Auditory Display (ICAD)*, Paris, France, June 24-27, 2008.

## REFERENCES

---

- [43] A. PIRHONEN AND H. PALOMAKI. **Sonification of directional and emotional content: Description of design challenges.** *in Proc. of the 14th International Conference on Auditory Display (ICAD)*, Paris, France June 24 - 27, 2008.
- [44] J. PONCE AND D. FORSYTH. **Computer Vision: A Modern Approach.** Prentice Hall, 2002.
- [45] R. PRADHAN, S. KUMAR, R. AGARWAL, M. P. PRADHAN, AND M. K. GHOSE. **Contour line tracing algorithm for digital topographic maps.** *International Journal of Image Processing (IJIP)*, vol. 4, no. 2, pp. 156–163, 2010.
- [46] M. PUCKETE. **The Theory and Technique of Electronic Music.** World Scientific Publishing, 2007.
- [47] C. RAMAKRISHNAN AND S. GREENWOOD. **Entropy sonification.** *in Proc. of the 15th International Conference on Auditory Display (ICAD)*, Copenhagen, Denmark May 18 - 22, 2009.
- [48] C. ROADS. **Microsound.** The MIT press, Cambridge, Massachusetts, London, England, 2001.
- [49] J. ROHRHUBER. **Sonification variables.** *in Proc. of the in Proceedings of the Supercollider Symposium*, 2010.
- [50] J. SANCHEZ. **Identifying and communicating 2D shapes using auditory feedback.** *in Proc. of the 16th International Conference on Auditory Display (ICAD)*, Sydney, Australia, July 6-9, 2004.
- [51] S. SHELLEY AND M. ALONSO. **On the use of sound for representing geometrical information of virtual objects.** *in Proc. of the 14th International Conference on Auditory Display (ICAD)*, Paris, France June 24 - 27, 2008.
- [52] S. K. SHEVELL, editor. **The science of color.** Elsevier, 2003.
- [53] B. L. STURM. **Synthesis and Algorithmic Composition Techniques Derived from Particle Physics.** *in proc. of the 8th Biennial Arts Tech. Symp*, Connecticut College, New London, CT, 2001.
- [54] T. TOLONEN, V. VLIMKI, AND M. KARJALEINEN. **Evaluation of Modern Sound Synthesis Methods.** Technical report, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of

## REFERENCES

---

- Acoustics and Audio Signal Processing, Report 48, ISBN: 951-22-4012-2, March 1998.
- [55] S. TORPEY, O. CURRAN, AND M. SCHUKAT. **Qualitative sonification of portrait pictures.** *Transactions on Engineering, Computing and Technology*, pp. 342–345, December 2004 ISSN 1305-5313.
- [56] KEES VAN DEN DOEL. **SoundView: Sensing color images by kinesthetic audio.** *in Proc. of the 9th International Conference on Auditory Display (ICAD)*, Boston, MA, USA, July 6-9, 2003.
- [57] KEES VAN DEN DOEL, D. SMILEK, A. BODNAN, C. CHITA, R. CORBETT, D. NEKRASOVSKI, AND J. MCGRENERE. **Geometric shape detection with SoundView.** *in Proc. of the 10th International Conference on Auditory Display (ICAD)*, Sydney, Australia, July 6-9, 2004.
- [58] K. VOGT. **Sonification of Simulations in Computational Physics.** *PhD thesis*, Institute for Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria. Institute for Physics, Department of Theoretical Physics, University of Graz, Austria, 2010.
- [59] J. WILLIAMSON AND R. MURRAY-SMITH. **Granular synthesis for display of time-varying probability densities.** *in Proc. of the International Workshop on Interactive Sonification*, Bielefeld, jan. 2004.
- [60] W. S. YEO AND J. BERGER. **Application of image sonification methods to music.** *in Proc. of the 11th International Conference on Auditory Display (ICAD)*, Limerick, Ireland, July 6-9, 2005.
- [61] W. S. YEO AND J. BERGER. **Raster Scanning: A new approach to image sonification, sound visualization, sound analysis and synthesis.** *in Proc. Digital Audio Effects (DAFx'06)*, Montreal, Canada, Sep. 2006, pp. 309–314.
- [62] T. YOSHIDA, K. M. KITANI, H. KOIKE, S. BELONGIE, AND K. SCHLEI. **EdgeSonic: Image feature sonification for the visually impaired.** *in Proc. of the The 2nd Augmented Human International Conference (AH 2011)*, Tokyo, Japan, March 12th - 14th, 2011.