Improving Visual-to-Auditory Cross-Modality Information Conversions

by Shern Shiou Tan

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy
November, 2018

Contents

Nomenclature

Acronyms / Abbreviations

ARM Advanced RISC Machine, a popular mobile processor

CMA-ES Covariance Matrix Adaptation Evolution Strategy

CMOS Complementary Metal–oxide Semiconductor

CNN Convolutional Neural Network

CPU Central Processing Unit

DoG Difference of Gaussians

EMD Earth Mover's Distance

EMD-KL Earth Mover's Distance and Kullback Leibler Divergence

fMRI Functional Magnetic Resonance Imaging

FOA Focus of Attention

GAN Generative Adversarial Networks

GC Garbage Collection

GMM Gaussian Mixture Model

GPU Graphics processing unit

GUI Graphical User Interface

HCI Human-computer Interaction

HCM Heuristic Colour Model

HRIR Head-related Impulse Response

HRTF Head-related Transfer Function

HSL Hue, Saturation, and Luminosity Colour Model

ICC International Color Consortium

IID Inter-image Distance

ISD Inter-sound Distance

JNI Java Native Interface

LoG Laplacian of Gaussian

LSTM Long Short Term Memory

LS-TTL Low-Power Schottky Transistor-Transistor Logic

MFCC Mel-frequency Cepstral Coefficients

MFC Mel-frequency Cepstrum

MIDI Musical Instrument Digital Interface

NDK Android Native Development Kit

NVDA NonVisual Desktop Access, a free 'screen reader'

OFAI Austrian Research Institute for Artificial Intelligence

OpenCV Open Source Computer Vision, a library for real-time image processing and computer vision tasks

OSC Open Sound Control, a network protocol commonly used for sound synthesizer and multimedia devices

OS Operating System

PCC Pearson Correlation Coefficient

Pd PureData, an open source visual programming language for sound generation

PET Positron Emission Tomography

PMD Photon Mixing Device

PM&R Physical Medicine and Rehabilitation

PSVA Prosthesis for Substitution of Vision by Audition

RAM Random Access Memory

RGB Red, Green, and Blue Colour Model

RNN Recurrent Neural Network

ROI Region of Interest

SCST Self-critical Sequence Training

SDK Software Development Kit

SSD Sensory Substitution Device

STK Synthesis Tool Kit, a set of open source audio signal processing library

SURF Speeded Up Robust Features

TCP Transmission Control Protocol

TDU Tactile Vision Substitution System

TED Technology, Entertainment, and Design Conference

TOF Time-of-Flight Camera

TVSS Tactile Vision Substitution System

USB Universal Serial Bus

VASS Visual-to-Auditory Sensory Substitution

VfW Video for Windows

VM Virtual Machine

VTSS Visual-to-Tactile Sensory Substitution

WHO World Health Organization

List of Publications

The following work was published/submitted for publication as a result of the investigations performed in the course of this thesis.

1. Shern Shiou Tan, Tomás Henrique Bode Maul, Neil Russell Mennie, and Peter Mitchell (2010). "Swiping with Luminophonics". In: *4th IEEE Cybernetics and Intelligent Systems (CIS)*. Singapore, pp. 52–57
2. Shern Shiou Tan, Tomás Henrique Bode Maul, and Neil Russell Mennie (Jan. 2013). "Measuring the performance of visual to auditory information conversion." In: *PloS one* 8.5, e63042
3. Shern Shiou Tan, Tomás Henrique Bode Maul, and Neil Russell Mennie (2015). "Luminophonics experiment: A user study on visual sensory substitution device". In: *PeerJ Preprints*

List of Figures

List of Tables

Abstract

Sensory substitution devices have been widely used as an assistive tool, mainly for the purpose of rehabilitation for people with disabilities. With the development of electronics and computing devices, the application of visual-to-auditory sensory substitution (VASS) is becoming widespread in sensory substitution devices for the visually impaired. These devices convert visual information from images into an auditory form, known as a soundscape, allowing listeners to visualize their surrounding by interpreting the audio representation they hear. Despite its potential benefits, the technology has not been gaining acceptance among the public because of its weaknesses, such as the interpretability of the soundscapes and the quality of the user experience. The aims of this study were to improve cross-modality conversions in areas that include interpretability, information preservation, and the generation of soundscapes that afford a better listening experience. The use of image processing methods for the purpose of visual feature extraction is demonstrated in order to help the user to better interpret the soundscape they hear. By combining audio synthesis with the sounds of musical instruments and mapping colours to these sounds, systems that generate soundscapes that not only contain more information than that produced by traditional devices but also afford a more pleasant listening experience are created. Finally, a new evaluation and optimization methods are proposed to allow better visual-to-auditory feature mapping and foster a more up-to-date means of developing such devices. According to the experimental results and user feedback, the performance of VASS systems created using proposed techniques, in general, improves compared to the traditional systems in terms of ease of usage and user utility.

It is encouraging that in the future improved devices can be developed following the direction proposed in this research coupled with more up-to-date techniques, such as machine learning.

Acknowledgements

I owe a debt of gratitude to my supervisor, Dr Tomás Henrique Bode Maul, who was not only a great source of guidance for me when conducting the research but also has acted as a life coach, offering useful advice that helped me through difficult times. Your brilliance in identifying opportunities in my studies and your deep understanding in the field inspire me to pursue and contribute more to science. I am very grateful for your patience, especially during the final phase of my studies. I will always remember all the good qualities you showed me and will continue to spread the positive attitude you taught me to the people around me.

My next sincere thanks go to my wife. Without your encouragement and praise, I would not be the person I am today. The motivation you give me goes a long mile and has helped me considerably in pursuing such a challenging career. You are always the first to listen and believe in what I do. I am also very grateful to my father, mother, and brother. The support I receive from all of you is like a safety net for my life. You are always there standing ready to catch me with wide open arms if I should fall.

I would like to thank my co-supervisor, Dr Neil Russell Mennie, and also Dr Peter Mitchell. You both showed me the fascinating world of psychology and aided me in many activities during my studies. In addition, I am also grateful to the Faculty members who participated in my experiments.

I would like to express my gratitude to my former employers for their understanding when I was doing my research and working at the same time. My final thanks go to all my friends for accepting who I am.

Chapter 1

Introduction

1.1 Background Studies and Motivation

1.1.1 Sensory substitution

Sensory substitution refers to a process in which the characteristics of a sensory modality are restored by compensating them with those of a different sensory modality. It is normally performed on the basis of converting signals from input stimuli to different output stimuli. For example, Braille utilizes tiny palpable bumps called raised dots to encode text, essentially allowing blind users to read using their fingers, that is, using tactile stimuli. It is anticipated that, with the help of sensory substitution, some functionalities of a defective sensory modality can be transferred to another better functioning sensory modality so that the input of a defective sensory function can remain relevant to the individual who has suffered a sensory loss.

Sensory substitution has benefited humans in many respects, but in the area of physical medicine and rehabilitation (PM&R) it truly excels. Sensory substitution devices (SSD) are used as a tool for improving the standard quality of life of people who suffer sensory loss, thus helping them in the rehabilitation process. As compared to current invasive devices, such as neuroprosthetics, the SSD has the advantage that it is non-invasive, that is, its use requires no major surgery (Michael J Proulx and

Harder, 2008). It has been reported that when users receive appropriate training, they can use some SSDs as part of their daily life (Ella and Guendelman, 2012; Maidenbaum, Abboud, and Amedi, 2014).

According to Bach-y-Rita and W Kercel (2003), sensory substitution occurs across human sensory system. Currently, SSDs exist that perform cross-modality conversion, such as sight-to-touch, or conversion within the same sensory domain, such as touch-to-touch. Humans have been depending on the technology of sensory substitution in their daily activities. Some may use the technology without realizing it. Bach-y-Rita and W Kercel (2003) suggested that even the act of reading can be considered as the earliest sensory substitution technology that humans invented. This is because reading is not a natural ability for humans, but rather they are taught to understand visual representations of speech (in auditory form). Through writing, the author is able to communicate with the reader, although they are not together in the same place. Essentially, reading is an auditory-to-visual sensory substitution technique. Additional modern sensory substitution systems are available on different media and also with the help of electronics and computers. All of these are made possible by virtue of the flexibility of the brain that can be moulded through the processes of reprogramming, remapping, and reorganization. This amazing adaptive capacity of the central nervous system is called 'brain plasticity' or 'neuroplasticity'.

1.1.2 Brain Plasticity

The human brain is an amazing organ. It weighs only three pounds and yet it is the most complex material in the universe that humans have discovered thus far. Constructed of billions of neurons, the brain is the decision maker and central communication centre of the body. The neurons that form the core components of the brain typically consist of three parts: the soma, dendrites, and axon. Physically, a single neuron may appear simple but its power lies in its ability to communi-

cate with other neurons in milliseconds through an electrochemical process called synapse transmission (Seung, 2012). The interconnected network of neurons is called the neural network (or sometimes the neural pathway). In his book The Organization of Behavior: A Neuropsychological Theory, Donald Hebb introduced 'Hebbian Learning', a theory that explains how the neural network adapts through synapses and spikes during a learning process (Hebb, 1949). This mechanism of synaptic plasticity that reweights its synaptic strength (either strengthens or weakens) based on synapses describes how our brain learns and makes decisions.

As the field of neuroscience progressed, neuroscientists realized the potential of the brain in that it is able to transform itself throughout its lifetime. This phenomenon is called 'brain plasticity' and refers to long-term structural reorganization and functional changes in the central nervous system. Like a wax tablet, as illustrated by Plato, the brain is malleable through episodes of plastic change in various situations. These changes may be caused by both external factors, such as brain injury, and internal factors, such as neuronal rewiring and synaptic plasticity.

In an experiment, Merzenich et al. (1983) proved by monitoring the changes in adult owl and squirrel monkeys' brains pre- and post-surgery that the adult brain has the ability to rewire itself structurally. The experimental results showed that two to nine months after the surgery, consisting of a transection of the median nerve, the cortical area that matched the monkeys' hand median nerve area was completely occupied by new and expanded representations of the surrounding skin fields. Following the surgery, the brain of the adult monkeys was able to maintain the topographic representations of the skin surface, where most cortical sectors, such as the relationship between the receptive field size and magnification, were reorganized normally. This discovery suggested that, contrary to popular belief, nerves are capable of repairing themselves after being destroyed, even when the host is well past the age of adolescence. It also indicated that processes identical to the original developmental organizing processes are operational throughout the lifetime of a primate.

In recent years, advanced tools, such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have enabled researchers to discover more examples of situations where the brain is able to reorganize itself structurally and more importantly remap its functionalities (Seung, 2012; Thomas et al., 2007; Thulborn, Carpenter, and Just, 1999). A study conducted by Weiller et al. (1995) using PET showed that adult patients who recovered from aphasia after a left-hemisphere stroke demonstrated right-sided activation for language processing. Similar experiments were executed by Thulborn, Carpenter, and Just (1999) using fMRI instead of PET to study the shifting pattern of cognitive workload that occurred in patients who had regained a language function lost as a result of a stroke. The studies of such phenomena were enhanced by exploiting the ability of fMRI to measure and characterize the activity of a large-scale cortical network and to noninvasively monitor any changes in its organization. Although fMRI has good spatial resolution, it is limited in its temporal resolution, which is at best 1 s for each feedback. Because of the low temporal resolution, most fMRI experiments can be conducted only on brain processes having a duration of at least a few seconds, and thus most neuronal activities that occur within a 100 ms timeframe are excluded. Thulborn, Carpenter, and Just (1999) observed that the activities of undamaged components, such as contralateral homologs, increased in order to share the workload of a large-scale cortical network that was damaged by stroke. In addition, patterns of compensatory cortical activation, such as increased activation in areas immediately adjacent to the lesion, may exist in patients a long time after a stroke. These observations further reinforced the idea that brain plasticity allows a function to be shifted from a damaged region to a nearby region through long-term adaptations of the neural network.

Sensory substitution is deeply connected with the idea of brain plasticity and most of the recent successes in this area can be attributed mainly to the improvement in our understanding of the brain and human cognition. Exploiting the ability of brain plasticity, SSDs are designed as tools to facilitate cross-modality region

rewiring to shift the functional representation from a damaged region to another functioning region of the brain. It is anticipated that with the help of SSDs, the user can utilize other functioning brain regions to replicate some of the functionalities of the sensory loss. For example, with appropriate training and long-term usage of SSDs, users claim that they can make a mental representation of an object's shape and surface texture and the location of their surroundings similarly to a person with normal sight (Bach-Y-Rita et al., 1969; Ward and Meijer, 2010). In an ideal scenario, the use of SSDs causes a plastic reorganization inside the brain, where the functioning region adopts external representations without sacrificing its original functional presentation (Jenkins et al., 1990). This can be seen when a user uses an SSD, such as a visual-to-auditory SSD. While the user listens to an auditory representation of visual stimuli, his/her other hearing functions do not cease to operate, and sounds from other sources remain audible. According to Bach-y-Rita (1995), this is probably because the brain's plasticity allows it to unmask the secondary input to the primary visual cortex. Unmasking is one of the probable mechanisms featured in late brain plasticity. After a neural lesion occurs, the pre-existing neuronal connections that were concealed are uncovered. If the user is provided with appropriate training sessions in the use of an SSD and a rehabilitation program, the masked connections in his/her brain can be reactivated because of the increase in functional demand or even as a result of the motivation and willpower of the user (Bach-y-Rita, Danilov, et al., 2005). Studies conducted by Ptito et al. (2005) showed that, after a few sessions with a visual-to-tactile sensory substitution system called the BrainPort interface, a PET scan revealed that the activity in the visual cortex of a blind person becomes prominent.

However, without proper information management and training, brain plasticity may nevertheless be a double-edged sword (Maidenbaum, Abboud, and Amedi, 2014). On the one hand, brain plasticity can be utilized to restore some functionalities through activating compensatory capabilities, but on the other hand, it poses some risks as a result of the alteration of the original sensory function caused by the

effort to restore functions. This functional reorganization may cause unknown side effects that require further investigation, especially in cases where the performance of other tasks (e.g., memory tasks) and habits on which individuals have learned to rely that use the same brain region are potentially hampered (Röder and Rösler, 2003). In the process of rehabilitation, users may experience changes in their life to which they need to conform in order to retain the newly acquired skill.

Another failure of PM&R involving sensory substitution is attributed to the existence of 'critical periods' in early childhood when the brain remains very plastic and susceptible to the development of basic functions. In cases involving congenital sensory loss and failure of appropriate childhood learning, the brain is prevented from fostering functional specialization (Bedny et al., 2011; Sathian and Lacey, 2007; Gougoux et al., 2005). A classic example Hubel and Wiesel (1970) was demonstrated in cats, the visual system of which remained dysfunctional if they did not learn to see during the first few years of their life. It should be emphasized that the success rate of rehabilitation through brain plasticity varies between individuals. Not only the structure of the physical brain plays a role; the past experience and willpower of the person are also important.

Rehabilitation through brain plasticity remains a complicated task that must involve many external factors to be successful. Thus, the use of SSDs alone does not guarantee a complete recovery. Although it may be almost impossible to achieve maximum functional restoration, specific training procedures should be provided to improve the effectiveness of SSDs in promoting the activities of brain plasticity (Bach-y-Rita, 1990). In addition, a conducive environment and therapy too help the sufferer cope with the recovery process is required.

1.1.3 Visual Sensory Substitution

Visual sensory substitution can be defined as sensory substitution technology that focuses on cross-modality information conversion of visual signals. Because the abil-

ity to see is a dominant part of the sensory system, the loss of vision is a severe impairment for the individual. This may be one of the main reasons why the majority of sensory substitution technology is applied to visual systems to rehabilitate the visually impaired population. According to a report of World Health Organization (2014) (WHO) , a total of 39 million people suffer from blindness (as of the year 2014). Visual impairment frequently has a catastrophic effect on the sufferer, with inevitable indirect effects on their standard of living, their ability to support themselves, and their caregivers' life. Furthermore, it has a significant economic effect on society, especially in developing countries where most of the visually impaired population currently resides (World Health Organization, 2003). These facts are the main driving force of this research (i.e., that of Luminophonics), where the main aim is to further develop and improve the technology of visual-to-auditory sensory substitution (VASS).

The tactile and auditory modalities are the modalities of choice for the output representation of converted visual signals. In the early days of sensory substitution, tactile representation was frequently used in exchange for visual stimuli, but recently an increasing number of auditory representations have appeared with the help of computer audio synthesizers. One of the earlier examples of visual-to-tactile sensory substitution that utilizes tactile stimuli as the output representation and is currently used widely is Braille. Braille, a tactile phonetic reading and writing system for the visually impaired, is maybe the most popular visual sensory substitution system invented. Louis Braille invented the Braille system to help people with sight problems to read by tracing different bump patterns using their fingers. In addition to Braille, blind people also rely on mobility canes, another visual sensory substitution device, as a navigation guide that indicates the surrounding environment through tactile sensation and an audible sound when the cane hits an object. The user hears and feels the feedback from the cane when he/she uses it to tap on surrounding obstacles. A cane is a rare example of a device that can convert a single stimulus source (i.e., a visual signal) into multiple output representations (i.e., both

tactile and auditory signals).

Mobility canes together with guide dogs and the Braille system have become the de facto assistive technology for the visually impaired, particularly for the purpose of navigation and reading, respectively. Because of their reliability and familiarity, these tools are well accepted by people suffering from visual impairment. However, the conditions of the modern world currently require an advanced tool that is robust and suitable for a fast-paced lifestyle and not limited to only a single purpose, such as navigation or reading. The visual substitution systems created to suit this lifestyle were frequently unmethodical until 1969, when a modern type of SSD was introduced by an American neuroscientist called Paul Bach-y-Rita (Bach-Y-Rita et al., 1969). After years of performing research in the field of neuroplasticity, Bach-y-Rita conceived this type of device as an aid to help blind people acquire visual information about their surroundings through tactile sensations. His device was one of the earliest attempts in this particular field and it triggered a new wave of visual rehabilitation tools based on the concept of sensory substitution. As a result, many different devices and software based on the idea of Bach-y-Rita are now being developed and made available.

From the earlier VTSS devices, which used vibrating plates attached to a chair, such as that initially proposed by Bach-y-Rita, SSDs have developed into various forms. There exists a VTSS in the form of a wearable device resembling a vest in which an array of vibrotactile actuators is embedded (Novich and Eagleman, 2015) and an electrode array that is attached to the tongue, called BrainPort, was developed by Danilov and Tyler (2005). Although the conversion of visual information to tactile stimuli was popular at the beginning of the development of sensory substitution for visual rehabilitation, currently not only visual-to-tactile conversion is implemented in SSDs. One of the more popular approaches for visual sensory substitution that has been gaining the attention of researchers worldwide is VASS. It uses auditory stimuli instead of tactile stimuli as the medium of interpretation. As well as being more sophisticated, it offers a few advantages over VTSS that are

suitable for the visually impaired.

As mentioned previously, this research project, Luminophonics, is focused only on VASS. Briefly, similarly to VTSS, VASS transforms visual information into an auditory representation (soundscape), primarily for the visually impaired. A VASS device that has attracted considerable public attention because of its practicality is called vOICe (Meijer, 1992). It was developed by Peter Meijer in 1992 as an experimental system for translating live camera images into sounds. Since then, it has become available in multiple forms, such as Android, Raspberry Pi, and Web applications. The device operates using as the input device a camera, which captures visual signals in greyscale and then converts them to auditory signals by manipulating the audio properties to match the pixel intensities of the input image frames. This framework, which consists of a camera as the input unit, a processing unit, and a speaker as the output unit, is common across most SSDs. A few other VASS devices exist, the design of which follows a similar framework (Maidenbaum, Abboud, and Amedi, 2014); some recent solutions include SonART proposed by Yeo, Berger, and Lee (2004), See ColOr proposed by Bologna, Deville, Pun, and Vinckenbosch (2007), EyeMusic proposed by Hanassy et al. (2013), and ETA proposed by Wong et al. (2000). Although they are identical in terms of their hardware design, these VASS devices differ in factors such as the implemented conversion technique, input-output features matching, form factors, and speed. Consequently, the performance of each system varies in terms of functionality, interpretability, information preservation, and usability. Each device has its advantages and disadvantages, as reported by researchers and inventors, and there exists no clear evidence of how they compare with each other .

The field of visual SSDs, whether VTSS, VASS, or other similar SSDs, is indeed growing slowly. Researchers worldwide continue to find means of improving visual SSDs. As the technology advances, the awareness of the existence of this technology is beginning to grow. With the help of coverage by mass media, such as news reports and articles on the Internet, people are beginning to become informed

about the benefits of SSDs for the purpose of rehabilitation. Recently, a number of sensory substitution researchers were invited to present talks in their field at popular public conferences, such as the Technology, Entertainment, and Design (TED) Conference . One of the more well-known talks on visual sensory substitution was given by Amir Amedi at TEDxJerusalem. Not long ago, with the help of its popularity gained at the TED conference, an auditory-to-tactile SSD, called VEST, designed by David Eagleman was successfully funded with the help of Kickstarter, a global online crowdfunding platform based in Brooklyn, New York (Eagleman, 2015). This constitutes further proof that the public is beginning to be receptive to the technology of sensory substitution, but more work is needed to further promote the public adoption of rehabilitation through sensory substitution. Following this trend, it is hopeful that visual sensory substitution can be widely adopted by blind people. Currently, only a small group of people with visual impairment are using the devices to facilitate their daily activities and most of the systems are reserved primarily for experiments in controlled settings. Our motivation is to help improve the performance of the VASS system so that in the future it can be a device that complements conventional assistive technology, such as the mobility cane.

1.2 Problem Statement

It is noteworthy that the World Health Organization (2014) reported on its Website that an estimated 285 million people are visually impaired, out of which 39 million are blind and 246 million categorized as having low-level vision. Owing to the decline in infectious diseases, the overall population of the visually impaired worldwide has decreased. However, work remains to be done to help and support them, most importantly in the area of rehabilitation. Through rehabilitation, the blind can be trained to acquire the relevant skills and capabilities that facilitate personal independence. Although technology has advanced, it can be seen that visually impaired people still rely on the tools that were created decades ago, such as

mobility canes and Braille. People with visual impairment can benefit from more up-to-date technologies, such as sensory substitution, in particular visual-to-auditory sensory substitution (VASS); however, these technologies are not widely adopted. Most VASS devices are operated only in controlled environments for the purpose of research activities. In brief, although VASS technology has the potential to improve the livelihood of the visually impaired through the process of rehabilitation, there remain deficiencies that prohibit the blind from using these devices in their daily life. Therefore, this research project is aimed to solve the problem by finding a means of improving visual-to-auditory cross-modality conversion in order to promote the adoption of VASS in the blind community.

In this section, possible problems that may affect the performance of VASS and thus lead to its slow public adoption are identified. Then, in Section 1.3, solutions and goals are formulated within the frame of this research to solve the problems through techniques that improve the conversion of visual-to-auditory cross modalities.

1. **Complicated interpretation**

   The soundscape produced by VASS is frequently confusing and not easily interpreted, especially by people who have not been trained in its use. The situation becomes more complicated when the conversion algorithm is attempting to encode a relatively large amount of visual information into an auditory representation. As compared to an auditory presentation, a 2D visual image can accommodate more varieties of information and its information content is relatively unpredictable. Humans perceive a considerable amount of information about their surrounding solely through their visual system, including an object's shape, location, depth, colour, and much more. Naturally, the human hearing system is not designed to perceive this information, because audio presentations and visual images serve different purposes and their structure and size differ. In other words, audio and images are not interchangeable in

nature, thus making it more difficult for a system to encode images into an audio representation so that the individual can 'hear' the images. Therefore, the conversion of an image to its corresponding auditory representation for the purpose of sensory substitution requires a complex procedure for encoding the visual information, which frequently results in an incongruous outcome, in this case a soundscape that sounds unnatural. In order to make sense of the visual information encoded within the soundscape, the listener must interpret the audio sounds and reconstruct them into the original visual form in his/her mind. Essentially, visual-to-auditory cross-modality conversion provides an additional form of communication channel where the VASS system is the encoder and the human's mind is the decoder. The complexity of the soundscape interpretation (decoding) depends on the amount of visual information being encoded in the soundscape. The interpretability and the amount of information being encoded are inversely exponentially proportional: the interpretability decreases as the amount of encoded information increases. As the quantity of information increases further, the complexity of the interpretability rises as a result of external factors, such as noise and the effect of cacophony. The problem of soundscape interpretability is a major one, which VASS researchers need to overcome or at least ameliorate. It would very considerably enhance the performance of a VASS device if the algorithm produced a soundscape that is easily interpreted or a scenario that is better represented.

2. **Information reduction during conversions**

   As mentioned above, the incompatibility of the structure and size of an auditory and a visual form seriously affects the procedure of visual-to-auditory conversion. Hence, it is necessary to map one property of the visual information to another property of the auditory information. This is a standard procedure in cross-modality conversion, in this case, the conversion from a visual to an auditory modality. It should be emphasized that one major drawback of this

type of conversion is the consequences of the information reduction that occurs during the conversion process. When designing cross-modality mapping, it is inevitable that certain properties of the source will be discarded for two major reasons. First, some information from the source cannot be represented completely in the target representation. The second reason is that the size of the target form may be limited and not able to contain all the information from the source. Unfortunately, the problem of information reduction within visual-to-auditory conversion is a result of both these factors. Thus, the conversion from the visual to auditory modality can be a demanding task if the source information is not selected appropriately. In many examples of the earlier generations of VASS devices, a considerable amount of important visual information was discarded. For instance, the inventor of vOICe chose not to include colour information in the encoding into the soundscape. Although an improvement in the interpretability of a soundscape is important, it is equally crucial to manage the information reduction during the conversion. A lack of relevant information will directly affect the performance of a VASS device negatively because the user receives a soundscape that only partially represents the visual form. For this reason, techniques need to be developed to optimize the visual-to-auditory conversion such that most of the relevant information is retained and the interpretability of the soundscape is enhanced.

3. **High learning barrier**

   Learnability was cited in multiple sources addressing sensory substitution research as one of the major hindrances to its use until the user becomes sufficiently acquainted with the sensory substitution device (SSD). In general, all SSDs require a certain training period before the user can use them correctly. Certainly, VASS devices are no exception. There is a strong likelihood that this problem is related to the brain's plasticity. As mentioned above, sensory substitution takes advantage of the plasticity of the brain, which allows it to

mould/rewire itself by empowering one of its regions to assume control of the functionality of another. It should be stressed that the brain does not rewire itself quickly and the time the process takes may vary from one individual to another. In order for a user to be able to understand the soundscape produced by a VASS device, the brain may need to rewire its auditory region to interpret the signals as visual information. However, the learning barrier of a VASS does not depend solely on the activity of the brain; it can also be affected by other factors. One of the factors that influence the learnability of an SSD is the procedures behind the cross-modality conversion. Thus far, each VASS algorithm presents a different learning barrier, and some systems are easier to learn and some more difficult. The elements that affect the learnability of a system include the quantity of information being encoded, the interpretability of the soundscape, and the number of different timbres used to represent the features of the image. In addition, the experience of individuals also significantly affects their learning process. To summarize, every VASS device has a serious learning barrier, which varies according to the individual using it and the conversion process. It is therefore essential to identify the underlying problem to reduce the learning barrier so that more people can adopt a VASS device with less effort.

1.3 Research Goals

The Luminophonics project was initiated to address the problems related to VASS with the aim of further developing and improving the technology in order to close the gap between research studies and practical visual rehabilitation implementations. The name Luminophonics is a wordplay combining ‘luminous’ (related to light as perceived by the eye) and ‘phonics’ (related to speech sounds), which essentially carries the meaning of translating visual signals into an interpretable auditory soundscape, such as speech. Although a few types of visual sensory substitution

systems exist, the research focused only on the conversion of visual information to an auditory representation. This decision was motivated by the advantages of VASS over its tactile counterpart. One of these advantages includes the future potential and flexibility that VASS can provide. As compared to visual tactile substitution systems (VTSSs), VASS has more possibilities and room in which to grow, because it requires only a camera and a speaker with which a regular smartphone is equipped, whereas VTSS needs specialized tactile output units, such as motorized actuators. Primarily, this is because the source of the power of a VASS device is the software, in contrast to a VTSS, which is mainly hardware-oriented. Moreover, a VASS system has a smaller footprint in terms of overall device size, and the devices' smaller overall size and cheaper manufacturing cost will facilitate their commercialization in the future. Finally, by focusing only on one type of output representation, more ideas can be generated during attempts to improve the performance of VASS devices.

The primary research question of this project is how can the performance of VASS be improved with better visual-to-auditory cross-modality algortihm? It is hopeful that with improved performance of VASS, the practicality of VASS device can be elevated such that they will be more suitable for the daily usage of people with visually impaired. Thus, the main goal of this research was to maximize the performance of cross-modality conversion from visual signals to auditory representation. Accordingly, the research scope is wide, but this research focused on the factors that would help promote the adoption of VASS devices. According to Maidenbaum, Abboud, and Amedi (2014), the reasons behind the low adoption rate of SSDs are two-fold. One is the problems related to sensory substitution itself, the performance and reliability of which do not meet the standards required for daily usage. The second reason, however, is related to the general limitations of visual rehabilitation itself that constrain the potential of sensory substitution. In general, VASS systems face a set of problems similar to that faced by most modern visual sensory substitution systems; however, a few problems exist that are unique to VASS systems. Because the research activities focused on improving the performance of

VASS devices, all the aims are organized to answer research questions related to the problems that affect the performance of a VASS system.

This research hoped to answer these sub-questions in order to arrive to the main goal.

- Can the listening experience of the soundscape be improved?
- How to convert more visual information into the soundscape?
- How to make the soundscape easier to be interpreted?
- Can we evaluate VASS fairly with automated measurement instead of user-based experiment?

1.3.1 Objectives

The objectives of this research were as follows.

1. **To improve the soundscape such that it sounds more natural to the user**

   Currently, most VASS systems manipulate audio frequency, associating it with visual properties to produce the soundscape. However, the results are unnatural and this has been shown to affect the pleasantness and usability of the SSD when the user is listening and even to induce mental fatigue after prolonged usage, especially in the case of high resolution image frames (Abboud et al., 2014). In some research studies, an attempt was made to improve the sound quality of the audio synthesizer (Cronly-Dillon, Persaud, and Gregory, 1999; Bologna, Deville, Pun, and Vinckenbosch, 2007; Abboud et al., 2014). However, this attempt introduced different problems, such as interpretation difficulties and sound cacophony in cases where the information was not transferred correctly from visual to auditory signals. Before implementing new idea of replacing the audio synthesizer, research is needed to solve these

problems so that important features are not sacrificed in favour of a more natural soundscape.

2. **To increase the amount of information that is retained without introducing the cacophony effect**

   Information loss occurs during the process of cross-modality information transfer from the visual to the auditory domain, because it is not possible to encode as much data in auditory signals as in visual signals. It is common for a VASS system to simplify the visual data before the process of conversion in which the audio properties are matched with the corresponding visual properties, in order to pack the data into the audio channel. For instance, earlier examples, such as vOICe, reduce coloured images to greyscale images before the conversion (Meijer, 1992), which significantly reduces the amount of input data, but also means that the output soundscape cannot represent colour information. However, because of the size of the information content of auditory signals, the effort to pack all the information of visual signals (which is larger) into auditory signals leads to the effect of cacophony. When the user experiences cacophony, his/her brain cannot interpret the soundscape because of the dissonance caused by a mixture of too many types of sound. In this research, several other options are explored, such as applying advanced image processing techniques to retain as much visual information as possible and discard other non-relevant noise to produce a soundscape that is interpretable and useful for the user.

3. **To explore means of enhancing the interpretability of the auditory representation**

   A soundscape characterized by a high level of interpretability requires the user to use less effort to interpret it. The soundscape becomes a visual mental map, which allows the user to better understand the auditory representation. It is important to apply a conversion algorithm that matches the correct visual

properties with the corresponding auditory properties in order to produce a highly interpretable soundscape. In addition, the characteristics of the soundscape must be suitable for the user's current situation. For instance, in a situation where the user needs to make a quick decision (e.g., during navigation), time is an important factor. Therefore, the soundscape produced for this scenario must be fast and concise as compared to slow and detailed, which is more suitable for a different scenario.

In addition, according to the results of studies by Brown, Macpherson, and Ward (2011), the soundscape's auditory characteristics produced by the SSD also play a part in influencing the user's perception, especially in the initial learning period. As for many other technologies, the learning period is crucial for enabling the user to fully understand the functionalities of the device. Moreover, sensory substitution relies on the plasticity of the brain in order to fully exploit the power of the technology. Hence, a good training module not only shortens the user's learning period but also teaches the user to interpret the soundscape correctly.

In this research study, the intention was to explore means of searching for a set of matching visual and audio properties that allow the normal user to interpret the soundscape easily. This is also crucial from the psychological aspect of users and their thought processes during soundscape interpretation, because such a set will determine an optimized visual to auditory translation configuration. The development of a training curriculum that helps promote the learnability of an SSD is also essential.

4. **To develop methods to evaluate a visual-to-auditory substitution system**

   The lack of a common evaluation framework for VASS systems is one of the reasons why it has not improved as much as anticipated, which has hindered its adoption among the public. Currently, sensory substitution research is

conducted in silo, where researchers develop their own solution together with an evaluation method of the performance of their prototype, which are then disseminated globally. This traditional path makes a comparative evaluation of the performance of SSDs difficult. In addition, experiments using human subjects are the solution for measuring performance that is frequently used, but they can be very expensive. The aim was to develop a common evaluation framework that can quickly evaluate an SSD as a precursor for a good VASS system. This framework must be transparent regardless of the SSD, comparing the output soundscape and input visual information to produce a quantitative measurement that is standard for all VASS systems. It is hopeful that with such a framework, VASS researchers can evaluate their SSDs without lengthy and expensive experiments and improve their solutions by fine-tuning the configuration based on the evaluation results.

1.3.2 Scopes

During the course of the research presented in this thesis, a few scopes were made. They are explicitly stated below.

1. This research does not aim to develop a VASS product immediately for everyday usage. There are many other complex design problems like human-computer interaction and feasibility studies that are best to be conducted by their domain experts. Instead, this research focused on improving the design of visual-to-auditory conversions algorithm.

2. There are different degrees of blindness, ranging from complete to partial blindness, and special situations, such as congenital blindness. Because VASS systems are intended to address the rehabilitation issue of the visually impaired, this research did not aim to target completely blind people, in particular people who never experienced vision. It is apparent that the severely visually impaired, such as those with no prior vision experience, and the partially blind

population require different treatment because of their contrasting perception and exposure to visual systems. Many more resources, such as psychological studies or even a completely new VASS system, are needed, in particular for people whose blindness is congenital because of their lack of prior vision experience. Therefore, this research is addressing those with minor visual impairment to acquire more knowledge that can be used to create a better system that is suitable also for the severely blind population.

3. The experiments in this research were all conducted in a controlled environment. For example, the navigation experiments were at an indoor room with supervision. where the situation could be controlled with suitable lighting conditions and less hindrance from noise. Furthermore, it is safer to perform initial experiments indoors so that the experimenter can protect the subjects from any unexpected accident and injury.

4. Due to budget constraint, the equipment and the software used in this research were all stated in this thesis. Although there are many high-performance hardware available, equipment like the camera was selected based on the criteria such that they are cheap and easily acquired. Software like audio synthesizers was all of the open-source in nature or freely available for educational purposes.

1.4 Contributions

The research presented in this thesis offers specific advancements in the following research fields.

1. **Swiping-based visual information pre-processing**
   Two out of the four major prototypes developed by Luminophonics are equipped with an algorithm that utilizes image processing techniques to process the input image frames prior to cross-modality conversion. The approach of applying computer image processing techniques inside the VASS system, as imple-

mented in a few of the most recently produced systems in order to produce more information from the input images, is beginning to be developed. However, the main contribution of our technique is the combination of a swiping mechanism and connected-component labelling for blob extraction. The main benefit of this technique is that it simplifies the entire soundscape by sonifying the blobs according to the location, thus creating a soundscape that is both quickly and easily understood.

2. **Inclusion of colour information in cross-modality conversions**
   Throughout the history of VASS, not many systems have included colour information in their elements of visual-to-auditory conversion. One of the main reasons for this is that it frequently makes the soundscape considerably more complex, which may lower the overall interpretability. Two new improvements are proposed in this thesis that are designed to prevent this reduction of interpretability. We propose a method to compute an optimum musical instrument set selection in which the timbre similarity is small so that the sound of each individual colour is easily distinguishable. The second improvement is the creation of a heuristic colour model (HCM) (see Section 3.2.3) to personalize the colour representation according to each user's perception.

3. **Use of a three-dimensional camera to capture depth information**
   With the advancement of camera technology, many types of camera have been invented that allow more information about a person's surroundings to be captured. In addition, a VASS system that uses solely the information provided by a 2D camera is not adequate for helping people navigate. Studies have shown that, in some circumstances, humans depend on depth cues from stereo vision during their navigation. Therefore, we are the first to propose using a time-of-flight (TOF) camera to capture an accurate depth map of the surroundings to provide depth cues as additional input information for our conversions algorithm. Furthermore, our depth implementation, whereby

more user control of the depth level selection during the course of soundscape sonification is allowed, is considered the first to implement with such way.

4. **Development of quantitative measurements to evaluate VASS through interpretability and information preservation**

   Because of the difficulty in comparing the performance of different VASS systems, we developed the basis of an evaluation framework that quickly gauges the important aspects of an SSD. In our initial work, the evaluation framework was used to measure the interpretability and information preservation of a VASS system. In the future, more features can be included in the framework as indicators of a good VASS system.

5. **Optimization of visual-to-auditory properties mapping**

   On the basis of the evaluation method developed previously, we propose a new means of optimizing visual-to-auditory properties mapping through the application of a genetic algorithm. The cost function of the optimization is based on the evaluation method, and therefore, the results produced can be optimized to provide the best interpretability and also information preservation. The benefit of this method is that it will facilitate future VASS development in that an excessive number of trial and error processes will not have to be executed to obtain the best conversion mapping.

1.5 Thesis Outline

In this thesis, the research activities are structured according to a broad to narrow approach as follows.

Chapter 1 First, the thesis presents an introduction to sensory substitution technology and a brief explanation of how cross-modality conversion operates in general and the relationship of these technologies with brain perception. This chapter also includes the motivation for improving VASS and the goals of this improvement.

Chapter 2 The research areas related to the problem framework of VASS are outlined in this chapter, including descriptions ranging from those of early visual SSDs to those of the state-of-the-art VASS systems. The advantages and disadvantages of the current research together with the areas of improvements are described in detail.

Chapter 3 Following a top-down approach, this chapter describes the process of designing VASS prototypes in order to further explain the current VASS system. The prototypes created include some new features, such as the incorporation of image processing techniques and instrument timbre for auditory representation.

Chapter 4 After the prototypes had been designed, their performance was tested through experiments using human subjects. The details of the experimental designs are presented in this chapter, together with an explanation of the experimental results.

Chapter 5 A set of quantitative measurements is introduced as a framework for comparing the performance of different VASS systems in terms of properties such as interpretability and information preservation. Using the quantitative

measurement, we introduce a technique for searching for an optimized VASS using a genetic algorithm.

Chapter 6 The overall results of the research are discussed and analysed in this chapter. The possible actions that can be taken to improve future research are also discussed .

Chapter 7 In this chapter, the future of VASS and some possible insights into the use of VASS in the future are discussed. The power of deep learning in shaping VASS technology is emphasized, especially through visual recognition, generating better soundscapes, and searching for the most optimized cross-modality mapping.

Chapter 8 The conclusions that were drawn are presented in this chapter, including an overall summary of the research activities and the contributions made towards advancing the field of sensory substitution.

Chapter 2

Related Research

2.1 Introduction

This chapter presents the related research on VASS concepts together with some of the more popular devices in the field. By virtue of research work worldwide, VASS systems have grown such that their current versions include more functions and are easier to interpret and more user friendly than the earlier versions, although the overall size of devices has been reduced considerably. These tremendous achievements can be attributed to two major factors that have contributed to the growth of VASS:

- Advances in electronics
- Better understanding in neuroscience.

2.1.1 Advancement in Electronics

Size is one of the advantages of VASS over VTSS. The small size of VASS is the result of its minimum hardware requirements, i.e., a camera, speaker, and a CPU, as compared to VTSS, which depends on an array of motorized actuators. Because of this advantage, researchers have been working to reduce the size of SSDs so that

they will be suitable for use as wearable devices that can be carried everywhere for daily activities.

At the beginning of VASS system development, the hardware components, such as the camera and processors, were rather large, which resulted in a device having a very large overall size. However, over the years, the size of these electronic components has been reduced to the point that the smallest VASS system currently existing is as small as a palm-sized mobile device. This can be attributed to the recent developments in electronics, especially in the mobile device industry. For example, the camera and audio speakers are packaged in a single integrated electronic circuit. A mobile processor, such as the Advanced RISC Machine (ARM) processor, is very energy efficient and yet its computing power is such that the life of a VASS device running on a battery is longer, allowing it to be carried anywhere. Finally, advanced mobile operating systems (OSs) , such as iOS and Android, make it easier for researchers to develop the necessary software without compromising the performance of the VASS system. Finally, the mobile device has evolved into a powerful smartphone through the combination of a powerful mobile processor, advanced mobile OS, and other electronics, such as the LCD touchscreen display and audio speaker. This has led to the increased portability of the VASS system. Its portability has helped promote the usage of this SSD, because the user can carry the device anywhere and its weight does not restrict the user to a single location or cause him/her discomfort during its use. This has further expanded the possibilities of VASS into areas such as navigation.

It is worth stressing that not only are the hardware requirements of VASS low but also most of the electronics needed to operate it are now widely available. Since the recent rapid adoption of smartphones, the majority of people in both developed and developing countries now own at least one smartphone. Because all the smartphones now available on the market are equipped with the three main components required by a VASS system, i.e., a camera, speaker and CPU, the development duration of a VASS device has been shortened. The effort expended on developing a VASS

device can now be focused on designing the software applications and improving the algorithm that powers it, whereas for the earlier VASS devices, the whole system had to be designed by researchers from the ground up, including both the hardware and the conversion algorithm.

Overall, the advancement in electronics has mostly benefited the field of VASS. The future of VASS systems is encouraging because of the shift from designing both hardware and software, so that the focus is now on improving the software. Following this trend, the VASS device is expected to perform better in terms of the conversion algorithms and also become more portable in the distant future. It is hoped that, by taking advantage of the opportunities resulting from the advancement in electronics, VASS devices can be widely accepted, especially by the community of the visually impaired.

2.1.2 Better Understanding in Neuroscience

While the advance in electronics has provided the necessary computing power and portable hardware, a better understanding of the brain has afforded researchers the knowledge of how it works, especially in the area of brain plasticity, and this has facilitated the development of improved visual-to-auditory information transfer. Software is an integral part of VASS devices, and a good VASS device requires the combination of powerful hardware and well-designed software that optimizes the information conversion. With the knowledge gained from interdisciplinary research in neuroscience and psychology, we are beginning to understand how humans interpret auditory information using VASS systems. Because of this, it is possible to increase the soundscape interpretability offered by modern VASS devices by using methods such as image processing, intuitive colour to auditory signal mapping, and the application of image saliency.

VASS has been in existence for at least 30 years. The earliest VASS system was developed in the 1990s. This VASS system implemented a simple algorithm that

utilized only the intensity of each pixel from the input images to modulate the sound frequency to produce a soundscape. Then, more complex algorithms were created by incorporating more visual computing into the conversion formulas. Figure 2.1 roughly summarizes the generations of VASS systems that have existed up to now, where the blue gradient represents the involvement of computer image processing during the visual-to-auditory conversions.

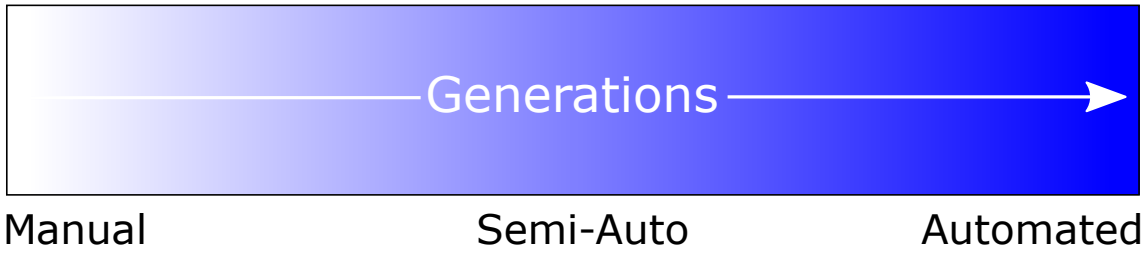


Figure 2.1: Generations of visual-to-auditory sensory substitution systems

In general, VASS systems can be categorized into three major generations, namely, manual, semi-automated, and fully automated. Most of the earlier VASS systems are identified as belonging to the manual generation, because they used minimal image processing in their algorithms. The characteristics of the soundscape of this generation are frequently straightforward and unnatural. The reasons why a simplistic conversion algorithm was used are probably the scarcity of computers with high computational power and a lack of emphasis on the visual-to-auditory conversion algorithm at the time of their development. Because the soundscape is converted through raw conversion without much manipulation, it relies more on the human brain power for interpretation. The user is required to make a mental map based on the sound frequencies to recreate the original input image.

With time, computing resources are becoming more abundant and therefore more researchers are trying to incorporate more computing power into their VASS systems to produce a better soundscape. Hence, image processing algorithms are starting

to be applied in the conversions for the purpose of feature extraction. Through the implementation of image processing algorithms, such as image segmentation, some VASS system developers have attempted to reduce the load on the user during the process of soundscape interpretation. This is achieved through increasing the involvement of computers during the encoding to synthesize a more intuitive soundscape that is easier to understand. For this reason, VASS systems in which advanced image processing algorithms are applied during the visual-to-auditory cross-modality conversion belong to the semi-automated generation.

In the future, it is possible that VASS systems will become fully automated and a machine learning algorithm can be used to recognize objects and explain the visual information to the user in human speech. For example, Karpathy and Li (2015) demonstrated a real-time image captioning application called NeuralTalk that runs on a multimodal recurrent neural network pretrained on a huge dataset of images with corresponding descriptions. At this stage, the computer takes complete control of the interpretation for the user, significantly reducing the need for the user to interpret the soundscape. Although the later generation has some improvements (e.g., better learnability and improved interpretability), there is no conclusive evidence supporting the notion that it is better than the older generation. Therefore, some of the manual VASS devices are still in use.

The following subsections examine some of the related VASS research and devices as part of the Luminophonics background studies.

2.2 Tactile Vision Substitution System

The TVSS is the product of the work done by Bach-y-Rita and his team at the Smith-Kettlewell Institute of Visual Sciences in the 1960s, which reignited interest in sensory substitution for rehabilitation purposes (Bach-Y-Rita et al., 1969; Bach-y-Rita, 1972). The TVSS has since evolved and grown through many research activities and projects, which have taught us many lessons and contributed knowl-

edge about building a good sensory substitution system. Although the TVSS is focused on the application of visual-to-tactile sensory substitution, many parts of the system were precursors for parts of the VASS system. Overall, the team established a basic modern sensory substitution framework through the application of their TVSS, which many research groups have been using.

The general sensory substitution framework introduced by Bach-y-Rita and his team follows a standard communication channel (see Figure 2.2). First, it streams the information captured by an input device into an encoder that transforms it into another modality as a medium for transportation to a decoder. In the TVSS, the team utilized the tactile sense as the medium for transporting the visual information that later is received by a tactile sensor, such as the skin or tongue. Then, finally the brain acts as a decoder to decode the information, restoring it to its original form. In general, the process is standard for all sensory substitution systems, but the usage of a computer to encode the information has only very recently been introduced into TVSSs. Using a TVSS, Bach-Y-Rita et al. (1969) successfully implemented a working prototype that transformed the face of sensory substitution. Since then, most modern SSDs have followed the same framework.

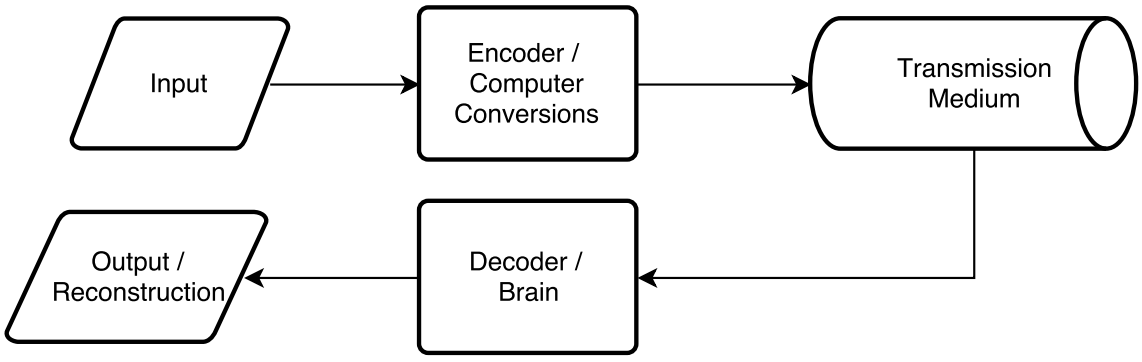


Figure 2.2: General sensory substitution framework

The first TVSS was very large. The user was seated on a large and sturdy

chair. In the back of the chair, 400 vibrators were installed aligned in a square configuration according to the resolution of the image captured by the input camera. The vibrators were wired to a computing unit that was connected to a camera and a monitor. While sitting in the chair with his/her back against the array of vibrators, the user held a camera, which he/she controlled, pointing it in the desired direction. The image was captured by the camera and transmitted to the computing unit for conversion. After the conversion, the array of vibrators was activated, and each vibrator vibrates according to a pixel location on the image. The user needed to recognize the visual images through the vibration felt in his/her back.

As can be seen in the artistic illustration of a TVSS in Figure 2.3[1], the overall size of the first TVSS was very large. The chair, camera, and computer that constituted the device together were the size of a room, which made its use as a mobile device impossible. As the technology advanced, the size of the entire device was reduced. However, the size of a TVSS is limited by the size of the vibrators. The number of vibrators increases in proportion to the resolution of the image to be converted. Moreover, so that the whole device is portable, a large battery is needed as a power supply for the computer and vibrators. Therefore, their size is a major disadvantage of TVSSs that needs to be overcome.

2.2.1 Tongue Display Unit (TDU)

The Tongue Display Unit (TDU) was invented by the same group of researchers to solve the size problem of visual-to-tactile sensory substitution systems so that they can be carried by the visually impaired. The TDU takes advantage of the fact that the human tongue has a more sensitive touch surface than other tactile sensors, such as the skin (Kaczmarek, 2011). Because of the crowded mechanoreceptors, such as Meissner's corpuscles, and the thin epidermis, the sensitivity of the tongue is higher in terms of both pressure sensitivity and spatial acuity. As a result, it can react

[1]Source: Tactile Communication and Neurorehabilitation Laboratory (TCNL), University of Wisconsin–Madison (https://tcnl.bme.wisc.edu/projects/completed/tvss)

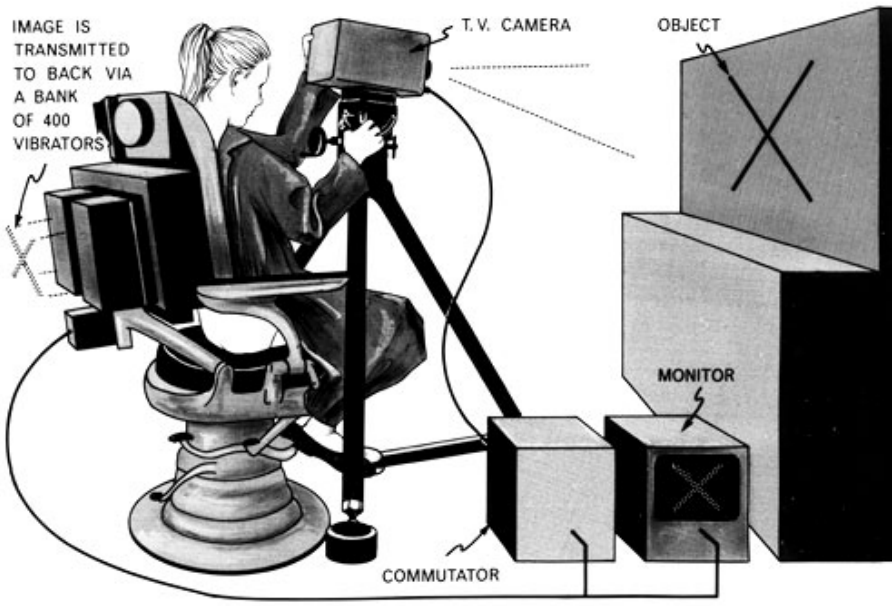


Figure 2.3: Artistic rendition of a tactile vision substitution system

better to smaller tactile stimuli, making it a surface suitable for the application of many tiny actuators. Another reason why the tongue is used is its protected location in the mouth, where hydration is more consistent than on the skin. When tested, the impedance of the electrode varies very little throughout the tongue. For these reasons, an array of electrodes forming an electrode-tongue interface is used to evoke electrotactile sensation (vibration, tingle, and pressure) to communicate temporal and spatial information, when other sensory channels cannot (Sampaio, Maris, and Bach-y-Rita, 2001; Bach-y-Rita, Tyler, and Kaczmarek, 2003; Kaczmarek, 2011).

A TDU comprises an array of electrodes in a flexible printed circuit substrate that interacts with the tongue to create a tactile sensation through electric stimuli. The first version of TDU had 144 electrodes capable of displaying a static 12×12 tactile pattern. As its development advanced, the number of the electrodes in an array grew, while the overall size of the substrate was maintained, making it capable of displaying information at a higher resolution. Currently, a tongue array containing

as many as 400 electrodes exists. The TDU can be coupled with a computing unit as an output device to substitute two type of sensory signals, visual, as provided by a tongue TVSS (Sampaio, Maris, and Bach-y-Rita, 2001; Kupers and Ptito, 2004) and vestibular, as provided by BalanceSensub (Tyler, Danilov, and Bach-y-Rita, 2003; Danilov, Tyler, et al., 2007). In a vision substitution situation, the TDU represents a captured greyscale image by activating the electrodes based on the location and intensity of the image pixels. White image pixels are mapped as strong tactile stimulation, grey pixels as medium level stimulation, and black pixels as no stimulation. The scenario of vestibular substitution is slightly different from that of visual substitution, where the position of the head stimulates a specific location of the TDU. For example, if the head is tilted back, a subarray of the TDU at the back is activated. The user then uses the information received from the tongue sensation to adjust his/her balance.

The TDU was not only demonstrated to be a good device in the research environment, but is also very successful commercially. As a result of a thorough commercialization process, TDU is now available to the public in a package that includes a video camera mounted on a pair of sunglasses, a hand-held controller, a tongue array, and a lithium battery that lasts up to 3 h with a single charge. Currently, it is sold as BrainPort to the public by a company called *Wicab Inc.* founded by Paul Bach-y-Rita and others (Danilov and Tyler, 2005).

2.2.2 Computer Graphics for the Visually Impaired

The researchers at the Tactile Communication and Neurorehabilitation Laboratory (TCNL), University Of Wisconsin–Madison, where Prof. Bach-y-Rita last worked, produced two spin-off devices of TVSS, which are designed to help the visually impaired visualize computer graphics, especially graphical user interfaces (GUI). They recognized that the visually impaired population needs to access and unlock the vast amount of information available on computers as the age of technology

advances. However, the devices need to be tailored for this purpose, with a specific algorithm that can handle most of the information.

One of these devices is in the form of a glove that generates a tactile sensation in the fingers according to the graphically-rich information presented on the computer screen. As reported by Tyler, Haase, et al. (2002), the complete system consists of a glove, which is tethered to the computer and which the blind person puts on one hand. The contact of the tactile actuators spreads from the fingertips all the way down to the top of the palm. The user feels the tactile sensation when scanning/brushing the top of a flat LCD monitor mounted facing upward. While scanning the image, through software control the user is able to control the size of the image by using a zooming feature. He/she can also control other characteristics, such as edge enhancement or black-white colour reversal. This haptic glove provides a good alternative to or even a replacement for the traditional haptic displays, such as Braille displays.

An additional haptic display proposed by the group at TCNL uses the technology of electrostatic stimulation, because this type of display has some advantages over the common haptic displays, such as those that use electrotaneous and vibrotactile stimulation (Agarwal et al., 2002). Its advantages include that batch fabrication using micro fabrication techniques is easy, its power consumption is lower, it can be used without being worn on the body, and it is less bulky. The display consists of a 4-inch thin silicon wafer that has multiple layers of a chemical compound that produces electrostatic tactile sensations. Like that of the TDU, this haptic display is thin and portable. Although its current usage is limited to displaying business diagrams, such as graphs, the initial experiments showed promising results, although more work is needed to reveal its potential.

2.3 Visual-to-Auditory Devices

2.3.1 The vOICe

vOICe is one of the better known VASS devices and is probably the device that is most frequently referred to in this field. The inventor, Peter Meijer, developed vOICe (the three middle capital letters stand for "Oh I See") in 1992 at the Philips Research Laboratories in the Netherlands as a device that offers live camera rendering in auditory form through cross-modality visual to auditory conversion (Meijer, 1992). Since its inception, vOICe has been improved significantly, particularly its form factor: it has evolved from a basic prototype that uses a special purpose computer into applications for multiple platforms that are currently available, including a smartphone application. Currently, on the Website (https://www.seeingwithsound.com/), five types of vOICe can be freely accessed: i.e., Windows Application, Android Mobile Application, WebRTC Application, Raspberry Pi Application, and NVDA Audio Screen. The popularity of vOICe has proven that its developers made a commendable marketing move when they provided the technology on the Internet at no cost on multiple platforms. This not only expanded the userbase of vOICe, but also helped promote the benefits of VASS systems to the public in the process.

As can be seen in Figure 2.4, the visual-to-auditory cross modal information transfer process applied in vOICe is basic and straightforward. This has advantages and disadvantages, but overall it has been demonstrated that vOICe can easily be implemented and the user requires some training to begin using it. Following the usual standard cross-modality conversion, the vOICe conversion process first extracts visual information and then executes visual-to-auditory property mapping. Finally, soundscape encoding synthesis is performed. As an SSD that is marketed as a device that is able to generate live sound representation in real time, vOICe acquires the stream of image frames from a video camera, such as a Webcam or head-mounted camera. The image frames then pass through a simplification process that reduces the amount of visual information. Colour is discarded from the input

image frames, turning them into greyscale images, and the image resolution is then reduced to 64×64 pixels (in the earliest version). This simplification not only reduces the soundscape's duration, making it closer to real time, but also decreases the complexity of the visual-to-auditory properties mapping. According to Meijer, the bitrate each ear can accommodate is about 15 kB/s. By limiting the total pixels per frame to 4096 pixels at 4 bits per pixel and scanning a single frame per 1 s, vOICe is able to produce the soundscape at a bitrate of 16 kB/s (Meijer, 1992; Jones, 2004). These simplification steps contribute to the effort to reduce the effect of cacophony that is commonly faced by VASS users. However, in advanced systems this process is substituted by image processing techniques.

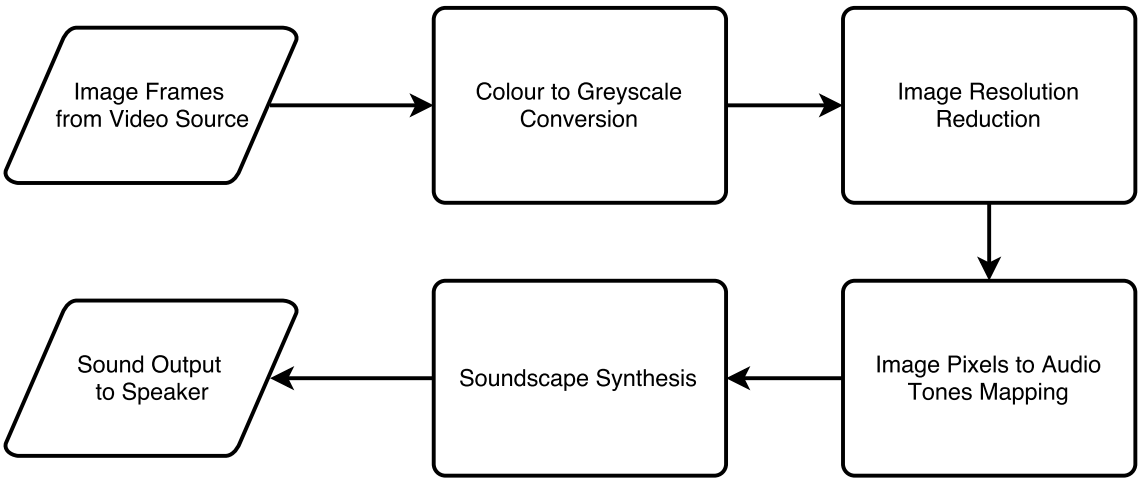


Figure 2.4: vOICe cross-modality conversion process

Each individual pixel p of an image frame (k) with the size M (height) $\times$ N (width) is mapped into 16 grey tones of different frequencies and amplitudes to create an audio representation $s(t)$, as

$$s(t) = \sum_{i=1}^{M} p_{ij}^{(k)} \sin(2\pi f_i t + \phi_i^{(k)}) \tag{2.1}$$

The vertical position i of the pixel affects the frequency f_i of the tone: the higher

the pixel, the higher is the tone. The intensity/brightness of the pixel changes the amplitude of the grey tone, represented by $p_{ij}^{(k)}$. The brighter a pixel, the higher is the amplitude/loudness of the tone. The image frame from the video is sounded in the left to right scanning order in the form of column j. The leftmost column is sounded first, and then, the next column $j+1$ is converted; this procedure continues until the rightmost column ($j = N$) is reached. After the frame is completed, a 'click' sound is appended at the end, denoted by $\phi_i^{(k)}$, an arbitrary constant for the generation of the synchronization 'click'. The process then moves to the next frame $k + 1$ and continues until no image frames remain in the video sequence.

Application

Figure 2.5 shows a screenshot of vOICe running on Microsoft Windows. The application is generating an audio representation of a greyscale image of a car on a road, scanning from left to right at a default rate of 1 s per frame. This application has a higher resolution than the earliest version with a 176×64 resolution by default. In addition, it can capture live views from most universal serial bus (USB) Webcams or head-mounted cameras available on the market that are compatible with Microsoft 'Video for Windows' (VfW). There are also very many options built into the application, including an auditory graphing calculator option, exercises for training, and a text-to-speech function. Another interesting feature that is included in vOICe Windows is the colour filtration option that can be toggled by the user. With colour filtration, the user listens to a selected colour by filtering out other colours. This is another means by which vOICe can handle colour information.

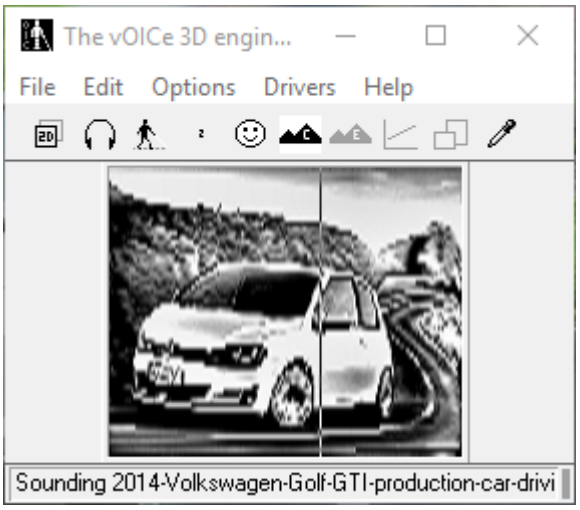


Figure 2.5: Screenshot of vOICe Windows

Figure 2.6[2] illustrates how the visually impaired use vOICe in a normal situation. The user wears a head-mounted camera together with a pair of speakers connected to a processing unit (normally a laptop computer carried in a backpack). In the picture, the user is attempting to locate the mug by grasping it with his hand while listening to the soundscape produced by vOICe. It is essential that the user wear the camera on the top of the head because he/she can then move the camera around, simulating how human eyes examine the surrounding.

Figure 2.6: User of vOICe

[2]Source: IEEE Spectrum Feb 2004

Discussion

vOICe was invented in 1992 when tactile sensory substitution for visual signals was the preferred modality rather than its auditory counterpart, which made it one of the earliest VASS systems. At the time it was conceived, researchers were still attempting to find the best means of accomplishing visual-to-auditory cross-modality conversion efficiently, because computation power was limited. Therefore, vOICe used a special purpose computer built using standard low-power Schottky transistor-transistor logic (LS-TTL) technology to translate visual information into an auditory representation by applying a simple and direct mapping method. Consequently, the user needs to use more manual effort to interpret the audio representation, which explains why vOICe belongs to the manual generation (see Figure 2.1).

Despite the advantages of the system that it is simple and easily implemented, interpreting its soundscape causes the user more fatigue over a long usage period, because it depends heavily on the user. Users need to decode the auditory representation themselves to restore the original visual information to a mental map. During the process, other auxiliary tasks, such as noise filtering and object recognition, are run in parallel, which can tire the brain. Thus, the outsourcing of some of the tasks to a computer through intelligent algorithms and image processing methods may lighten the user's cognitive load, and hence, reduce his/her fatigue and also improve the overall usability of the SSD.

In addition, it is difficult to learn to use VASS devices (especially those belonging to the manual generation). In order to increase the user's proficiency so that he/she can use vOICe in daily life, structured training needs to be provided, as well as frequent usage, to increase the efficiency of soundscape interpretation. Although several basic training materials are provided on the Seeing with Sound Website, which help the user to start using the device, it is important to emphasize that a structured tutorial would help the beginner learn to use the device in practical environments in a shorter time.

2.3.2 Prothesis Substituting Vision by Audition

PSVA stands for Prosthesis for Substitution of Vision by Audition. PSVA is a VASS system developed by Capelle, Faik, et al. (1994) of the Université Catholique de Louvain. The system is one of the earliest VASS prototypes, having been developed in 1994, two years after the introduction of vOICe. Although it was established in the early 1990s, PSVA paved the way for the application of image processing techniques to extract more information from a normal 2D image in order to closely model human vision. Hence, it is regarded as a semi-automated VASS system. Another earlier feature that PSVA introduced into VASS systems is its emphasis on operating in real time as a means of affording the user an efficient sensory-motor interaction. Because of this, it was designed from the ground up to be a real-time VASS system for the purpose of vision rehabilitation. The resulting PSVA manages to achieve a maximum total data processing (from video grabber to speaker) duration in a laboratory setting of 100 ms (Capelle, Trullemans, et al., 1998).

PSVA follows the same framework as other VASS systems for converting video signals into soundscape. It acquires visual information from an image frame (in 64×64 pixels) captured by a video camera. As opposed to vOICe, which only converts colour images into greyscale images, in PSVA an attempt was made to model humans' primary visual system by implementing features such as lateral inhibition and graded resolution. In the human visual sensory system, lateral inhibition increases the contrast of a certain region by reducing the signals from its neighbouring regions. To simulate lateral inhibition, in PSVA a Laplacian of Gaussian (LoG) convolution filter is implemented and then the detection of zero-crossing in the input image is performed. The filter suppresses signals other than those of the edge to produce an input image with edge detection.

In a normal situation, human vision utilizes the focus of attention to segment visual data into multiple regions, thus reducing the information load to be processed by the brain (Hubel and Wiesel, 1977; Balasuriya and Siebert, 2003). Inspired by

this, in PSVA graded resolution is implemented using a technique called multiresolution artificial retina. An artificial retina with two levels of resolution pyramid consists of an 8×8 pixels low resolution grid with 2×2 pixels at the centre grid, subdivided into another 8×8 pixels, amounting to a total of 160 pixels (from the original 4096 pixels). This process can be continued to four levels of resolution pyramid, which can be as detailed as a total of 208 pixels. This graded resolution both reduces the duration of conversion and greatly lessens the effect of cacophony in the soundscape.

The conversion process first uses the information from the previous image processing stage and applies it in visual-to-auditory mapping. Using a basic coding scheme that is based on the association of pixel intensity and audio frequency, PSVA presents two types of mapping codes, segregated code and interlaced code. Both codes have the following basic features.

- Pixel vertical position $\longrightarrow$ audio pitch
- Pixel horizontal position $\longrightarrow$ binaural intensity and phase differences
- Pixel brightness $\longrightarrow$ audio loudness or amplitude
- The frequencies are chosen such that the columns resemble melodious sound and the rows resemble harsh sound.

Segregated code assigns a different frequency value to each pixel location with f_0 as the base frequency (normally 50 Hz), following

$$f = f_0 \cdot 2^{p/32} \tag{2.2}$$

Interlaced code assigns the frequencies so that the fovea (central pixel region) and the periphery (outer pixel region) are not harmonic with each other. The formula for generating frequencies is particularly flexible, ranging from two to four levels of resolution pyramid. With f_0 as the base frequency, the frequencies for two levels of resolution pyramid are generated using

- For periphery grid,

$$f_{periphery} = f_0 \cdot 2^{p/8} \cdot 2^{1/16} \tag{2.3}$$

- For fovea grid,

$$f_{fovea} = f_0 \cdot 2^{p/32} \cdot 2^{1/64} \tag{2.4}$$

Because PSVA uses a pregenerated frequency mapping code, in this system the need to recalculate the audio frequencies for each conversion is eliminated, resulting in a fast and efficient conversion. To further reduce the duration of soundscape generation, two dedicated printed circuit boards (PCBs), containing the control circuitry and music processors, respectively, were built so that the entire process can be executed in real time by offloading the processes from an Intel 486 processor. The PCBs are connected to a speaker to output the soundscape.

2.3.3 See ColOr

Bologna, Deville, Pun, and Vinckenbosch (2007) advanced the development of VASS systems by incorporating many new ideas in his project, Seeing Colours with an Orchestra (See ColOr), after years of research in the field of assistive technology for the blind. Prior to See ColOr, Bologna and Vinckenbosch (2005) created the Ambisonic 3D-sound field, which utilizes an eye tracker to assist the capture process of the visual input device based on inherent attention combined with musical instrument sounds to encode colour pixels. Visual substitution through the Ambisonic 3D-sound field introduced several innovative ideas into the research field, including the differentiation of colour by different sounds of musical instruments, head-related impulse response (HRIR) to simulate 3D surround sound, and an eye tracker to limit and crop visual input based on the attentional field.

The semi-automated generation of the VASS system appeared around the year 2005, as computers began to play a role in assisting the processing of visual information to increase the efficiency and interpretability of the auditory representation (Bologna and Vinckenbosch, 2005). These sequences of events were spurred by the

increase in available computing power and the efficient image processing algorithms contributed by the field of computer vision. A computer can both assist in extracting more information from a raw image and be used to filter out information that is irrelevant to the user. As demonstrated by Ambisonic 3D sound, it can be used to imitate humans' attentional visual field by cropping the entire raw input image into a smaller image by computing the importance of the region.

See ColOr is a VASS prototype that is focused on guiding the visually impaired and is thus configured specifically for navigation purposes. Several aspects that were previously used in VASS systems are implemented in See ColOr. In particular, it includes colour information, which the system is able to encode as musical instrument sounds during the cross modal conversion. In the past, many VASS systems were designed to sonify colour in the soundscape, but frequently this introduced more noise and the cacophony effect, which causes deterioration in the user's interpretation performance. Colour information is important for recognition because the description of coloured objects and their textures helps the user build consistent mental images of the environment. With colour information, the user is able to recognize objects more easily than when he/she distinguishes objects purely by their texture in greyscale, because colour information decreases the level of ambiguity during the process of recognition. Moreover, colour is important for navigation because traffic and safety signs utilize colour to impart a warning message to the user (i.e., red signifies a prohibition/danger alarm, yellow/amber a warning, blue a mandatory instruction, and green an emergency escape/first aid/no danger message).

In addition, See ColOr demonstrated the ability to encode depth through a special spatialisation method using the depth map of a stereoscopic camera. Because the amount of information collected from a stereoscopic camera is frequently very large, it is a norm to simplify the visual information before executing the conversion in order to avoid a situation where the user is overwhelmed by all the different sounds in the soundscape. In order to resolve the issue of cacophony, in See ColOr image simplification is implemented by means of segmentation and guiding the focus

of attention (FOA) through the computation of visual saliency. The inspiration behind FOA, which is implemented in See ColOr, is humans' natural attentional visual field. As shown in Figure 2.7, See ColOr performs image simplification in parallel with computing the FOA using the raw coloured image frame. In Bologna, Deville, Pun, and Vinckenbosch (2007), mean shift, K-means and quadtree algorithms were tested to obtain a suitable algorithm that provides the greatest number of regions of optimum size without sacrificing computing time. In general, the images produced by the simplification process contain less noise, because visually similar pixels are merged together to form multiple segments. In contrast, some fine details are discarded in the process, resulting in the loss of texture details in favour of a simpler image.

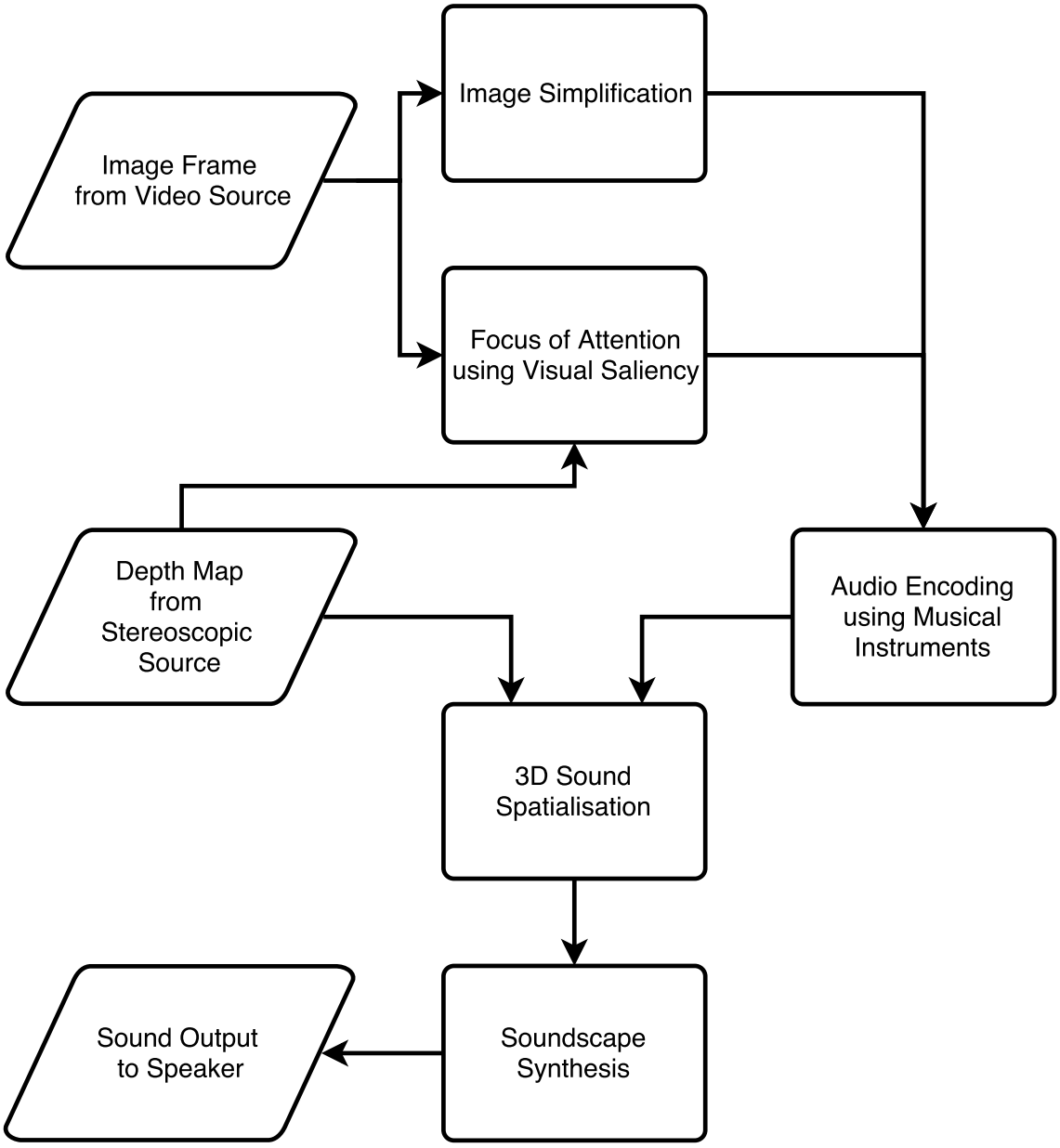


Figure 2.7: See ColOr cross modal conversion process

Visual saliency is computed from the same raw image frame. According to Landragin (2004), visual saliency refers to the visual mechanism that is linked to the emergence of an object from a background that captures the attention of the indi-

vidual. Humans frequently depend on visual saliency to determine the region on which they should focus, so that the brain does not need to process the entire field of view, and thus, the cognitive load imposed on the brain is reduced. Factors such as the iconic features of the scene and cognitive preference determine the value of visual saliency. Many computerized methods to compute visual saliency exist, but only a few are able to combine both physical features and cognitive factors to determine the value of visual saliency. Methods that focus only on physical factors are called bottom-up approaches, whereas cognitive-based ones are termed top-down approaches. Considering that the See ColOr system is not aimed to replace the cognitive abilities of the blind user, the top-down attention model was selected to understand the captured scene (Deville et al., 2008). The saliency regions detected are based on conspicuity maps computed using methods that focus on specific characteristics of images, such as entropy or blobs (Kadir and Brady, 2001), such as difference of Gaussians (DoG) or the speeded up robust features (SURF). Interestingly, See ColOr utilizes a depth map acquired from a stereoscopic camera to determine objects of interest and filter out those that can be ignored. The importance of the objects is based on the distance between them and the user. The closer the objects, the more their importance increases. However, because See ColOr is designed to be used in conjunction with a mobility cane, objects located at less than a minimum distance, d_{min}, and reachable by the cane are discarded. Objects that are located at a distance that exceeds the maximum distance, d_{max}, set by the user, are also discarded. Using p_z as the depth value of the pixel p, the distance feature map, F_d, can be simplified using

$$F_d(p) = \begin{cases} d_{max} - p_z, & \text{if } d_{min} < p_z \leqslant d_{max} \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

Motion is another important information feature that can be derived from the depth map. Using Equation 2.6, the object's motion is computed from the gradient of depth ∂p_z over time t frame by frame as received from the camera to determine

whether it is moving towards or away from the user. Objects that are moving closer to the user are retained and the others are discarded.

$$F_{\nabla}(p) = \begin{cases} -\frac{\partial p_z}{\partial t}, & \text{if } \frac{\partial p_z}{\partial t} < 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

In order to obtain the corresponding conspicuity map of each feature map, DoG is applied to the feature map (e.g., depth feature map $F_d(p)$ and motion feature map $F_{\nabla}(p)$). DoG here acts as a band-pass filter to discard all values outside σ_1 and σ_2, which essentially reduces the amount of information by excluding non-relevant information, including noise. Finally, a saliency map is formed from the weighted sum of all the computed conspicuity maps.

To produce the auditory representation, the pixels' values in the saliency map are encoded using colour-to-musical instrument mapping developed for See ColOr. Two experimental mappings are used in See ColOr, namely, flat audio encoding and 3D sound spatialisation. Flat audio encoding encodes into an audio representation by transferring the values from colour pixels matching a set of musical instruments, while 3D sound spatialisation enhances flat audio encoding to create a perception of localised sound sources called Ambisonic through the application of personalised head-related transfer functions (HRTFs). The hue, saturation, and luminosity colour model (HSL) , a cylindrical-coordinate representation of colour, is used to extract the colour value from the pixel instead of the original red, green, and blue colour model (RGB) colour model supplied by the camera. As a symmetric double cone, HSL has the advantage of having symmetrical lightness and darkness at both opposite ends, while the hue determines the colour in degree (from red to purple) and saturation measures the concentration of the colour. As compared to RGB or other colour models, such as the hue, saturation, and luminosity colour model (HSL), HSL is closer to human vision and more intuitive. The system translates hue to eight different instruments and saturation to four different notes (C, G, B♭, and E) and uses luminosity to determine whether a double bass note or a singing note is used.

To create the perception of depth in the soundscape, the duration of the sound is calculated based on the depth value p_z received from the stereoscopic camera, converted to the equivalent metric meter, D. This emulates sound travel, where the sound of an object situated farther away or closer takes a longer or shorter time, respectively, to reach the user. Detailed mapping values are presented in Tables 2.1, 2.2, 2.3 and 2.4.

Table 2.1: Mapping of hue, H, to musical instrument by See ColOr

Hue, H	Instrument
Red ($0 \leqslant H < 1/12$)	Oboe
Orange ($1/12 \leqslant H < 1/6$)	Viola
Yellow ($1/6 \leqslant H < 1/3$)	Pizzicato Violin
Green ($1/3 \leqslant H < 1/2$)	Flute
Cyan ($1/2 \leqslant H < 2/3$)	Trumpet
Blue ($2/3 \leqslant H < 5/6$)	Piano
Purple ($5/6 \leqslant H < 1$)	Saxophone

Table 2.2: Mapping of saturation, S, to note by See ColOr

Saturation, S	Note
$0 \leqslant S < 0.25$	C
$0.25 \leqslant S < 0.5$	G
$0.5 \leqslant S < 0.75$	B♭
$0.75 \leqslant S < 1$	E

Table 2.3: Mapping of luminosity, L to note by See ColOr

Luminosity, L	Note	Type
$0 \leqslant L < 0.125$	C	Double Bass
$0.125 \leqslant L < 0.25$	G	Double Bass
$0.25 \leqslant L < 0.375$	B♭	Double Bass
$0.375 \leqslant L < 0.5$	E	Double Bass
$0.5 \leqslant L < 0.625$	C	Voice note
$0.625 \leqslant L < 0.75$	G	Voice note
$0.75 \leqslant L < 0.875$	B♭	Voice note
$0.875 \leqslant L < 1$	E	Voice note

Table 2.4: Mapping of distance, D, to sound duration by See ColOr

Depth, D in m	Sound duration in ms
Undetermined	90
$0 \leqslant D < 1$	160
$1 \leqslant D < 2$	207
$2 \leqslant D < 3$	254
$3 \leqslant D < \infty$	300

Application

See ColOr is designed to guide the visually impaired for navigation purposes, and therefore, it is not expected to perform well in heavy feature recognition tasks, such as object identification, because of the lack of feature information that is encoded. As reported in Bologna, Deville, and Pun (2008), an experiment involving two scenarios was conducted to measure the performance of See ColOr in terms of assisting the user in the scenarios. The first scenario tested the ability of the soundscape to describe colour correctly using musical instruments. The experiment was performed

using seven blindfolded adults who were asked to find the matching half of a pair of socks of the same colour. Instead of associating the colour of the socks by listening to the sound, they were asked to find the matching pair of socks, because it is difficult for first-time users to make the colour-sound association after one training session. Before the actual experiment, each participant received a two-step training session. The first step consisted of static training. The participants were shown static images on a laptop computer while a trainer explained the colour associated to the sound the participant heard. The second training step involved an activity that was closer to the experimental task. The participant held a pair of socks and learned to associate the colour with the sound. During the actual experiment, five pairs of socks having the colours black, green, low saturated yellow, blue, and orange were presented to the participant for matching and the time taken by each participant to complete the task was recorded. On average, a person can match a pair of socks correctly in 25 s, or 4 m for a total of five pairs of socks. An interesting finding from this experiment was that a participant (who was one of the authors of See ColOr) was able to complete the task in 2.2 m, which was almost twice as fast as the next fastest participant. This further proved that training and prior knowledge does indeed increase the user's efficiency when using an SSD.

The second scenario of the experiment involved the participant following a coloured serpentine path laid on the floor in an outdoor environment. This scenario was designed to investigate how effectively See ColOr can assist a user in a navigation task. The same seven participants as took part in the previous experiment were invited to take part in this test. They were required to wear a Webcam connected to the See ColOr system installed on a laptop computer. They were instructed to follow the red coloured tape stuck on the floor, called the 'serpentine path', by listening to the soundscape produced inside their earphones. As usual, each participant was provided with a short training session in which they were guided to listen to a sound pattern that contained an oboe sound (representing red) and a double bass sound (representing grey, the colour on both sides of the red tape) while following the ser-

pentine path. During the testing phase, the time it took a participant to walk from the point of origin to the destination while listening to the soundscape was measured. In order to complete a navigation task as demonstrated in this scenario, the user needs to learn to use at least three components: the sound position of objects, awareness of head orientation, and alignment of the body and head. The results of this experiment showed that, when colour information is implemented correctly, users receive more information, which makes it easier for them to determine their surroundings. It is essential that this information be included in a VASS system to assist a user to navigate.

Discussion

In See ColOr, the application of VASS was advanced by incorporating interesting ideas, such as colour, musical instruments, image processing, and a stereoscopic camera, but it was far from perfect. Because the concept was relatively new at the time of its development, there were not many systems with which to compare See ColOr. It is unconfirmed how the system compared with other systems, because it was tested in an experiment designed by the researchers. An improved general testing framework is needed that can fairly evaluate and quantify the performance of multiple VASS systems in terms of different aspects. In addition to evaluating the performance, such a general framework could reveal the advantages and disadvantages of a system. It would be beneficial in a situation such as that described above, because it would be easier for the developers of a device such as See ColOr to identify an idea that does not perform well without excessive testing.

The new implementations in See ColOr were still preliminary, and there are areas that can be improved to increase the effectiveness of the system. For instance, what is the optimum mapping of colours to musical instruments? As shown in Table 2.1, Bologna, Deville, Pun, and Vinckenbosch (2007) proposed a set of seven colours (red, orange, yellow, green, cyan, blue, and purple) to be matched with seven musical instruments (oboe, viola, pizzicato violin, flute, trumpet, piano, and saxophone).

There are several questions related to this proposed mapping scheme, including why a non-standard colour set was chosen instead of a rainbow colour set. In addition, the mapping of colours to musical instruments according the human psychological perception of colour might improve the soundscape's interpretability and help the user learn the colour association during the training period. Let us consider the case of traffic signage. Warm colours such as red are frequently associated with a warning or hazard. It might be more intuitive for the user if the colour red was mapped to a musical timbre that is related to hazard, such as a horn. Moreover, the computation of visual saliency needs to be re-examined so that the algorithm does not discard important information, which might hamper the system's usability and endanger the user during navigation.

To conclude, more studies are required, especially interdisciplinary ones involving human psychology and neuroscience, to design a better system with high usability that produce a soundscape that is more intuitive and easily interpreted.

2.3.4 EyeMusic

The two most recent directions of VASS research are aimed at overcoming the problems of the unpleasantness of the soundscape and the non-inclusion of colour information in the soundscape. Following the trend, EyeMusic, a VASS system developed by a team of researchers at The Hebrew University of Jerusalem, is aimed at solving these two problems with a single solution (Abboud et al., 2014). Similarly to See ColOr, EyeMusic uses musical notes to encode colour visual information in order to convey shape and colour information in an auditory form that is pleasant to listen to (Hanassy et al., 2013). This proposed VASS system can be considered to belong to the manual generation because, like vOICe, it simplifies the input image by reducing the image resolution but retains the colour information.

Colour vision is an important feature of sight that provides humans or other organisms with the ability to distinguish objects based on the different wavelengths

of the light reflected from the object. Humans frequently depend on colour information to perform various tasks in their daily life. The reports of Bramão et al. (2011), Goffaux et al. (2005), and Yip and Sinha (2002) further confirmed that humans use colour vision for tasks such as object and face recognition, scene reconstruction, and navigation. Without colour, humans take longer to execute these tasks or even fail to complete them. For this reason, it is imperative that a VASS system include colour as a part of its vision-to-auditory cross modal conversion in order to help the visually impaired in their vision rehabilitation. This motivated the developers of EyeMusic to solve the problems that arise as a result of the incorporation of colour information such that the overall performance of VASS is not sacrificed.

The objective of the developers of EyeMusic was to create a general purpose VASS system to sonify the surrounding shapes, including their colour information, in a pleasant manner. In contrast, the developers of See ColOr, one of the systems most similar to EyeMusic, focused on building a VASS system with a single specialized function: navigation. The developers of EyeMusic approached the conversion by processing each image column-by-column in a single direction from left to right. This technique is called the sweep-line technique. This approach is identical to the swiping technique that is first proposed in an article describing the first prototype in this study (see Tan, Maul, N. R. Mennie, and Mitchell, 2010). Figure 2.8 succinctly shows the process flow of EyeMusic, which is similar to that of vOICe (see Figure 2.4), except that EyeMusic features a colour clustering algorithm to retain the colour information instead of the colour to greyscale conversion that is used in vOICe. In the process, first image frames are taken from an input video source, such as a camera or computer screen. Then, the image frame passes through a resolution reduction process so that the amount of visual information is decreased to lessen the effect of cacophony in the soundscape. The resolution of the image is reduced by expanding each pixel using a pixelation technique into a larger pixel having a width of 40 pixels and height of 24 pixels. With the resolution, the image is split into multiple columns 40 pixels in width. A soundscape is then synthesized for each image

frame column by column in one direction from left to right. At the start of each soundscape, a cue sound is played to signify the beginning of the soundscape. The visual-to-auditory mapping follows the HSL colour model to determine the colour of the pixels. Each larger pixel in a column is mapped to a single musical instrument according to the colour taken from the hue value, H. The pitch of the musical note is set according to the Y-axis of the larger pixel and the volume of the musical note is set according to the luminance level, L, of the larger pixel.

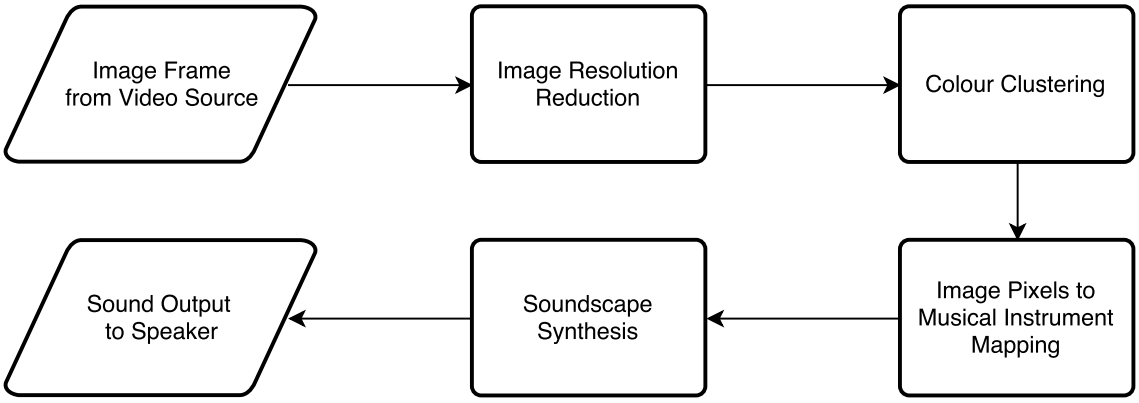


Figure 2.8: EyeMusic cross modal conversion process

Application

At the time of writing, EyeMusic can be found in multiple implementations, including mobile applications in the form of iOS and Android. On its official Website, managed by one of the authors (http://www.amedilab.com/), there are several interesting applications, including a tutorial that teaches a beginner how to use EyeMusic, a Web application that converts a static image into an EyeMusic soundscape, and an HTML5 game based on EyeMusic. It is refreshing to find a simple game based on VASS technology. The game is designed as a shooting game in which a cannon situated at the bottom of the screen attempts to shoot a target at the top of the

screen. In order to win the game, the player needs to correctly shoot at the target, which is located at the top right hand corner, in the middle, or at the top left hand corner, based on the soundscape. This simple game has demonstrated that thus far VASS has achieved much, considering that the most recent VASS implementation is able to accurately encode the information of feature, colour, and location combined together.

The researchers who developed EyeMusic have publically promoted their device extensively in the past few years. This is new and very welcoming, because sensory substitution technology was largely unknown in the past, as the research of sensory substitution frequently occurs inside the laboratory, where most users use the technology in a closely monitored environment. The activities of the group of researchers at The Hebrew University of Jerusalem have successfully raised the public's awareness of this technology and helped the visually impaired population understand the benefits of VASS. They followed the good example set by the developers of vOICe in that they set up a Website that provides tutorials and information and released their software to the public who are interested in testing it. Moreover, the primary researcher, Prof. Dr. Amedi Amir, gave an interesting talk at TEDxJerusalem (an independently organized TED event in Jerusalem) entitled 'Seeing with the Ears, Hands and Bionic Eyes' in which he briefly summarized how sensory substitution technology can benefit the rehabilitation of visually impaired people[3]. In addition to promoting and demonstrating EyeMusic, he conveyed a considerable amount of useful information about sensory substitution to the audience. Public reaction to the presentation was very encouraging, as the audience realised the advantages that sensory substitution brings to the world of assistive technology.

[3]Link to 'Seeing with the Ears, Hands and Bionic Eyes' by Amir Amedi at TEDxJerusalem: https://www.youtube.com/watch?v=jVBp2nDmg7E

Discussion

The new ideas realized in EyeMusic may be promising, but more actual tests need to be conducted to confirm their feasibility and evaluate the performance of EyeMusic. For example, the choice of instruments used in EyeMusic will remain questionable until there are proven results to support it. The reason for this is that some musical instruments sound similar, especially those that belong to the same group, such as the cello and guitar, which are both string instruments. Using musical instruments that sound similar will degrade the quality of the soundscape, causing the user confusion because of the effect of cacophony. Furthermore, questions exist regarding the selection of colours; for example, is their reflection of the surroundings beneficial to the user? Moreover, does the number of colours also affect the interpretability of the soundscape? More colours might better describe the surrounding but would add to the time needed for interpretation. It is essential to conduct systematic tests to find the appropriate number of colours together with the correct choice of musical instruments that optimize interpretability without sacrificing the functionalities.

While the work of the group led by Prof. Dr. Amedi Amir has improved VASS technology, most importantly their public relations work has raised the awareness of this technology. During the TEDx talk, the public reception of the technology proved that they previously had little or no knowledge of this type of technology. After the talk, there was wide coverage of this technology on both mass media and social media, which spurred the public to discover more about sensory substitution. Through their promotional activities, the EyeMusic research group may have solved one of the key problems that contribute to the low level of public adoption of SSDs. In order to encourage public usage, not only the performance of the device need to be focused on but also the marketing efforts to promote the technology to the public.

2.4 Discussion

There is much to be learned from the popular visual sensory substitution systems, including both the TVSS and VASS. The work of researchers has contributed significantly to the field of sensory substitution and has facilitated the further development of the systems. It should be noted that, in most of the published papers, it is agreed that visual sensory substitution can benefit people with visual impairment; however, gaps remain that need to be filled to allow its adoption by the public.

A common feature of all the visual sensory substitution systems is that they share the same framework. The general sensory substitution framework, as described in Section 2.2, was inspired by the traditional communication channel, which consists of three major components: encoder, transmission medium, and decoder. For instance, in the VASS system an algorithm is applied to encode visual information into a soundscape, which is transferred through a medium; in this case, the sound of the audio representation is transmitted over the air. The ears of the user play the role of the receiver: the user listens to the soundscape and his/her brain attempts to interpret the visual information contained in the soundscape by decoding it using the visual-to-auditory mapping that was used by the computer during the conversion. Because this framework has been proven to be effective in all VASS systems thus far, it was decided that this research too would follow the same framework as the general structure in terms of designing the basis of the prototypes.

However, the majority of VASS systems have their own unique process for visual-to-auditory conversion, from the most simple, which is extracting pixel intensities (as in vOICe and PSVA), to those that include image processing techniques to extract more visual information (as in SeeColOr and EyeMusic). Although the systems have their own complexity in terms of handling visual information, they also have their own advantages and disadvantages, which is why no approach is clearly superior. For other up-to-date systems, such as SeeColOr and EyeMusic, means of including more information in the conversion process were explored. For example,

the developers of SeeColOr were the first to propose that it is necessary to include colour information in the conversion mix, which was rarely considered in the earlier VASS systems. Moreover, Bologna, Deville, Pun, and Vinckenbosch (2007) were the first to utilize a 3D camera to provide an accurate depth map to assist in the visual saliency computation for visual-to-auditory conversions. The effort invested by Bologna, Deville, Pun, and Vinckenbosch (2007) to attempt to improve the performance by increasing the input information is commendable; however, the cost of this improvement is a considerably more complex soundscape. The complexity of the soundscape affects the interpretability of the output of the system, because it may generate more signal noise and cause the soundscape to be cacophonous. When designing an improved algorithm, the Luminophonics team examined the approaches taken in all the related research and their effect on the soundscape in order to find the balance between the richness of the soundscape and its interpretability.

In the area of sound synthesis, in the earlier VASS systems a direct approach was used according to which the audio frequency was attenuated. Although this approach simplified the entire process, it generated an unnatural soundscape. The up-to-date approach however utilizes an advanced sound synthesizer library to create a natural yet vibrant soundscape by utilizing models of the timbres of various musical instruments. Moreover, the use of different timbres has the advantage that more sound representation is achieved, which leads to an increase in the amount of information that can be encoded.

As visual sensory substitution is still at the development stage, many researchers are experimenting with different methods and formulas to create a better visual-to-auditory conversion, and thus, a better soundscape. Despite the fact that good results based on user evaluation have been reported for VASS systems, there is no clearly correct method for designing a good VASS system. This is probably due to the lack of standardized measurements for evaluating these systems. Currently, research groups that design their own system also create their own evaluation criteria to measure the performance of their device. As a consequence, these evaluations in

silo have hindered the growth of VASS systems. This motivated us to create a standardized evaluation platform that has the ability to quantify the performance of every VASS device.

Chapter 3

Prototyping

3.1 Introduction

As the Luminophonics project was new at the time the research started in 2009, not many free resources were readily available and few similar systems existed. Instead of studying solely on VASS, the search for information about SSDs was expanded to include other types of SSD, such as those featuring visual-to-tactile conversion. Despite the effort, most SSDs are closed-source and on top of that, they are frequently difficult to replicate internally because of hardware and software constraints. Therefore, it was decided that a top-down approach was applied as the overall strategy for this research. In a top-down approach, so that more insight into the inner functions of the existing systems is gained, they are decomposed, that is, broken down into smaller components. Different working systems such as vOICe and See ColOr were examined down to the individual process level in order to fully understand the features implemented in them. Then, several prototypes were built based on the combination of the promising features from the pool of features that were being investigated. Following the prototyping process, discussions were conducted on adjusting and fine-tuning the features that were implemented in the prototypes. The main purpose of this was to solve the problems and improve the performance of the

systems built in this studies. In parallel, the prototypes underwent a series of tests, including user-based experiments, to measure their individual performance. This section covers only the activities of the prototyping process and the methodology behind the development phase of the Luminophonics project. Chapter 4 and Chapter 5 document the details on the experiments and the performance measurements of the prototypes that were conducted.

3.1.1 Initial studies

A number of studies were conducted on the majority of SSDs including Meijer (1992), Capelle, Faik, et al. (1994), Balakrishnan et al. (2005), Bologna, Deville, Pun, and Vinckenbosch (2007), and some other minor systems. They were done with the purpose to understand the design of an SSD and also the advantages and disadvantages of each implementation. Currently, the existing systems can be categorized into three different forms, namely, open-source systems, closed-source systems, and those that have been presented in the literature, including in conference papers and journal publications.

Of all the systems, open-source systems are the easiest to study because their underlying algorithms and the conversion process are detailed in the source code. In general, they can be downloaded from the official Website of the system or major source code repositories, such as Github, Bitbucket, and Sourceforge. The downloaded material frequently includes instructions on how to set up the software and the relevant configurations needed. With the correct hardware, the systems can be re-implemented in their entirety. Examples of the open-source systems that are available online include OpenSonify[1], Wavy by Nicolas Louveton of the University of Luxembourg[2], Sensub by Stefan Strahl of the UCL Ear Institute[3], and other generic implementations from open-source contributors on the Internet. Although the avail-

[1]OpenSonify Repository: https://sourceforge.net/projects/opensonify/
[2]Wavy Repository: https://bitbucket.org/nblouveton/wavy/overview
[3]Sensub Repository: https://code.google.com/archive/p/sensub/

ability of source code facilitated the investigation, most open-source SSD material covers only the basic sensory substitution features that are not cutting-edge. This has a major effect on their performance, which is frequently much lower than that of their closed-source counterparts. In addition, the purpose of providing open-source SSDs is mainly to demonstrate to the general public the ability of sensory substitution. Therefore, the open-source SSDs are developed for the public, utilizing basic consumer hardware, such as Webcams and smartphones. Because the open-source SSDs are lacking in terms of sophisticated software and specialized hardware, the maximum capabilities of sensory substitution that they achieve is not equal to those of the existing high-quality SSDs. However, this study did not exclude the open-source SSDs because it is essential to examine them for the sake of overviewing SSDs and the conversion algorithms. Moreover, from these systems, one can learn the basic structure of the software design that powers an SSD and how best to implement an SSD from the source code. The investigation of SSD did not end here. The effort to explore and study systems other than open-source SSDs continues to understand the features that contribute to building a better VASS system.

As well as open-source SSDs, many studies were conducted to cover closed-source devices and the SSDs that have been presented in the literature. As compared to the open-source SSDs, the performance of these systems is frequently much better as a result of the effort and the resources that were invested by the researchers and inventors. However, for the same reason, some of the internal and system design is not disclosed, mainly because of future commercialization opportunities. However, it is necessary to invest a considerable amount of effort in analysing the systems because they incorporate many cutting-edge features with up-to-date algorithms. The systems that fall into this category are described in detail in Chapter 2. They include the The vOICe, Prothesis Substituting Vision by Audition, See ColOr, EyeMusic, and the VTSS. Of these five systems, vOICe and EyeMusic exist in the form of an executable software application that the public can download and use. These executables provided some avenues to test the functionalities of the individual systems

in detail and to examine the process behind the visual-to-auditory conversion.

The final category consisted of systems, the software of which is not available to the public, such as See ColOr and PSVA. The information about these systems is available in the form of papers published by the authors of the systems. The only possible means of examining these systems is to re-implement the mechanism behind them. Some potential features, such as colour-to-auditory mapping and image segmentation processes were re-implemented in some of the prototypes of this research to gain a better understanding of their operation. However, intricate parts, such as the hardware-software integration and the physical aspect of the entire system, are difficult to imitate because they require that both the hardware and software operate according to specific configurations and settings. This information is frequently not disclosed in publications. It is possible that the performance of the systems is elevated by the configurations, as well as by the hardware-software integration tuned specifically for the system. Although the reimplementation did not achieve the highest performance as specified in the publication, the advanced design and modern ideas included in the systems provided many insights into means of the improvement of the VASS system.

To conclude, many extensive studies on similar systems were conducted during the exploratory stage. In these studies, the main intention was to search for the features and design that affect the overall performance of the system. These studies included examining the source code, using the software repeatedly, conducting rigorous system testing, and reimplementing some of the features in the prototypes. Then, efforts were expended to improve the performance of the VASS system by upgrading these features in combination with some innovative approach, to reduce the effect of the features that negatively influenced the performance of the system.

3.2 Common Features

In total, four prototypes and one mobile prototype (in the form of an Android smartphone application) were produced in this studies. All the prototypes were developed according to a single common framework and using the same software architecture, with several additional features implemented individually that differentiate the prototypes from each other. In this section, the common features shared across the prototype and their implementation are discussed in detail. These features include the architecture of the software, the type of hardware used, and the processing of colour information.

3.2.1 Software Architecture

Figure 3.1 shows in detail the software architecture that was utilised in the prototypes. In general, the architecture consists of three layers of the process flow. The information flow in sequential order starts with the imaging layer (left), followed by the image processing layer (middle), and finally the sound synthesis layer (right). To convert the visual information into a soundscape, in every Luminophonics prototype the same three-layer process flow is applied. Each individual layer contains several components (as represented by the rounded rectangles inside the layer) that cooperate to achieve the functionality of the layer. Because of certain constraints, in some prototypes, different software packages and coding languages were used, but for all the prototypes the same framework and process flow were followed. By unifying the prototypes by using the same concept/framework, it was easier to manipulate and make changes to the prototype for the purpose of performance enhancement. The decision to use a modular framework instead of building each application has been proven to be useful for future work, although it takes more time to establish the framework initially. As a result, the architecture saves time and reduces some of the engineering efforts by eliminating the need to rebuild the entire application for each different prototype.

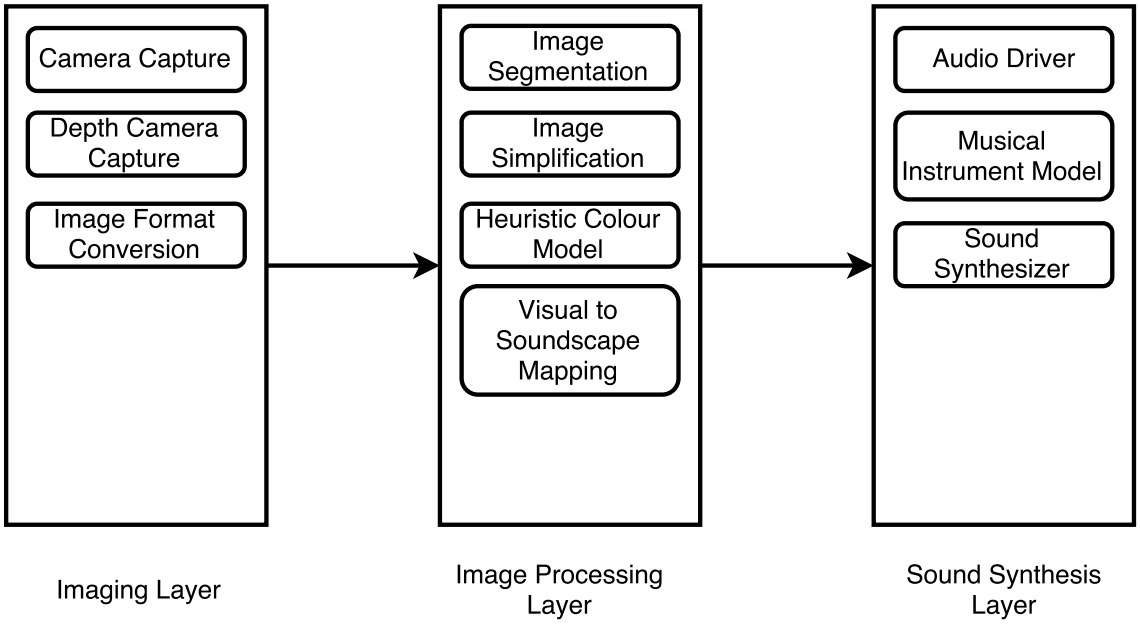


Figure 3.1: Luminophonics software architecture

Moreover, the modular framework facilitates the optimization process, which is further discussed in Chapter 5. In summary, the main purpose of using this layered architecture for Luminophonics prototypes is to facilitate the prototyping process, which allows one to make multiple minor changes to the prototypes without needing to rewrite the entire software application.

Imaging Layer

Because of the complex nature of a VASS system, attention should also be paid to the process of camera selection. Certain camera specifications, such as sensor size and colour sensitivity, may affect the general performance of a VASS system. However, given the many different cameras that are currently available, it is frequently difficult to select the camera most suitable for the VASS prototypes. Moreover, the format of the visual data varies according to the manufacturer and the type of the camera.

It is not feasible to build an image capture module for each camera that is going to be tested. Hence, decision was made to isolate the image capture module in order to create a general purpose image capture function that enables the prototypes to support a wide range of cameras. The result of the process is this imaging layer, which is both modular and robust. The sole purpose of creating a modular imaging layer was to allow the visual information captured using cameras that are widely available on the market to be used. The image data captured are then transformed into a standardized container suitable for use as the input at the next stage for further processing. The benefits of using a modular imaging layer are two-fold. It both allows most commercially available cameras to be used as the input component through common ports such as a USB and removes the need for format conversion in the later stage.

In order to achieve a modular image capture process, it is necessary to use a suitable image processing library that is not only widely acceptable, high in performance, and reliable, but also, most importantly, able to operate with a vast selection of cameras. Hence, this imaging layer relies on OpenCV [4], an open source image processing library originally written by Bradski (2000). OpenCV was designed mainly for image processing tasks, including video capture and image segmentation, with a strong focus on real-time applications, which makes it very suitable for VASS systems. OpenCV is free (released under a Berkeley Software Distribution license), and most importantly is highly efficient because the core components are written in optimized C/C++. In comparison with other image processing libraries of high-level scripting languages, such as the Image Processing Library of MATLAB and the Python Imaging Library of Python, OpenCV performs better, mainly because of the compiled code and more efficient memory management.

The use of OpenCV allows the prototypes to operate with many video sources because the library is compatible with most popular OSs (e.g., Microsoft Windows, Linux, and Mac OSX) through their internal video capturing API, VfW, AVFoun-

[4]OpenCV Official Website: http://opencv.org/

dation, FFmpeg, and GStreamer. Although this design decision means that the system can support most video cameras, the use of a universal video capture driver limits the systems' ability to utilize advanced camera features, such as faster frame rates and sensor control, which are offered only through a proprietary driver. However, it fulfilled the project's one of the important objectives, which was not to be limited to using a single camera model because it is important to evaluate as many cameras as possible to determine which is the most suitable for the prototypes and conversion process.

In addition to being able to manage data produced by the 2D cameras, the compatibility of which with the system was originally provided by OpenCV, one of the requirements of the Luminophonics project is the ability to manage also the data produced by the proprietary 3D depth sensor of DepthSense. To manage 3D depth maps, which is not within the scope of OpenCV, a custom wrapper was written on top of the OpenCV video capturing function. It acts as an automatic switch to capture 2D image data from a 2D source using the OpenCV video capture function and to obtain the 3D depth map whenever the software is connected to a depth sensor. Because the only depth sensor used is supplied by DepthSense, the wrapper utilizes the DepthSense software development kit (SDK) functions that are provided with the sensor. When the algorithm requests a depth map, the wrapper activates the depth sensor and retrieves the relevant depth map frame from it. It is fortunate that the driver supplied by DepthSense is sufficiently robust to be able to work with both Microsoft Windows and Linux, which are the main OSs on which the prototypes run. Thus, a wrapper was written to add an additional avenue to OpenCV's 'videoio' class, which provided an easy interface to allow most 2D cameras to work with the DepthSense sensor through the proprietary SDK.

An additional important requirement of the imaging layer is that it be able to convert most image data (in various formats) into a standard format that is acceptable by the image processing algorithm. The image data from commercially available cameras frequently differ in three aspects: image format, container format,

and pixel arrangement. In addition, the DepthSense depth sensor uses its own format, which is YUV2 for 2D images and an array of floating values for depth maps. While most conversions of image data are executed inside OpenCV 'videoio' into OpenCV's standard cv::Mat container through the 'videoio' class, the final image format has to be converted into the HSL colour model that is used for the Heuristic Colour Model (HCM) (see Section 3.2.3). Therefore, custom functions have to be written to convert both the cv::Mat container and DepthSense container into a container represented by an array of 32-bit floating point pixel values. It also converts cv::Mat's standard BGR pixel arrangement and also DepthSense's YUV pixel arrangement to the HSL colour model.

Image Processing Layer

For each image frame, the imaging layer outputs a container instance that contains all the metadata of the image frame, together with the raw image data that have been converted for further processing. The instance is then directly channelled to the subsequent layer, which is called the image processing layer. This layer is the most crucial layer in the entire architecture, because in this layer most of the computational tasks are executed. In the architecture diagram (see Figure 3.1), it can be seen that the image processing layer is positioned between the imaging layer and the sound synthesis layer, acting as the intermediator for the previous and the subsequent layer. Basically, it converts the visual information arriving from the imaging layer (located to its left) to the auditory modality that is synthesized in the sound synthesis layer (located to its right). In general, the input of this layer arrives from the imaging layer and its output flows into the sound synthesis layer. Hence, in order to preserve the modularity of each layer, the format of the input and output of the image processing layer is standardized to avoid information conflict if changes occur in the layers.

In total, the image processing layer contains four modules: image segmentation, image simplification, the HCM, and finally a decision engine called visual-to-

soundscape mapping. The processes in this layer depend mainly on OpenCV as the primary image processing library. They utilize several OpenCV image processing functions together to achieve their function. However, in some cases where the algorithms are not available in OpenCV, additional custom functions were written. This section does not cover the intricate details of each module in the layer. Because of their importance and complicated nature, they are further described in detail and explained in a later section individually. Furthermore, the interaction and the usage of the modules are discussed further in the sections where the conversion of the visual-to-auditory cross-modality information executed by the prototype are detailed.

Briefly, the modules in this layer are independent of each other. Each has different functions that can be combined to fill a larger role in the conversion process. Uniform across all prototypes is the decision engine, which receives the extracted visual features from the other modules and translates them into the appropriate auditory format through its mapping. The results are then packaged into an array of auditory data that are later synthesized into a soundscape in the subsequent layer.

Sound Synthesis Layer

The final part of the visual-to-auditory conversion process is the generation of sound from the converted visual information that was captured and extracted previously. Hence, the final layer of the Luminophonics architecture, called the sound synthesis layer, serves one purpose: it generates the soundscape based on the output of the image processing layer. Exactly like the other layers, this layer accepts a standardized input arriving from the previous layer. By enforcing a standardized input into this layer, the modularity of the layer is maintained, regardless of the different conversion processes used in the different prototypes.

Contained inside this layer are the three main components: an audio driver, a set of musical instrument models, and a sound synthesizer. As elements of the sound synthesis layer, they work together to fill a larger role as the audio generator for

the VASS prototypes. In more detail, the sound synthesizer uses a set of musical instrument models that is preloaded at the start of the application to generate the soundscape in the form of audio. The result (the audio signal) is then channelled to the audio driver, which interacts with the OS in order to play the audio through the speaker connected via an audio jack. Each component plays an integral part in producing the soundscape. If one of the components fails, a soundscape cannot be generated accurately, which may result in the user hearing distorted sound, or worse, total silence.

Exactly like the previous layers, the sound synthesis layer is not written from the ground up. Instead, it relies on software or a library at its core that provides most of the audio-related functions. The prototypes were designed such that they are compatible with the three main OSs (Microsoft Windows, OSX, and Linux), so that it is possible to play well the audio that is produced on any of the OSs and, most importantly, the soundscape sounds exactly the same on each of the three OSs. Hence, the audio library that was selected for this layer had to be cross-platform and able to run on the x86 architecture for the OSs mentioned. A major change was made in the sound synthesis layer during the period of prototyping, involving switching to a different core audio library. Initially, Pure Data (Pd) [5], an open source visual programming language for sound generation, was used (Puckette, 1996). In the later part of the research, the sound synthesis layer used the Synthesis Tool Kit in C++ (STK) [6] produced by the Stanford Center for Computer Research in Music and Acoustics (Cook and G. Scavone, 1999; G. P. Scavone and Cook, 2004). The decision to change the core audio library was due mainly to the performance requirement of the optimization process, which is discussed further in Chapter 5. As the format of the input to the layer remained the same, the processes of the other layers were not affected.

The decision to use Pd as the core audio library for all of the prototypes was

[5]PureData Official Website: https://puredata.info/
[6]STK Official Website: https://ccrma.stanford.edu/software/stk/

motivated by its visual programming paradigm, because it is easy to learn and sufficiently robust to allow customization. By presenting audio concepts in visual boxes, such as algorithmic functions as objects, a patching window as canvas, and data flow connectors as cords, Pd enables users to create software graphically through its GUI. The first sound synthesizer was developed by dragging and dropping visual boxes and connecting them with cords in the canvas provided. This was easily accomplished in the Pd application. Furthermore, Pd is launched as a separate process, which is beneficial for the prototype because it does not interfere with the conversion process. In other words, the sound synthesizer runs in parallel with the conversion process. In order for the two applications to communicate with each other, a networking protocol is used. The networking sound protocol is called Open Sound Control (OSC). OSC is used within the communication of the prototypes to transfer the soundscape information from the image processing layer to the sound synthesis layer. It is very similar to the musical instrument digital interface (MIDI), which is another transport protocol for musical instruments and sound synthesizers. Although both OSC and MIDI are supported for Pd, OSC was preferred for the prototypes because of its advantage over MIDI in that it allows multiple datatypes, including 32-bit integers, floating point numbers, strings, and more, to be transmitted. OSC also includes a high-precision timestamp with picosecond resolution, which is crucial for applications such as VASS systems, where speed and accuracy can greatly affect the user experience. Using the OSC protocol format, the image processing layer encodes the soundscape results into OSC data packets. The soundscape that was encoded in OSC format is transferred using the transmission control protocol (TCP) to the sound synthesis layer. Before starting the VASS system, both the applications (that contain the earlier layers) and the sound synthesis layer have to be initiated simultaneously. During the initiation, the sound synthesizer instructs the Pd engine to load all the required musical instrument models. Immediately after the musical instrument models are loaded, the sound synthesizer assumes a standby state while listening to the incoming data through the dedicated OSC network in-

terface. Upon receiving the data, the sound synthesizer begins to operate. First, it unpacks the data into time series format forming a complete soundscape. Audio is generated by reading the data chunk by chunk and the dataflow is directed to the appropriate instrument according to the information stored in the chunk. The Pd engine controls the instrument by attenuating the properties of the sound (e.g., pitch and volume) to match the instructions provided. The audio is then played on the speaker, which is connected to the audio jack of the computer.

In the latter stage, an optimization process was planned to be introduced to increase the performance of the VASS system. However, despite its benefit for the prototype development, the performance of Pd is deemed not suitable for inclusion as a part of the optimization process. The main reason for this is that Pd is launched as a separate process and the communication between the application and the sound synthesizer was through the internal network using the OSC protocol. The separation introduced a lag between image processing and sound synthesizing, which slowed down the optimization process very considerably. To eliminate this bottleneck, it was decided to merge all the three layers into a single application using a single code base. Hence, STK was chosen as the immediate replacement for Pd as the engine that powers the sound synthesizer. Because the implementation behind audio signal processing is similar to Pd and STK, the changes in the process flow were not very large. A similar process flow was ported from Pd to STK, but the layer was then coded in its entirety from the ground up using the C++ language (which is the same language used for both the previous layers). Instead of using OSC to represent the soundscape, an instance of the soundscape from the image processing layer is directly passed to the sound synthesis layer through the manipulation of internal memory. As in the process in Pd, the musical instrument models, which are preloaded, are used to synthesize the soundscape and then are played through the speaker. As a result, the soundscape can be generated more quickly and is more memory efficient, which makes it suitable for use with the optimization algorithm.

3.2.2 Hardware

Hardware is an additional important aspect of a VASS system, together with the internal software that executes the visual-to-auditory conversion. For the normal prototypes (Prototype 1–4), they are built on an x86 computer on which the three supported OSs (Microsoft Windows, OSX, and Linux) are installed. Although the software of the prototypes operates with all computers in different form factors, it is recommended that a relatively small computer (preferably containing a battery) such as a laptop be used for the sake of portability. Hence, in the experiments (see Chapter 4), the participants were instructed to carry a backpack containing a laptop computer that was running the prototype.

As the VASS prototypes are expected to be operated in real time so that the soundscape is heard immediately after an image frame is captured, the specification of the computer hardware has to be powerful. The two main aspects of the hardware that should be emphasized in order achieve a reliable real-time performance are the processor and RAM. The role of a processor is mainly to operate the software, which includes heavy computation tasks, such as manipulating image pixels and image segmentation. Therefore, it is recommended that a multi core processor with a relatively high clock speed be used. The lowest processor specification that was tested that runs well with the prototypes is the Intel® Core™ 2 Duo E6550 with 2.33 GHz. RAM also plays a role in ensuring the soundscape generation is as fast as possible. It is crucial mainly in three areas in the process: memory storage to keep the image frames that are captured before processing, as a cache to speed up image processing tasks, and finally to hold the soundscape information while it is being synthesized. Hence, the larger the memory, the better the performance of the VASS system. A minimum RAM size of 4 GB is recommended to minimize the effect of audio lag between each soundscape.

A VASS system also requires input and output (I/O) devices. The main functions of these devices are to supply the visual information to the system and to produce

the audio for the soundscape. In order to achieve these two functions, the system requires at least a camera and a pair of speakers. Because the research prototypes are designed for use by the general public, there are no strict requirements regarding the types of camera and speaker. Throughout the research activities, a range of standard USB powered Webcams was used to supply the image frames. As for the speakers, a stereo speaker is needed, because the conversion algorithms utilize the binaural feature of a sound to represent the visual information. Sennheiser over-ear headphones as the headphone of choice headphones because they can be worn comfortably on the head and, most importantly, they produce a clearer sound by minimizing the external noise from the surroundings.

Additional hardware used in the research included a TOF depth sensor to provide depth information and also a smartphone for the mobile version of the VASS prototype. These devices are described later in more detail, in their individual sections, the depth sensor in Section 3.7 on Prototype 4, and the smartphone in Section 3.8, on Mobile Prototype.

3.2.3 Colour Information

As explained in Section 1.3 of Chapter 1, one of the research goals was to include colour information in the visual-to-auditory cross-modality information. The colour information feature is an essential part of the conversion process and therefore it is shared across all of the research prototypes. The purpose of this feature is to include the colour information in the soundscape such that it can closely represent normal human colour vision. According to DeValois and Webster (2011), colour vision can be defined as the ability to differentiate the light through its wavelength composition. It is instrumental for human vision and serves multiple important purposes in our daily life. For example, colour is applied in object recognition. Humans identify an object's shape, location, and texture through its colour tones and shades. It is also useful for scene reconstruction, especially natural scenes that are cluttered with

many objects and sometimes include lighting effects such as shadows.

In the earlier VASS systems, the developers chose to discard the valuable colour information in their conversion process for several reasons, which were not limited to the lack of CPU processing power and the cost of electronics components at that time. The decision was also driven by the results of experiments, which showed that a simpler soundscape is considerably preferable because of its speed and ease of interpretation. Therefore, in the first generation of VASS systems, there was a tendency to emphasize visual features, such as texture and shape, which are expressed in greyscale images. However, it was anticipated that, with better designs coupled with improved hardware and software, the implementation of colour information will not degrade the performance of the soundscape but rather enable us to design a system that provides a balance between information richness and interpretability. The inclusion of colour information in VASS systems would allow more possibilities and improvements. The soundscape can be more varied and users can utilize the additional information to aid their interpretation. In order to include the colour information without reducing the quality of the soundscape, additional precautions have to be taken when designing the conversion algorithm. It is known that, as the complexity of the information increases, other negative side effects may be introduced in the transmitted medium, which in this case may be the cacophonous sound effect. Hence, in the prototyping phase, many possible means of overcoming the problems that arise when the soundscape is synthesized with colour information were explored.

The incorporation of colour information in visual-to-auditory conversion for the prototypes is discussed in detail in the following subsections.

Colour and Personalization

One of the problems that arise when implementing colour information in visual conversion is the consistency of colour recognition. In fact, the variance in colour perception is common across all colour reproduction devices, such as colour mon-

itors and printers. VASS too suffers this type of problem. In some of the initial tests of smaller research prototypes, the users reported many incidents related to colour inconsistency. In some of the minor tests, they reported that that some users frequently incorrectly identified the colours; in particular, they confused two colours that are very close to each other, e.g., blue and indigo or red and orange. When comparing the actual ground truth (input images) with the output soundscape, some users noted that the colour captured was misrepresented in the soundscape. In addition to the problem of the colour reproduction devices, colour constancy also affects the human perception of colour (Krantz, 2012). Briefly, colour constancy is the ability of humans to recognize a colour regardless of the colour of the light source. For instance, a green plant appears green under white daylight but reddish during a sunset when the main light source is predominately red. Therefore, the colour has to be adjusted such that VASS can maintain a stable colour appearance across light sources and users.

We hypothesize that there are two major factors that contribute to this problem. One is that colour preference differs from one individual to another. The other contributory factors are external. They consist of a slight variance in the colour captured by input devices and the surrounding illumination conditions. Therefore, a colour profile for VASS systems similar to the International Color Consortium (ICC) profile was proposed. The colour profile created by this study is called the Heuristic Colour Model (HCM). ICC profile that is introduced by the International Color Consortium is widely used to describe the colour attributes of a particular device. Although somewhat similar, the role of the HCM is to set the colour representation for VASS in the soundscape such that it matches a profile calibrated for the input device and users.

The HCM was developed based on the HSL colour model as opposed to the popular RGB colour model for both imaging and display. Although most electronic devices, such as digital cameras and monitors, use the RGB colour model, the research prototypes use the HSL colour model as the basis of the visual conversion.

The HSL colour model is preferred to the RGB colour model because it represents colour more naturally. Being a cylindrical-coordinate representation of colour, HSL defines the colour such that it closely resembles the colour as perceived by the human retina. H, which stands for the hue value, specifies the base colour in degrees (°) from 0°to 360°in a full circle. The other two values, S (for saturation) and L (for luminosity), respectively specify the saturation and brightness of the colour. S follows a scale ranging from the lowest saturated colour, 0, to the highest saturated colour, 1 (in floating points). L follows the same scale, where 0 refers to a very dark colour and 1 to a very light colour. Using the three components (hue, saturation, and luminosity), the HCM was developed by combining different threshold values for each component calibrated according to the preference of different individuals.

Figure 3.2 shows a flowchart of the colour calibration process of the HCM. The results of the colour calibration profile determine the colour based on several threshold values. The process starts with the value of S. Using two threshold values, S is used to determine whether the colour should be classified as greyscale or non-greyscale. If the S value of a colour is very low (below the lowest threshold) or very high (above the highest threshold), the colour is considered greyscale. There are three outcomes of greyscale colour: white, grey, and black. The reason for using the S value to simplify the colour into greyscale or non-greyscale is because humans frequently fail to recognize a colour if its saturation is too low or too high.

Next, the value of L is used to decide whether the colour falls in the category of white, grey, or black. This is done because, when the luminosity of the colour is too high, humans tend to regard it as white and when the luminosity is too low as black. Therefore, if the L value of the colour is below the lowest threshold, it is classified as black because it has a very low brightness, whereas if the L value is higher than the highest threshold, it is classified as white. When the value of L falls in between the high and low thresholds, the colour is determined as grey.

In the flowchart shown in Figure 3.2, the left side of the branch determines the colour through the H value. The non-greyscale colours are segregated into seven

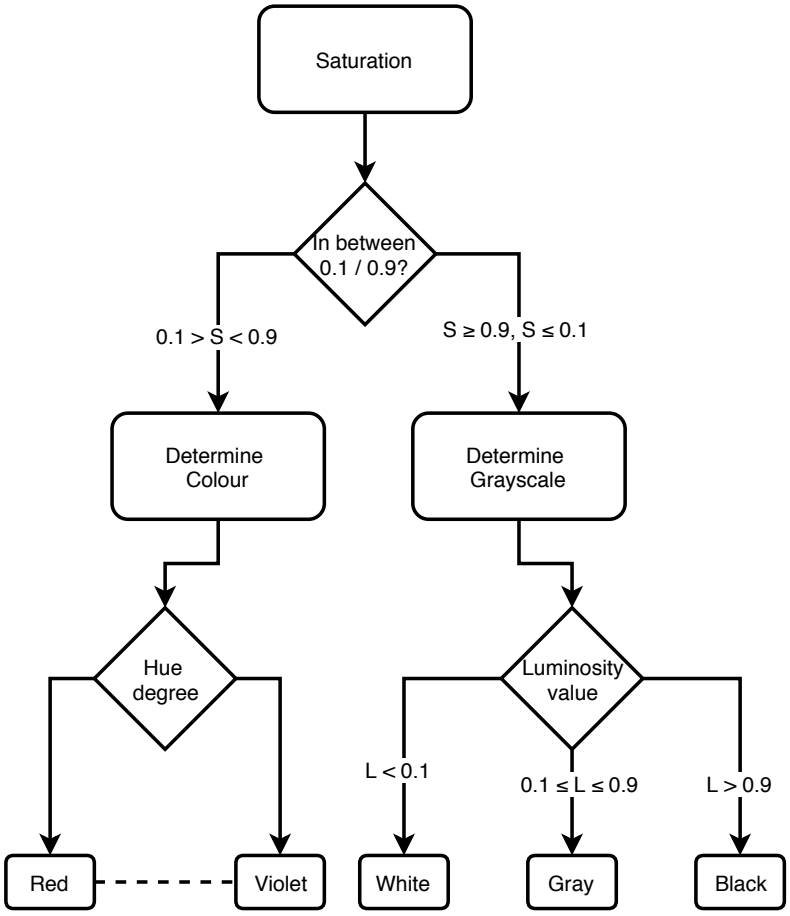


Figure 3.2: Heuristic colour model chart

basic colours according to the rainbow scale (red, orange, yellow, green, blue, indigo, and violet). Because the H value is of a circular shape, seven threshold values are used to separate the pie into seven segments, with each segment belonging to one of the seven colours. Starting from 0° until the next threshold, the colour is classified as red. The thresholding continues and ends at 360°, where the final segment belongs to violet. Using this calibration, a total of 11 threshold values are recorded as the HCM calibration profile, the role of which is to determine the colour preference of each individual.

In order to produce an HCM calibration profile, users are required to undergo a simple test assisted by a facilitator. First, a greyscale colour is presented. The facilitator slowly adjusts the brightness of the colour from white to black. One high threshold and one low threshold are recorded from the test. The second test is conducted to determine the seven segments of the hue pie chart that segregates the H values into seven different colours. Repeating the same process, the facilitator adjusts the H values from 0° to 360° while monitoring the response of the user in order to obtain the seven threshold values that separate the pie chart into segments. However, the high and low threshold values of S are not obtained through testing. The high threshold is predefined as 0.9 and the low threshold as 0.1. After the 11 threshold values are obtained, they are saved into a configuration file that is later loaded and applied in the conversion algorithm for colour determination.

Colour and Timbre Selection

After the colours are distinguished using the HCM, the visual-to-auditory conversion attempts to map the colour information into an auditory property that does not cause the user confusion when interpreting the soundscape as a result of an effect frequently referred to as cacophony. According to the Oxford dictionary, cacophony can be defined as a harsh discordant mixture of sounds. A cacophonous sound is mostly experienced by the user as noise and dissonance, which is undesirable because it confuses the user and reduces the interpretability of a soundscape. When

generating the soundscape for VASS, it is best to minimize the cacophony effect so that the soundscape presented to the user is clear and characterized by minimum noise and high interpretability.

In addition to aiming to minimize the effect of sound cacophony, Luminophonics plans to enhance the soundscape further by making it as natural as possible. Currently, in the majority of VASS systems the approach of frequency modulation by attenuating the audio signal to encode the visual information into the soundscape is applied. The result is a soundscape that sounds unnatural and that can exhaust the energy of the user if used for a long time. Listening to unnatural sounds is tiring because human hearing is not accustomed to this type of sound. The final soundscape to which Luminophonics aspires is ideally a soothing and natural sound experience similar to music performed by a musical orchestra. Therefore, it was decided that musical instrument timbres should be used as the basic sound components in the synthesis of the soundscape. This approach is not new: the developers of See ColOr proposed a similar approach, wherein different timbres are used to represent different colours (Bologna, Deville, Pun, and Vinckenbosch, 2007). In a manner analogous to See ColOr, the Luminophonics project used the 10 different colours from the HCM, i.e., red, orange, yellow, green, blue, indigo, violet, white, grey, and black, and paired them with 10 different musical instrument timbres.

Given the number of musical instrument sounds that are currently available, choosing the appropriate set of timbres that suits the prototype was a major problem. Ten different musical instruments have to selected and the set of timbres has to conform to the requirement that they should be easily interpreted. Furthermore, the individual timbres must be distinctive. If two timbres are very similar, they may contribute to the effect of sound cacophony, because it is more difficult for the user to distinguish the sounds as they clump together forming a singular sound. It is not advisable to randomly choose a set of timbres because this may hamper the interpretability of the soundscape in the event that multiple competing timbres are grouped together. Therefore, the selection of the timbre set is a difficult task and is

important considering that, if it is not accomplished correctly, it may degrade the entire performance of a VASS system.

The easiest means of building an optimal timbre set is to examine the timbres one by one and test the selection using human subjects to determine their suitability. However, this process is exhausting. For instance, with 20 potential musical instruments and 10 different colours, each subject is presented with 200 test cases. To obtain a better sample size, the 200 test cases have to be tested on a minimum of 20 users. In total, to test 20 musical instruments, the study would have to conduct 4000 test cases spread over 20 users. Conducting 4000 tests is expensive in terms of both time and money. Thus, human subject tests for timbre selection can be a major hindrance to a research project with limited resources. Therefore, an alternative method to systematically select the best timbers was proposed in which each sound signature is analysed and compared with other timbres within the set. Using this method, a set of distinctive timbres can be found. Through the application of the timbre set, the effect of cacophony is reduced significantly in the soundscape produced by the prototypes.

The method used for timbre set selection was built using MATLAB on top of a toolbox that provides functions to compute music similarity, which was developed by Elias Pampalk (Pampalk, 2004). Two algorithms, mel-frequency cepstral coefficients (MFCC) and Earth Mover's Distance (EMD), are used from the toolbox for these audio similarity measures. MFCCs are a collection of coefficients that are used to form a mel-frequency cepstrum (MFC). A full MFC is frequently used as a representation of audio because it is able to express a sound structure through its short-term power spectrum calculated from the non-linear mel scale of its frequency (Mermelstein, 1976). Because of the robustness of MFCC, which is able to detect audio signatures regardless of volume or noise, it is frequently used as an audio feature extractor for applications such as music similarity measures and audio information retrieval. EMD is the second algorithm used alongside MFCC. EMD was created to measure the minimum distance between two different probability distributions.

For this purpose, the distance between two sound signatures (in the form of MFCC distributions) is measured using the EMD formula, as shown below:

$$\int_{x=-\infty}^{\infty} |F_a(x) - F_b(x)|\, dx \tag{3.1}$$

The value of their EMD indicates the extent to which they are dissimilar. Using the EMD value, one can determine the inter-sound distance (ISD) between two timbres and then filter out the timbres that are deemed to be too close to each other.

Algorithm 1 Timber set selection pseudocode

```
for all 10 Timbres do
    Convert the timbre sound from stereo to mono
    Scale down the timbre bitrate to 11025 Hz
    Compute the MFCC of the timbre
    Perform frame clustering (FC) on the timbre's MFCC
end for
while i ⩽ 10 Timbres do
    while j ⩽ 10 Timbres do
        Compute the EMD of timbre i vs timbre j.
    end while
end while
```

Overall, the optimization of the timbre set selection comprises a two-stage process. In the first stage, similarity measurement of the audio samples is performed, and in the second, when the similarity of the timbres is below a threshold, they are replaced. The process continues until all the available timbres have been used or all the similarity measurements are well within the acceptable threshold. Initially, a random set of 10 different timbres is chosen out of the entire set of 20 timbres. An audio sample is generated for each timbre using the sound synthesizer. Similarity measurement of the generated audio samples is then performed. So that the dif-

ference measurements are focused only on the timbre signature, the volume, pitch, and duration of the sound are fixed as the control variables. Each audio sample is generated with the highest volume using La or A as the pitch playing continuously for 5 s. The two-stage process is repeated until an optimal timbre set is found. After multiple iterations of optimization, the final timbre set must contain timbres that achieve a minimal similarity that has a sound separation above a threshold. The pseudocode of the optimization process is shown at Algorithm 1.

Results of the Timbre Set Optimization Process

The proposed optimization process for timbre selection shows promising results. After four iterations of similarity measurement and timbre replacement, a set of desirable timbres is obtained. To simplify the process, the similarity of the audio samples was grouped into four levels, $\leqslant 5000pt$, $5001pt \rightarrow 7500pt$, $7501pt \rightarrow 10000pt$, and $\geqslant 10000pt$, where the lower the point, the closer in similarity are the members of the timbre pair. The goal was to eliminate timbres that have a similarity of $5000pt$ and below. The figures below show all four iterations of similarity measurement and timbre replacement for the available timbre samples. As the iterations proceeded, the number of similar timbre pairs was reduced and the similar timbre pairs were replaced by other instruments until it reached Iteration No. 4 in which most of the similar timbre pairs were replaced.

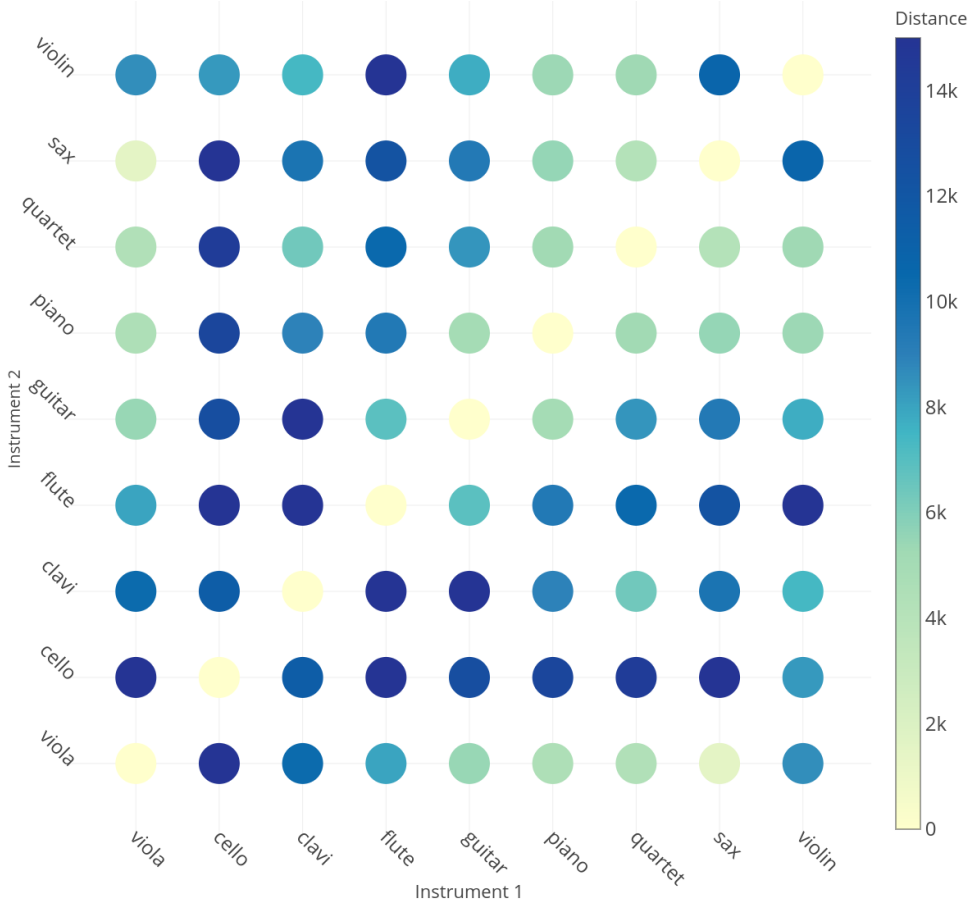


Figure 3.3: Optimizing timbre set, Iteration No. 1

Iteration 1

The viola timbre was found to be very close to at least three other timbres, including piano, quartet, and saxophone. In Figure 3.3, the similarity of the viola to the piano, quartet, and saxophone is $\leqslant 5000pt$. Therefore, it was selected for replacement in the next iteration.

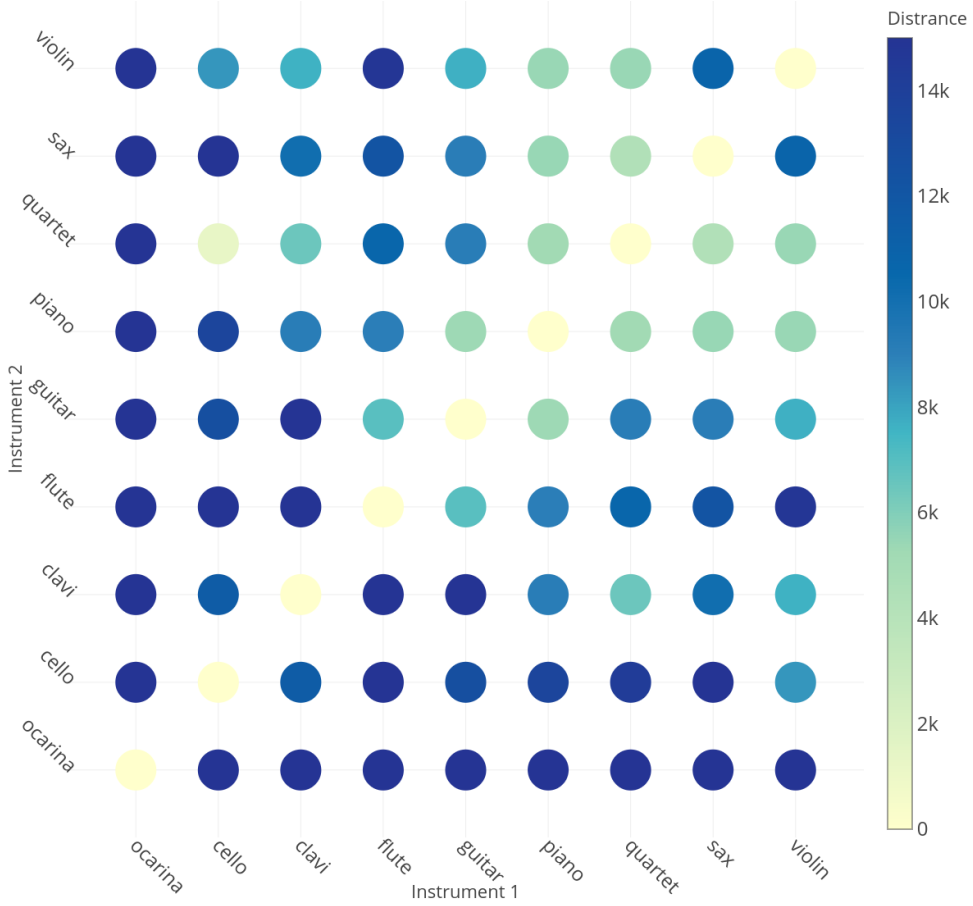


Figure 3.4: Timbre set optimization, Iteration No. 2

Iteration 2

In this iteration, the viola was replaced by the ocarina, which has a very distinctive sound, scoring above 10000*pt* when matched with other timbres. However, it was found that the saxophone is very similar to the quartet. Therefore, the quartet was selected for replacement in the next iteration.

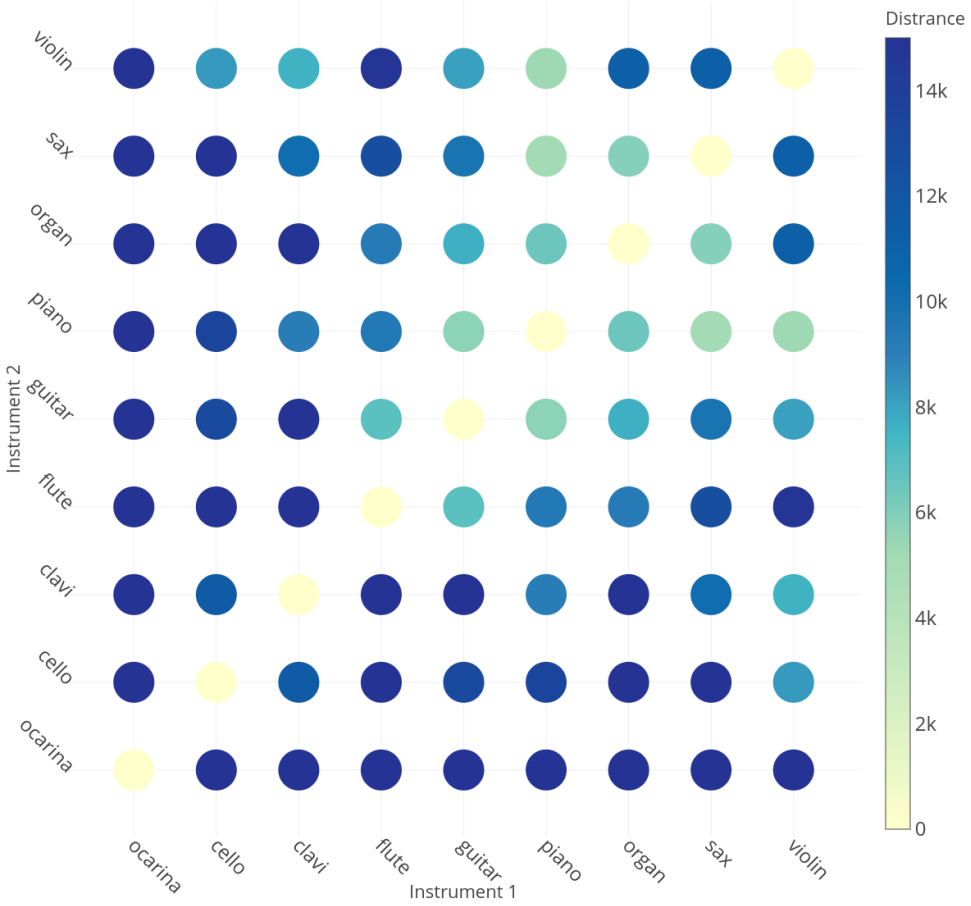


Figure 3.5: Timbre set optimization, Iteration No. 3

Iteration 3

After replacing the quartet with the organ, an improved result was obtained. However, the piano timbre was unsatisfactory because the similarity measure showed that the piano timbre is very similar to that of three other timbres (organ, saxophone, and violin), scoring very close to 5000pt.

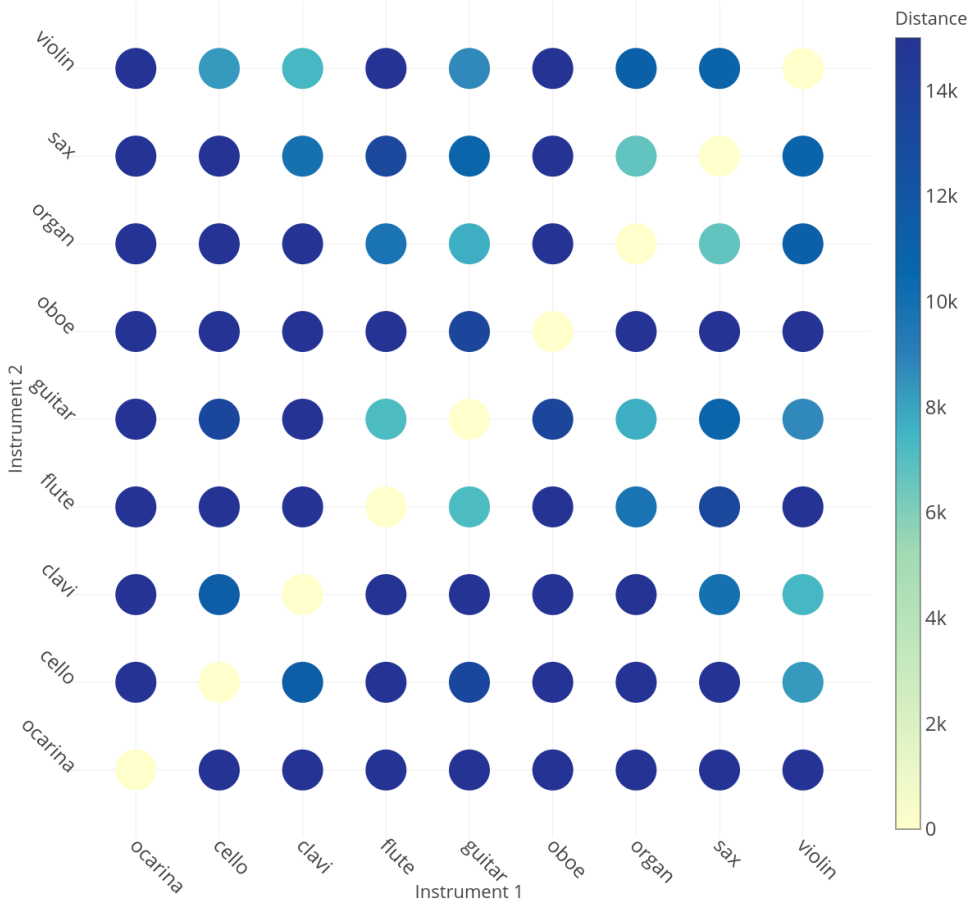


Figure 3.6: Timbre set optimization, Iteration No. 4

Iteration 4

In the final iteration, the piano was replaced with the oboe. Like the ocarina in Iteration 2 (in Figure 3.4), the oboe was a good replacement because its sound signature differs greatly from that of other timbres. After four iterations, the final timbre set was within the threshold of distance measurement.

The final set of timbres that satisfied the distance measurement is presented in Table 3.1:

Table 3.1: Final timbre set after four iterations

Timbre
Cello
Clavichord
Flute
Guitar
Harmonica
Oboe
Ocarina
Organ
Saxophone
Violin

3.3 Overview

In order to create a better VASS system through continuous improvement, a common software engineering approach called prototyping was used. A prototype is a draft version of the final product that is built to demonstrate one or more features and concepts before time and resources are invested in initiating the entire development of the concept. This approach allows implementations and examinations of the feasibility of new VASS features in prototypes before the actual final implementation. During the prototyping process, different types of prototype are built, ranging from simple designs, such as drawings and flowcharts, to smaller working prototypes that integrate both hardware and software. Prototypes allow researchers, to communicate ideas during brainstorming sessions and determine whether a certain feature will help to achieve the research goals. Adjustments to and comments about the idea on which the prototype is based can be made before moving on to the next stage. After the validity of the ideas is confirmed, features and functions are combined for development into a bigger and functional prototype. The prototyping process

eventually became an evolutionary process in which the prototypes that have the potential to perform well were selected to be developed. Prototyping optimizes the development process by helping to reduce the costs incurred. A substantial amount of time and money is saved when bad concepts are discarded before the actual development phase.

Figure 3.7 shows the overall flow chart of the prototyping process of this research project during the early phase in which multiple prototypes were built according to the flow. By following this process and continually evolving the prototypes from initial concept, a total of four working prototypes and some other smaller prototypes were produced for this research studies. In the figure, it can be seen that the ideas resulting from brainstorming are evaluated and later become prototypes. Then, the development phase began, during which software was coded and hardware was assembled to develop the concept into a working prototype. The process continued until four prototypes were achieved. The prototyping concluded with a series of experiments and analyses to measure their performance.

The following sections present the details of all the four working prototypes that were created in the Luminophonics project. Each section is divided into four subsections: Description, General Process Flow, Conversion Mapping, and Usage. The subsection on description outlines each individual prototype and the differences between them. In the section on general process flow, a flow chart is provided to illustrate the basic process flow of the particular prototype. While different prototypes have their own visual-to-auditory conversion, the conversion mapping subsection describes how the prototype converts the visual information. Finally, the practical usefulness and the benefits of the prototype for the user are clarified in the usage subsection.

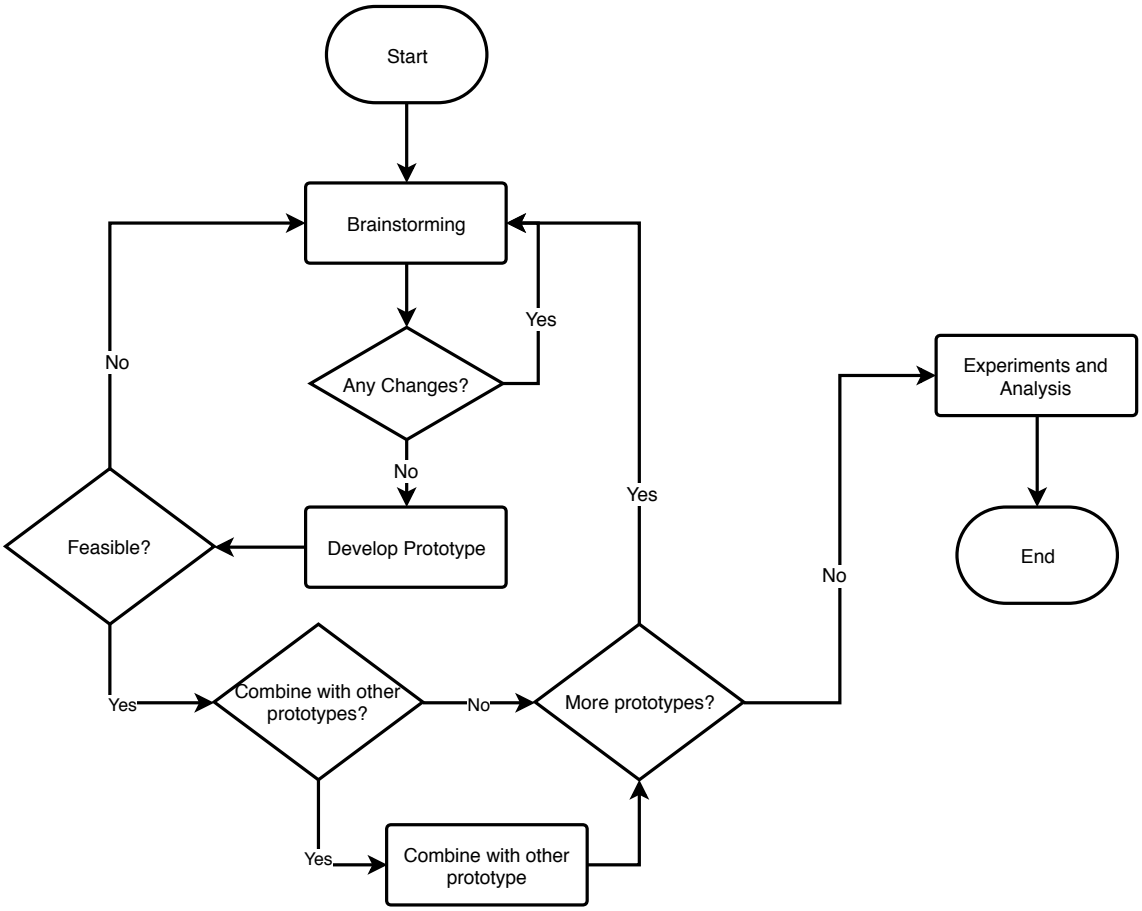


Figure 3.7: Prototyping phase

3.4 Prototype 1

At the end of the exploratory phase of the research, several potential features in the form of smaller prototypes had been achieved. A basic initial concept was formed by combining some innovative ideas that were conceptualized with several common features found in the existing SSDs, in particular in VASS systems. In order to realize the true potential of a concept, a working prototype must be built. The main purpose of building a fully working prototype is to evaluate the concept and examine the strengths and weaknesses of the features. Hence, after finalizing the ideas, the development process began by building the first working prototype. Presented in this section are the details of building Prototype 1 and an explanation of its operation.

3.4.1 Overview

Prototype 1 is the first working VASS prototype produced by the Luminophonics project for the purpose of examining the proposed features. One of the features that was examined was the incorporation of additional image processing techniques into the visual-to-auditory conversion. Recently, researchers from a similar field started to harness the potential of image processing and computer vision techniques in some of the latest VASS projects, such as See ColOr and EyeMusic. The main motivation for applying these techniques in the cross-modality conversion was to increase the feature extraction of visual information in order to improve the interpretability of the soundscape. As well as using image processing techniques to improve the visual feature extraction, the inclusion of colour information in the conversion was proposed. By including colour information, the soundscape can provide more information to the listener than can the traditional VASS soundscape, which encodes only colourless visual texture information. Finally, an additional aim of the creation of Prototype 1 was to improve the quality of the generated soundscape. To achieve this, the prototype relies on the timbres of musical instruments as the basis for its

sound synthesizer, in the hope that the synthesized soundscape would sound more natural so that it can be more easily interpreted and more comfortable to listen to.

In addition to the above three potential features, a new approach that was introduced in Prototype 1 is the usage of the swiping mechanism. In summary, the swiping mechanism implemented in Prototype 1 splits an image into multiple rows and scans the rows individually in one direction (from left to right). The purpose of applying the swiping mechanism is to introduce the concept of time delay to represent one of the visual features. Utilization of the temporal property of audio allows an additional option to be mapped to the visual features that have been extracted in Prototype 1. The details of the system and experiments were published in one of my earlier papers titled "Swiping with Luminophonics" (Tan, Maul, N. R. Mennie, and Mitchell, 2010). In the following sections, the processes of Prototype 1 and the development efforts are discussed. On the other hand, the process of evaluation is covered in Chapter 4.

3.4.2 Process Flow

Figure 3.8 illustrates the general conversion process flow applied in Prototype 1. Prototype 1 follows the general framework that is commonly used for most VASS systems. In this process, first the visual data are received from an image frames grabber unit, such as a Webcam or a digital camera. The capturing process is performed in the imaging layer, which is represented by the blue box in Figure 3.8. As explained in the previous section, this layer executes all the heavy tasks, capturing the image and processing the data into the suitable format.

After the image frames have been captured, the data are delivered directly to the next layer, which is the image processing layer (represented by the yellow box in Figure 3.8). The purpose of this layer is to extract the visual features from the image data and to translate/map them into the corresponding auditory form. Several standard image processing techniques are used in this layer, which executes

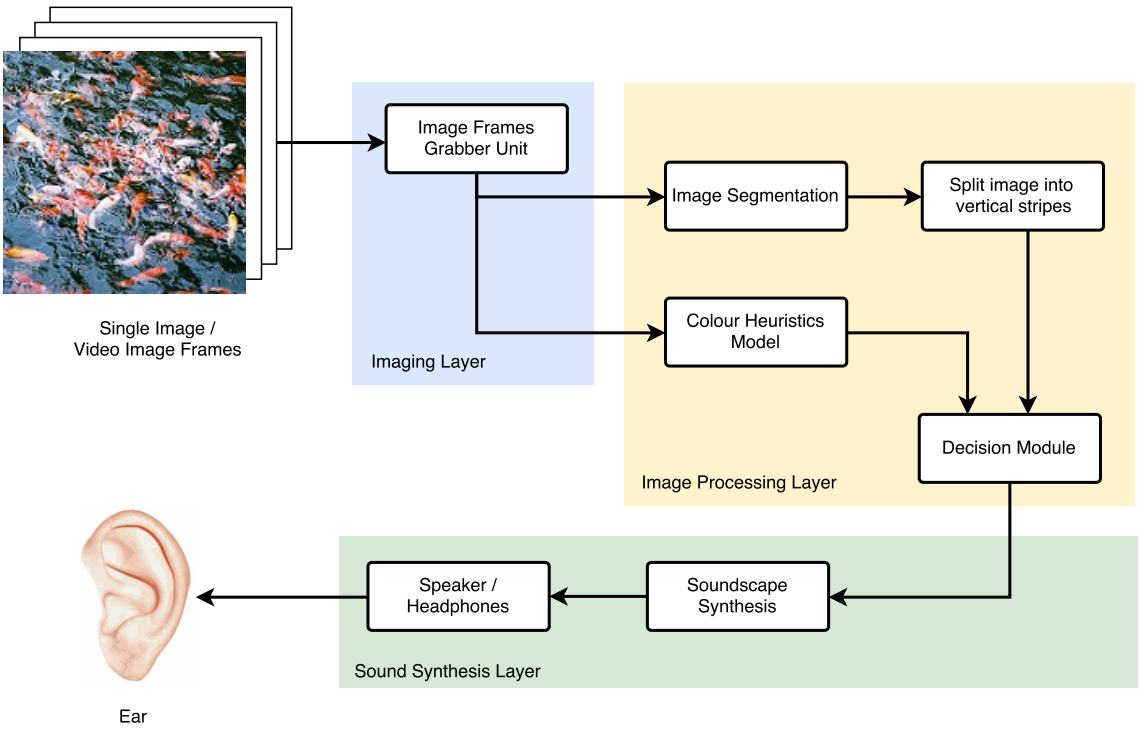


Figure 3.8: Prototype 1 conversion process

mainly image segmentation, as well as colour processing. The visual features that are extracted in this layer include colour information, location, and the size of the segmented blobs. To perform the task, the layer contains four modules, each of which is designed to manage a different task. As the image data enter the layer, they are processed by the image segmentation module, which converts the image data into several different blobs. This process is also called blobbing. The blobs then pass through the HCM (as explained in Subsection 3.2.3), which determines the colour of each blob. The locations of the blobs are determined by a third module that stripes the image into multiple rows and arranges the blobs in their specific rows for the swiping mechanism. All the information is then fed into the final module in the layer, called the decision module. This module, as its name suggests, makes the decisions for the cross modalities conversion process. It decides which visual property matches the corresponding auditory properties such that a good soundscape is produced. In addition, the swiping mechanism is implemented in this module. The decision module receives input from the two previous sources combined in a streaming format of blobs tagged with their colour code. It then decides to output an individual sound for each blob according to its size, colour, and location. Each blob has a different sound that is predetermined by the mapping coded inside this module. Finally, a soundscape is generated by combining the sound of each blob, where the blobs residing in the leftmost column are sounded first, followed by the blobs in the subsequent columns, until the rightmost column is reached. This briefly describes the operation of Prototype 1's swiping method.

In the final step, the soundscape generated from the image processing layer is transferred to the final layer, the sound synthesis layer. This layer contains a sound synthesizer that is implemented using Pd or the STK library. Before the application is started, a set of timbre models of different musical instruments is preloaded in the sound synthesizer. It then uses the timbre models to generate the soundscape according to the message passed from the previous layer. Finally, the user listens to the audio (using a speaker or headphones) created in real time by the synthesizer

through the audio driver inside the OS. Prototype 1 provides an additional function that saves the soundscape into an audio file so that it can be played again in the future.

3.4.3 Software and Hardware

During its lifetime, Prototype 1 underwent two revisions before reaching the final version. In its initial version, the software of Prototype 1 was written as two separate applications, one of which executed the imaging and image processing and the second the sound synthesis. The two applications ran in parallel.

The diagram in Figure 3.9 shows the general processes and software of Prototype 1 Beta, the first revision of Prototype 1. In the diagram, two big rectangles separated horizontally, on the left and right hand side, can be seen. The rectangles represent the two separate applications used in Prototype 1. Between them, an envelope symbol is drawn to represent the communications channel between the two applications. Stated simply, Prototype 1 comprises two applications running in parallel and communicating with each other using a common message format called Open Sound Control (OSC). The first two layers (the imaging layer and the image processing layer) run inside the first application. Using the functions supplied by libraries, such as OpenCV, the application captures image frames from the camera and translates them into a soundscape, which is contained in OSC format. The OSC message is then transferred to the second application through a TCP network channel. Upon receiving the OSC message containing the information about the soundscape, the second application starts to decode the information in the message. Using the functions provided by Pd as its core engine, it synthesizes the soundscape into its audio form. Pd also fills the role of playing the audio through the audio driver. The user listens to the soundscape in real time using the headphones connected to the computer.

Following Prototype 1 Beta, the final version of Prototype 1 was created after a

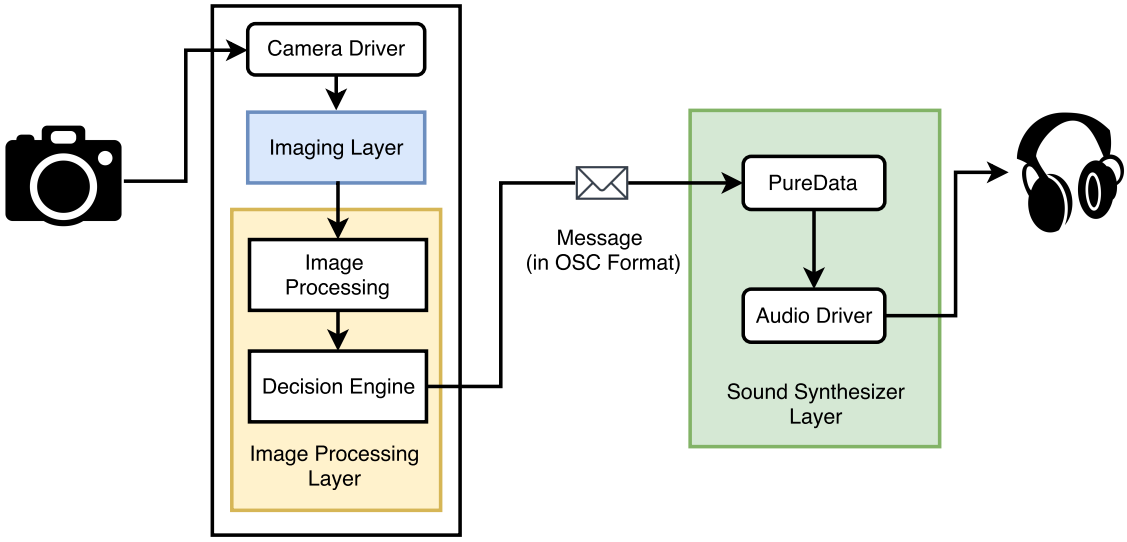


Figure 3.9: Prototype 1 Beta software

major change had been made in the software and the library used. In Figure 3.10, the diagram of the second revision of Prototype 1, which is also its final version, is presented. Although the versions produce similar results, parts of the software (in the second revision) were completely rewritten. As opposed to Prototype 1 Beta, the final version of Prototype 1 is written as a single application. All the three layers, the imaging layer, image processing layer, and sound synthesis layer, are merged into one application. A major difference between the versions is the sound synthesis layer. The final version of Prototype 1 no longer uses Pd as the core sound synthesizer. Instead, it uses a C++ sound library called the STK library as its replacement. Basically, in the context of the prototype, STK functions in the same manner as Pd as a real-time sound synthesizer. The main reason behind this major rewrite was to cater for future tasks, especially the optimization of the conversion, which is discussed in Chapter 5. Furthermore, by using STK as the audio library, the entire application can be written in a single language (C++). This increases the efficiency of all the software by eliminating the bottleneck that occurs when

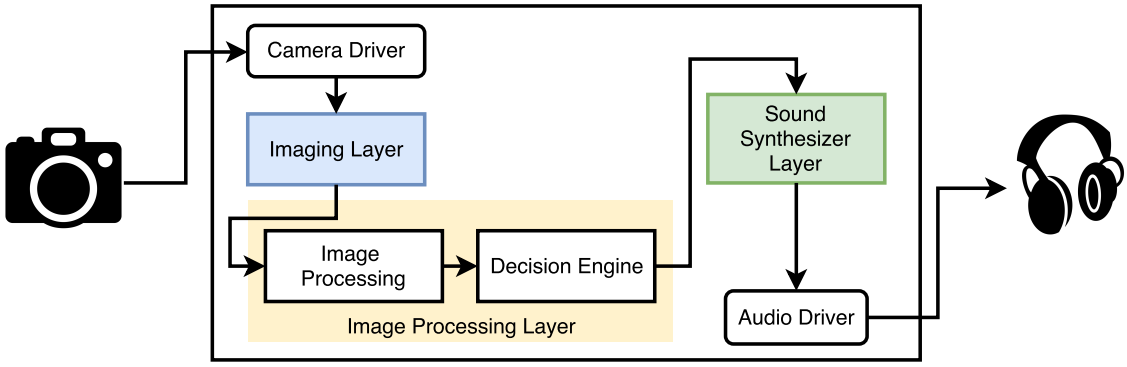


Figure 3.10: Prototype 1 software

transferring the soundscape through a TCP network channel.

3.4.4 Image Segmentation (Blobbing)

Following the previous introduction of the general process and software used for Prototype 1, this section discusses the core process of the visual-to-auditory conversion implemented in the prototype, which is the feature extraction function. This function resides in the image processing layer, which is located between the imaging and sound synthesis layers. It receives the visual data from the imaging layer and converts them into the auditory soundscape format that is transformed into audio in the sound synthesis layer. The main purpose of this feature extraction function is to obtain the relevant visual information, such as colours, textures, and shapes, so that it can be represented in auditory form. In Prototype 1, an image segmentation technique called blobbing is applied as the main feature extraction component.

Basically, upon receiving the image data, the image processing layer calls the algorithm to segment the image. The algorithm slices the image into multiple blobs. A blob can be defined as a group of image pixels that are considered to belong to the same source/object. The process is sometimes called blobbing. In order to identify the blobs in the image, the algorithm scans the image and labels the

pixels by computing their connected components as specified by the contours. A few image segmentation algorithms for digital images are available, each of which has its own strengths and weaknesses. In this image segmentation module, the main image segmentation algorithm used, called "A linear-time component-labeling algorithm using contour tracing technique", was proposed by Fu, Chen, and Lu (2004),. The component-labelling algorithm operates by tracing the contours in the input greyscale image to detect both the external and internal contours of the components. By iterating through the list of contour points, multiple blobs can be detected by identify the components that belong to the same contour and can be grouped. The main advantage of using this simple yet efficient algorithm is that it segments the image in linear time, which is very fast for a small image. The algorithm is ideal for the prototype because the image frames used are small (with a maximum of 640×480 image resolution). Therefore, although image segmentation is computationally heavy, the process operates in near real time with a minimal time lag being incurred by the entire conversion process. Furthermore, the algorithm produces additional blob properties, such as size (width and height) and exact location. The implementation of this fast image segmentation algorithm enhances the user experience of the prototype because it minimizes the lagging effect during the conversion. Because the conversion operates in near real time, it also helps the user make quicker decisions. Moreover, the additional blob information, such as its size and location, facilitates the visual-to-auditory property mapping during the conversions process.

However, the image segmentation module in Prototype 1 does not entirely depend on the connected component contour tracing algorithm. The module includes additional image processing functions for pre-processing the image in order to increase the accuracy and the performance of the feature extraction. Figure 3.11 shows that two pre-processing steps are executed before the main segmentation using connected-component contour tracing segmentation. In the pre-processing, first, the image frames are cloned and converted into greyscale colour space (as shown in

Figure 3.11b). The conversion of the colour images into greyscale images is a prerequisite for the contour tracing algorithm. However, in order to maintain the colour information for colour extraction in the later stage, the image frames are cloned before the greyscale conversion is performed. At this point, two image frames (colour and greyscale) are split from a single source.

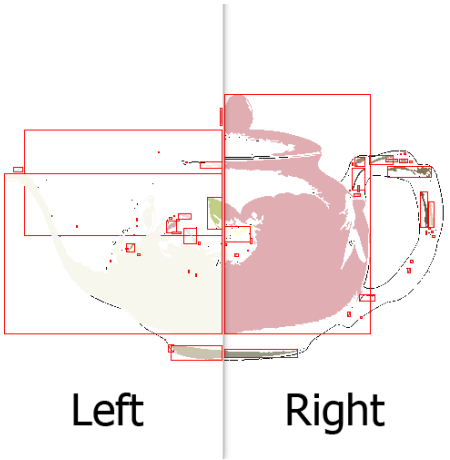
Next, a K-means clustering algorithm (with the number of clusters set to 10) is applied to the greyscale images. The purpose of the clustering algorithm is to further simplify the images by grouping similar pixels together. Figure 3.11c illustrates an example image of a teapot after K-means clustering. The K-means algorithm clusters the image into several segments and further reduces the noise in the image. If the K-means algorithm was not applied, the contour-based image segmentation method would produce a considerable amount of noise comprised of many tiny pixels that may greatly degrade the quality of the soundscape. After five iterations of K-means, the noise is greatly reduced, where some of it is seen to be absorbed into a larger segment. The post K-means binary images are then divided into two parts of equal size horizontally (left and right hand) forming two regions of interest (ROIs). The function of creating two ROIs out of a single image serves two purposes. It not only speeds up the process of image segmentation but also caters to the two channel stereo audio aspect of the prototype. In this prototype, blobs in the left and right hand regions are sonified individually in different channels. Assuming that earphones/headphones are being used, the left ear can hear only the objects present in the left hand ROI and the right ear can hear only the objects in the right hand ROI. If the object is large so that it stretches over both regions, its corresponding audio properties are played in both channels so that they are heard in both ears. This implementation allows the user to determine the X-axis of objects.

The last step in the image pre-processing phase is the generation of the contours for both ROIs. This is an important process, because the contour information is one of the dependencies of the subsequent image segmentation algorithm. The extraction of contour information is relatively fast and easy using the built-in function provided

(a) Example image

(b) Example image (greyscale)

(c) Example image (K-means)



(d) Example image (connected component segmentation)

Figure 3.11: Blobbing steps for an example image

by OpenCV library. The contour function is executed twice, once for each ROI. At the end of the pre-processing phase, two sets of contour information, together with their greyscale image data, are obtained for the left and right hand ROI of the image, respectively.

The feature extraction process ends with the image segmentation that was described above. Both the ROIs and their corresponding contour information are fed into contour-based component-labelling algorithm proposed by Fu, Chen, and Lu (2004). Blobs are then generated for both the left and right hand region. The results of the segmentations are arranged into a list of blobs according to their location in terms of a vertical starting point (Y-axis). The description of each blob is stored in the individual element of the list. It includes the size and the starting location of the blob (X and Y), and its width and height. The final piece of information about the blobs, which is their colour, is determined using HCM process. At this stage, the blobs generated by the image segmentation algorithm are mostly of a single colour with slight variations. The HCM module that was discussed in the previous section is utilized at this stage. To determine the colour of each blob, the RGB values of every pixel in the blob are averaged. The HCM module determines the colour of the blob using the averaged RGB value. The results of the HCM are appended to the list of blobs. The final result of the feature extraction is a list of blobs with their metadata, including size, starting point, width, height, averaged RGB value, averaged HSL value (from HCM), and colour.

3.4.5 Conversion Mapping

The primary visual properties that were investigated in this prototype are the colour, size, and location of objects. In order to synthesize the soundscape, the visual properties are extracted and then converted into the corresponding auditory properties. The conversion process follows a set of rules that determines the matching of the visual and auditory properties, which is called visual-to-auditory mapping. Superfi-

cially, the mapping appears to be simple and direct, but it is crucial to understand both the input and output properties that allow the prototype to maximize the information conversion to produce a good soundscape. A good and intuitive mapping both improves the learning process so that the user can start using the technology more quickly and maximizes the information retention across visual to auditory modalities, as well as increasing the interpretability of the soundscape. This subsection documents the activities and design process that are used in Prototype 1.

Figure 3.12 shows the results of the conversion. After the image is divided, each individual blob is extracted from one of the stripes. In the figure, the green blob in the 8^{th} row is highlighted. The blob then passes to the conversion mapping, which produces four information items describing the blob: colour, colour brightness, location (X-axis), location (Y-axis), and the size of the blob. Referring the example teapot image, the following sections describe the process of the conversion using this mapping.

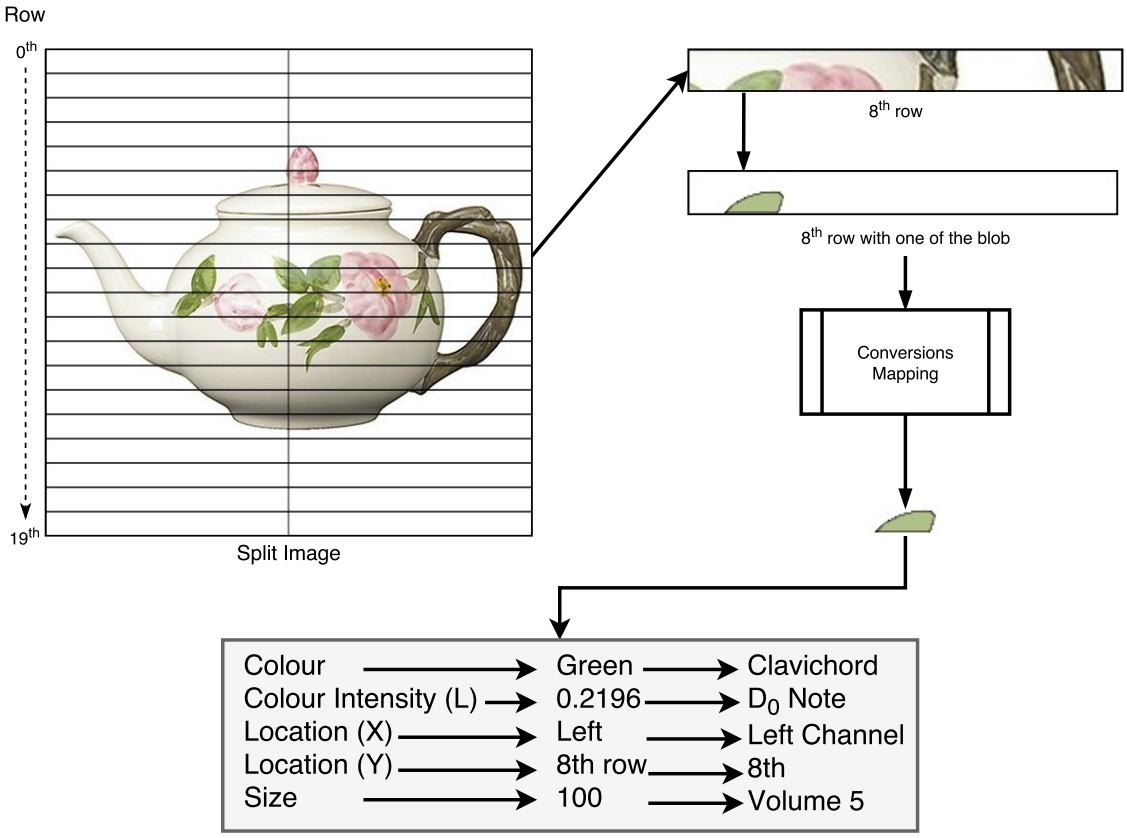


Figure 3.12: Prototype 1 conversion for example image

Colour Mapping

After undergoing the HCM process, the colour of each blob is labelled from the range of 10 colours (red, orange, yellow, green, blue, indigo, violet, white, grey, and black). To represent it in the soundscape, each colour is directly mapped to a musical instrument from the set of instrument timbres that was preloaded into the sound synthesizer. Initially, a set of musical instruments was selected based on the intuition of psychological association between timbre tone and colour (similar to the traffic light scenario, where stop is associated with red and caution is associated with amber). However, the general feedback collected from the users indicates that the

instruments produced a weak soundscape with too much overlapping sound, which resulted in a soundscape that was difficult to interpret.

This problem led to the creation of the process for optimizing the timbre set selection, discussed in Subsection 3.2.3. The general idea is to maximize the distance between each timbre by emphasizing the uniqueness of the sound signature and thus form a set of timbres that are distinguishable when played together. The result is the set of optimized timbres presented in Table 3.2. The same timbre set is used for all the prototypes in this studies.

Table 3.2: Prototype 1 Colour Mapping

Colours	Instrument
Red	Saxophone
Orange	Cello
Yellow	Harmonica
Green	Clavichord
Blue	Oboe
Indigo	Guitar
Violet	Ocarina
White	Organ
Gray	Flute
Black	Violin

Colour Mapping: Brightness

An additional benefit of using the HSL colour model as the basis of HCM is the availability of colour brightness information. The brightness of a colour is determined by the L component in the HSL value. Hence, using the value of the L component, the brightness of the blob is able to be encoded in to one of the audio properties. The decision to encode the brightness of the blob in the soundscape was made because its

brightness is a crucial aspect of a colour, in addition to the colour itself. In certain situations, decision have to be made based on the combination of the colour and its brightness. For example, a colour can be perceived as white when it is too bright. The same principle also applies in the opposite situation, where the colour can be considered black when the brightness is not sufficient. In order to better describe colour in terms of normal human vision, it is essential to include colour brightness in the soundscape.

The conversion mapping in Prototype 1 maps the brightness to the audio frequency. Because in general human hearing is effectively limited to a range of audible frequency (around 20 to 2000 Hz), the L value has to be converted to a value within this range. In order for the audio frequency to be more pronounced, the algorithm converts the L value to a range of audio notes in a musical scale. For the purpose sound synthesis, the Dorian scale was chosen. In a standard octave, the notes in the Dorian scale are C_0, D_0, $E\flat_0$, F_0, F_0, A_1, $B\flat_1$, and C_1. Table 3.3 shows the notes of the Dorian scale and the corresponding frequencies used for this conversion. In the example in Figure 3.11, the colour of the extracted blob is green when the L value is 0.2196. Therefore, the converted note is D_0, which is 73.416 Hz. Combining the colour conversion (above) and this brightness conversion, the sound of the blob is played using the clavichord in D_0.

Table 3.3: Dorian scale and its frequencies

Note	Frequency (Hz)
C_0	65.406
D_0	73.416
$E\flat_0$	77.782
F_0	87.307
G_0	97.999
A_1	110.000
$B\flat_1$	116.541
C_1	130.813

Location Mapping: X-Axis

For Prototype 1, the information describing the horizontal location (in terms of the X-axis) of a blob is reduced to only two resolutions, the left and the right hand side. This allows the sound to be directly mapped to the stereo (dual channel) audio used to synthesize the soundscape. Each channel corresponds to the ear in which the user can hear the sound. Sound that is produced in the left or right channel can be heard only in the left and right ear, respectively. Therefore, if and only if the blob resides in the left hand side of the image, is it heard in the left channel of the soundscape. The same applies to blobs located at the right hand side of the image: they can be heard only in the right channel of the soundscape. If the blob is so large that it spans across both sides of the image, it can be heard in both the left and right channel and is thus played in both ears. In the example in Figure 3.11, the green blob is located at the right hand side of the image, and therefore, the sound of the blob can be heard only in user's right ear.

Location Mapping: Y-Axis

The swiping mechanism proposed for the prototypes is used to encode the vertical location (in terms of the Y-axis) in the soundscape. Because Prototype 1 swipes from the top to the bottom, creating one horizontal stripe after the other, the vertical locations of the blobs are translated into temporal information in the soundscape. To simplify, the blobs in the first stripe are heard first and subsequently the blobs in the stripe below that stripe, until the swipe reaches the end of the image. The start time (t_n) of the blob sound is calculated as

$$t_n = \frac{T}{N} \cdot n \tag{3.2}$$

The start time (t_n) of the blob in the n^{th} stripe is the total time of a soundscape (T) divided by the total number of horizontal stripes (N) times the index of the stripe (n). In the example in Figure 3.11, the blob is located in the 8^{th} row. If the total duration of the entire soundscape of an image is 2 s, the blob can be heard 800 ms after the sound of the image. This is because there are a total of 20 horizontal stripes in the image, and therefore 100 ms is allocated for each stripe according to Equation 3.2.

Size Mapping

When the blobs are extracted from the image frames, the component labelling algorithm supplies the size of every blob by calculating using the coordinates of the rectangle bounding box. Each blob's size, S, can be obtained from the two coordinates of opposing corners, the top right hand corner (x_0, y_0) and the bottom left hand corner (x_1, y_1) using

$$S = (x_1 - x_0) \cdot (y_1 - y_0) \tag{3.3}$$

For Prototype 1, it was decided that the size of a blob should be mapped to the sound volume. In order to describe the size of a blob, the two elements (the blob's

size and the sound volume) are directly proportional. The intention was that the larger blobs would sound louder than the smaller blobs. However, the volume range is normalized within a limited sound range to ameliorate the problem that the sound of a bigger blob covers the sound of smaller blobs. Ten levels of sound volume are used, from 0 (no sound) to 9 (maximum volume). The sound volume of the blob is calculated using its size as compared to the total size of a single horizontal stripe. For example, if the total size of a horizontal stripe is 800 pixels and the blob size is 200 pixels, the sound volume of the blob will be 3. The result is obtained by rounding up the value (2.5) from the fraction of $\frac{200}{800}$ multiplying by 10 volume levels.

3.4.6 Usage

Prototype 1 must be used with a personal computer, whether a desktop or a laptop. It can be used with or without a camera attached to it. However, without a live camera, the application can sonify only when loaded with a static image. In this mode, the user puts on the earphones/headphones, opens the application, and loads an image. The computer plays the audio of the soundscape in the earphones/headphones when it has completed the conversion.

In the live mode, Prototype 1 must be tethered to a USB-powered camera. It operates with any Webcam or the DS311 depth sensor produced by DepthSense. Normally, in experiments the camera is tethered to a laptop computer because it is lighter and easier to carry. When in live mode, the camera constantly feeds the captured image frames to the application. The application reads the image frames and converts them into the soundscape serially. Between the frames, a short monotone sound is played to signify the end of the current frame while the application is processing the next frame. As usual, the user listens to the audio of the soundscape through the earphones/headphones.

To interpret the soundscape, the user needs to concentrate on the sounds heard in both the left and right ear. The blobs at the right hand side appear in the right

sound channel and the blobs located at the left hand side appear in the left sound channel. As Prototype 1 swipes from top to bottom, the blobs that appear first are located at the top. Subsequently, the blobs appear one by one depending on their location on the Y-axis. To pinpoint the location of the blob approximately, the user first determines the Y-axis and then determines whether the blob appears on the left or right hand side. If the blob appears on both sides of the sound channel, the blob stretches over both sides. After obtaining the location of the blobs, the user can interpret the remaining details of the blobs according to their mapping. For example, the colour is determined by the type of timbre, the shade by the tone used, and the size by the volume of the sound.

In a practical scenario, two options are provided for the position of the camera. The user is free to choose whether to mount the camera on top of his/her head or to hold the camera in his/her hand. Each option has its own strength. The head-mounted position provides vision that is closer to that which the human eye sees, whereas the handheld position provides more degrees of freedom because the user can swing the camera. Whichever camera position is chosen, the usage is similar. The user points the camera to the direction he/she intends to visualize and the soundscape is played in his/her ears. One advantage of mounting the camera on top of the head is that it frees up the user's hand to do other things. When the soundscape of the visual frame has been played, the direction of the camera is changed slightly to capture a slightly different angle of the scene. While listening to the soundscapes, the user gradually builds a mental image of his/her surroundings by interpreting and comparing the soundscapes.

3.5 Prototype 2

3.5.1 Overview

Prototype 2 is the second iteration in the Luminophonics project prototype series. It was derived from Prototype 1, most of the basic processes of which were retained with some changes to introduce an additional new feature. Although the two prototypes are similar in terms of basic functionalities, Prototype 2 was created to examine a major shortcoming in Prototype 1: the lack of resolution of the blob's horizontal position. In Prototype 1, the information of the blob location is translated to temporal delay to represent the vertical position and to the audio channel in the stereo audio to represent the horizontal position. As an image is split into 10 separate stripes, the vertical position of a blob has 10 possible discrete locations. However, because the prototype relies on stereo audio, there are only three possible options for the translation of the horizontal position. The blob can be heard in the left audio channel if it is located at the left hand side of the image or otherwise in the right audio channel. It can also be heard in both channels if the blob stretches from the left to the right hand side of the image. In total, the blobs in Prototype 1 have only three possible horizontal positions: left, right, and centre. Superficially, the interpretation can be very simple, but it can be imprecise. The main objective of the development of Prototype 2 was to address the problem by increasing the number of possible horizontal positions of the blobs to improve the precision of their location.

In order to increase the number of horizontal positions, Prototype 2 implements a concept called the head-related transfer function (HRTF) (, which takes advantage of humans' ability to localize sound through their hearing. This amazing ability of humans allows them to locate a sound source by listening to the volume difference in the binaural sound. This technique has been commonly applied and exploited in the entertainment industry to create an immersive theatrical experience. Popular cinema halls are frequently equipped with multiple speakers (5.1 or 7.1) of all sizes

located around the audience, which play multi-channel audio to simulate sound coming from different surrounding locations. The concept of applying HRTF in VASS systems is not new: Bologna and Vinckenbosch explored this idea in one of their research projects, namely, Ambisonic 3D-sound field (Bologna and Vinckenbosch, 2005). The Ambisonic 3D-sound field device utilizes sound localization to simulate a 3D surround sound environment by locating individual sounds through augmentation according to the location from which the objects are coming. By means of this immersive surround sound, the user can approximately determine the source location of objects. By exploiting this human stereo hearing ability, sound localization may have a large potential for improving visual-to-auditory conversion.

A simple HRTF is implemented in Prototype 2. Basically, the volume of each blob is varied according to its horizontal position. Thus, more resolution is added to the horizontal representation by using the volume resolution. However, by associating the volume to the horizontal position, Prototype 2 lost the ability to represent the blob size by volume, which was a feature of Prototype 1. This overview describes the brief implementation and the intuition behind Prototype 2. The details are further elaborated in the 'Conversions Mapping' section (Subsection 3.5.4).

3.5.2 Process Flow

The process flows of Prototype 1 and Prototype 2 are similar, as shown in Figure 3.8. First, the image grabber captures the current image frame and stores the data in the image container for subsequent operations. Then, an image processing module, inherited from Prototype 1, proceeds to process the image data, producing blobs and colour information. It is noteworthy that it is the subsequent part of the process that distinguishes Prototype 2 from its predecessor. The information from the image is split into multiple horizontal stripes according to its location. The soundscape is generated by swiping each horizontal stripe top-down, matching each visual property to its corresponding audio properties. The resultant soundscape (in

the form of an array) is then transferred to a sound synthesizer so that the audio can be played by means of the audio driver in the computer. A headphone/speaker connected to the computer at the audio jack plays the audio to be heard by the user.

3.5.3 Image Segmentation and Blobbing

In Prototype 2, the same blobbing method is implemented as in Prototype 1, as described in the previous section (Subsection 3.4.4). Since the swiping direction of the two prototypes is the same, the manner in which the image is divided is also the same. In both prototypes, the segmentation function divides the image into two vertical columns (left and right). The two columns are then further divided horizontally into multiple stripes for the top-to-bottom swipe. However, the major differences between Prototypes 1 and 2 lie in the conversion mapping, which is described in detail in the following section.

3.5.4 Conversion Mapping

The differences between Prototypes 1 and Prototype 2 lie in their conversion mapping. They are revealed by a side by side comparison of Figures 3.12 and 3.13. The figures illustrate the conversion of each prototype performed on the same example image. The results of the conversion are clearly different because of the changes effected in Prototype 2. One obvious difference in the Prototype 2 conversion results is the number of conversions: Prototype 2 encodes only four types of information, whereas Prototype 1 encodes five types of information. The types of information used in Prototype 2 are colour type, colour intensity, horizontal location (X-axis), and vertical location (Y-axis). The absence of blob size in the soundscape produced by Prototype 2 is compensated by emphasizing the horizontal location of the blob. Prototype 2 was built such that the horizontal location is defined better by adding more positions to the X-axis information as compared with only three positions

(left, right, and centre) used in Prototype 1.

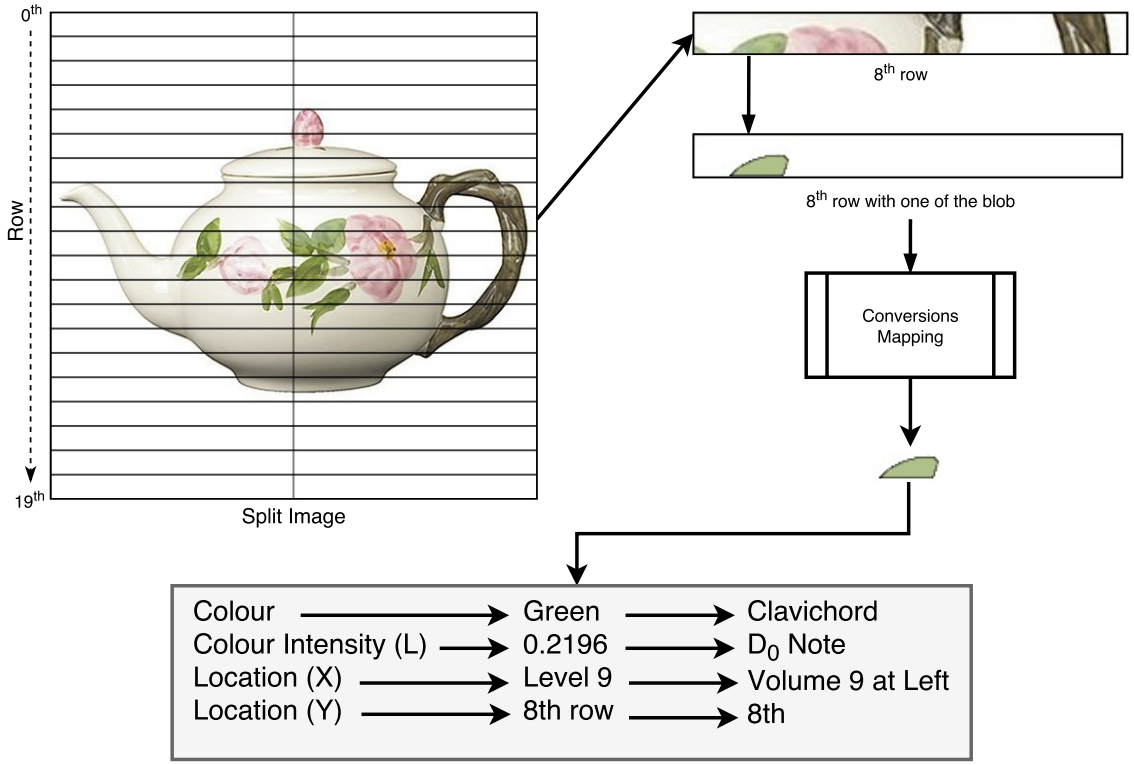


Figure 3.13: Prototype 2 conversions for example image

Colour Mapping

Prototype 2 inherited the colour-to-instrument mapping from Prototype 1. The details of the colour conversion are discussed in Subsection 3.4.5. The same colour-to-instrument mapping shown in Table 3.2 is used.

Colour Mapping: Intensity

Because Prototypes 1 and 2 use the same colour mapping, the mapping of colour intensity to audio frequency remains the same. Prototype 2 converts the colour intensity to the Dorian scale, as shown in Table 3.3. As illustrated in Figure 3.13,

the result for the green leaf in the teapot is translated as a D_0 note played using the clavichord.

Location Mapping: X-axis

As mentioned, the implementation of the horizontal location information of the blob (in terms of the X-axis) is the major difference between Prototype 2 and Prototype 1. The reason for this change is that the horizontal location of a blob is poorly represented in Prototype 1. The horizontal information of the blobs in Prototype 1 allows three positions (left, right, and centre). Changes were made in Prototype 2 to increase the horizontal resolution in order to increase the accuracy when sonifying the location of the blobs. A more precise location representation greatly enhances the conversion of the features of an image frame.

To increase the horizontal resolution of the location, a different sound property from that used in Prototype 1 is utilized. This was because the current implementation (representing horizontal location with the stereo channel) cannot accommodate this information expansion. Therefore, in Prototype 2 the support for size mapping was omitted so that the volume property would be available. Moreover, by representing horizontal location by sound volume, a user experience that is similar to HRTF can be created when the user listens to the soundscape.

The sound volume of a blob is calculated using its distance from the centre of its horizontal stripe. The closer to the centre the blob is situated, the higher is the sound volume of the blob. Similarly, if the blob is situated at the far end of the stripe, its sound has the lowest volume. The sound volume, V, of a blob is calculated as

$$V = 10 \cdot \left\| \frac{\frac{W}{2} - \left| x - \frac{W}{2} \right|}{\frac{W}{2}} \right\| \tag{3.4}$$

Figure 3.13 shows that the volume of the green leaf blob according to result of the equation above is 9. Its volume is 9 because the horizontal location of the blob,

x, is pixel 340 while the centre of the stripe is pixel 400, making the distance of the blob from the centre 60 pixels. With this distance, the blob falls in the second of the 10 slots, and therefore it has the second highest volume (9). Finally, because the blob is at the right hand side of the image, it is heard in the right channel of the soundscape.

Location Mapping: Y-axis

Finally, in Prototypes 1 and 2 the same swiping technique is implemented: they both swipe from top to bottom. Therefore, the vertical location of the blob is translated to temporal information in the soundscape. Stated simply, blobs that appear in the first row of the image are heard first. Subsequently, the blobs in the next row are heard, and so on until the swipe reaches the final row in the image.

3.5.5 Usage

The differences in terms of usage between Prototypes 1 and 2 are minimal. They both have two modes, a static image mode and a live mode. The user is still required to wear earphones/headphones to listen to the soundscape. The general process is the same for the two prototypes: the prototypes receive visual data and play the converted soundscape through the speaker. However, the user needs to relearn the interpretation of the soundscape for Prototype 2.

In Prototype 2, the volume of the sound is no longer used to represent the size of a blob. Because the size of the blob is not represented in this prototype, users have to reinterpret the information using their mental map. The lack of size information is compensated by the increased resolution of the blob's horizontal position. The volume of sound is reassigned in Prototype 2 to represent the horizontal position of the blobs. Previously, in Prototype 1 the blobs had only three positions (left, right, and centre). For instance, a blob that is played in the left channel is situated on the left hand side of the image, whereas a blob that is played in the right channel is

situated on the right hand side of the image. The difference in Prototype 2 is that the user is required to listen closely to the volume to determine the location of a blob. The volume of the sound now inversely correlates to the calculated distance of the blob's horizontal position from the centre point of the image. In other words, the louder the sound volume of the blob, the closer it is to the centre of the image. For example, when the user hears the sound of a blob at low volume in his/her right ear, this indicates that the blob is located at the right side of the image far away from the centre.

To conclude, the usage of Prototype 2 remains similar to the general approach, where the user points the camera to the intended direction and allows the prototype to sonify the entire scene. However, Prototype 2 offers a soundscape with a higher horizontal resolution than Prototype 1 because the horizontal information is translated to the volume.

3.6 Prototype 3

Unlike Prototype 2, Prototype 3 is not a direct evolution from its predecessors. There is a major difference between it and both Prototype 1 and Prototype 2. While Prototype 3 still retains the swiping mechanism originally implemented in Prototype 1, a much simpler image segmentation technique is applied for visual feature extraction. In Prototype 3, the contour-based blobbing technique for segmenting the input image into objects/blobs is not used. Instead, an approach for resolution reduction similar to that used in many of the first generation VASS systems (e.g., vOICe) is used. The source of the idea behind Prototype 3 was many of the earlier VASS systems, where the focus was on producing soundscape from raw visual features, such as pixels and colour intensities, rather than from blobs. The purpose of implementing a much simpler algorithm for image segmentation instead of the complex contour-based segmentation technique used in Prototypes 1 and 2 was to determine the advantages and disadvantages of both techniques through comparison. Each

technique has its own unique features: the blobbing technique produces a simpler soundscape focusing on each object as a whole but sacrifices information, such as detailed features. Although the soundscape based on raw pixels may retain much of the visual information, it may degrade the interpretability of the soundscape. By comparing the two techniques, the effect of each on the user experience in terms of interpretability and learnability and also on the robustness of the conversion can be determined. Using the results of the comparison, more insight into how to build a better image processing algorithm that is suitable for a VASS system was obtained and then further elevates its performance. This section describes in detail the internal operation and the development of Prototype 3 and the inspiration behind it.

3.6.1 Overview

Prototypes 1 and 2 utilize a contour-based image segmentation technique to extract the features of the objects/blobs in an image. The soundscape is then generated based on the information of the extracted blobs. This mimics human vision as closely as possible because humans recognize a scene through the objects in the frame. By using this approach, the power of the computer is harnessed to process the image in advance to lighten the interpretation burden of the user.

The VASS devices of the earlier generation (i.e., vOICe and PSVA) applied a much simpler image segmentation technique, emphasizing human interpretation of the soundscape instead of computer assisted interpretation with image processing. They do have certain merits, and thus they sometimes outperform Prototypes 1 and 2 in certain situations. Because they directly map the properties of the pixels to the audio properties, the devices are able to maintain most of the visual properties, such as the features, shapes, and textures. In situations that greatly favour visual features (e.g., object recognition), such a VASS system shows a very good performance. Users are able to utilize the features to recognize an object through its shape and texture.

As compared to the prototypes that use contour-based segmentation, many visual details were discarded during the image processing phase in favour of the shape as opposed to the texture. Although the soundscape sounds simpler and is more easily interpreted, the user finds it difficult to use it in a situation that requires finer details. For example, a user could not easily distinguish between two objects of the same shape, such as a basketball and an orange. The detailed results and further explanations of the comparison are presented in Chapter 4.

The source of an additional inspiration that encouraged us to pursue direct pixel mapping was the modern MIDI controllers that are widely used by soundmixers and disc jockeys, such as Ableton Push[7] and UNTZtrument by Adafruit[8]. Images of the two controllers are shown in Figure 3.14. It can be seen that the devices feature multiple buttons arranged in a square shape, where each row can be programmed as a different instrument and each column can be specified at a certain time interval. A disc jockey composes a tune by pressing the buttons so that the instruments are played in a specific timeframe in a desired key. Connected to a sound synthesizer such as the Ableton Push, the MIDI controller then sweeps from left to right, column by column, repeating the beats to form a tune. It should be noted that the swiping mechanism proposed in this study implemented in the previous prototypes operates in a similar fashion. This one-direction swiping either from left to right or from top to bottom utilizes the time factor to play different keys to form a tune/soundscape. Thus, by re-imagining each image pixel as an individual button on the MIDI controller, a soundscape can be synthesized using the same procedure as implemented in the MIDI controller.

[7]Information on Ableton Push can be found at: https://www.ableton.com/en/push

[8]Information on UNTZtrument can be found at: https://learn.adafruit.com/untztrument-trellis-midi-instrument/overview

(a) Ableton Push (b) UNTZtrument produced by Adafruit

Figure 3.14: Images of popular modern musical instrument digital interface controller

3.6.2 Process Flow

Figure 3.15 shows the conversion process of Prototype 3 from capturing the input images to synthesizing the soundscape. The main difference between this process and that implemented in Prototypes 1 and 2 is the absence of an image processing module. This module is replaced by a simple function to reduce the resolution of the image from the original resolution, which depends on the type of camera used, to a resolution of 20×20. The HCM, which previously ran in parallel with the image processing module, now operates after the image resolution is reduced. The reason for locating the module after the resolution reduction is that the HCM operates by looping each pixel row by row, which is very time consuming when applied on a high resolution image. Hence, after resolution reduction the process of colour determination using the HCM is faster and less heavy.

When the application has been initiated, it captures image frames using the camera connected to the computer. The frame grabber then converts the visual data into the colour model accepted by the OpenCV library and stores it inside a frame container. An instruction is then transmitted to the frame container to reduce the resolution of the image data inside it. Then, the frame container passes

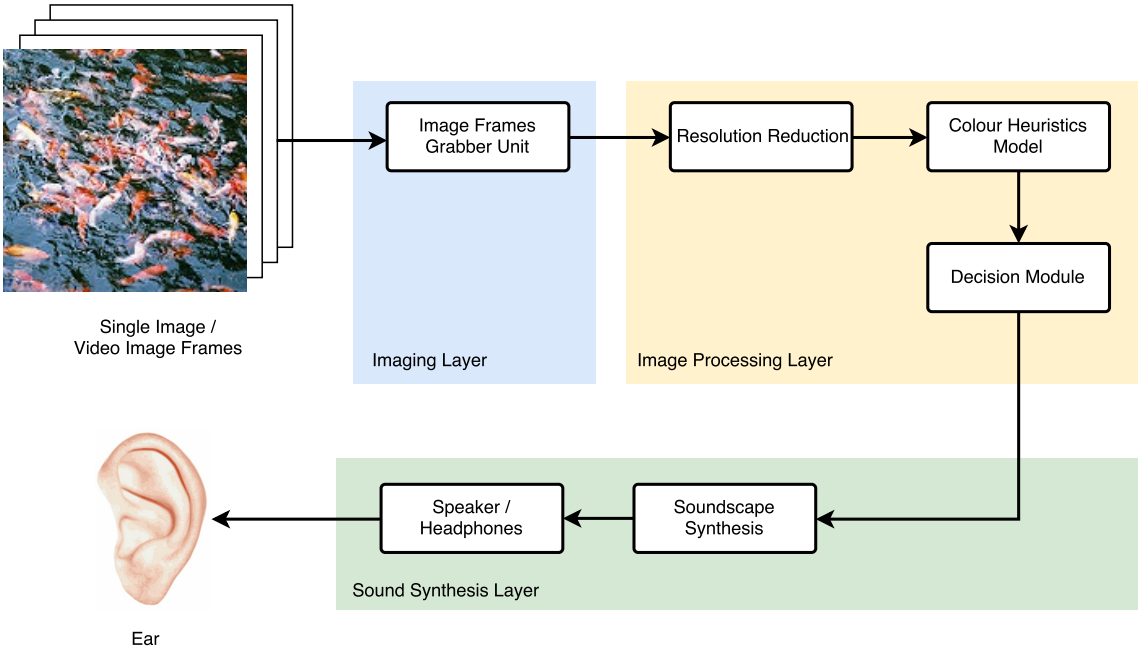


Figure 3.15: Prototype 3 conversion process

through the process of the HCM. Additional data, which contain an array of colour information based on each pixel, are embedded into the container after the process. The decision module then processes the image data together with the array of colour information to transform them into a time-series structure. The sound synthesizer, in which the musical instrument models are loaded, uses the time-series structure to generate a soundscape accordingly. Finally, the user can listen to the soundscape being played on the computer using headphones plugged into its headphone jack.

3.6.3 Image Segmentation/Pixelation

As compared to Prototypes 1 and 2, Prototype 3 takes a step backward in that it uses a much simpler visual feature extraction method to focus on the texture and the details of an image instead of on the shapes and objects. However, the outcome would be very poor if the conversion mapped every pixel in the image to a sound.

The resulting soundscape would be very noisy and difficult to interpret. In order to solve this problem, the input image frames must be simplified so that the conversion algorithms can produce a better soundscape. I referred to vOICe for the solution (Meijer, 1992). Meijer (1992) proposed down-sampling an image by grouping the neighbouring pixels to form a larger pixel. The process is called pixelation.

Pixelation is a method of scaling down the resolution of the image by applying a filter to obtain the value of a central pixel by averaging the values of its neighbouring pixels. Through pixelation, although the resolution/quality of the image is reduced, the approximate texture of the content is still retained. Therefore, Prototype 3 uses the technique of pixelation to reduce the amount of pixels to be converted while maintaining the texture information of the image. To maintain the aspect ratio of an image, Prototype 3 reduces the input image using a fixed size average box filter. For example, if a 30×30 box filter is applied to an input image with a resolution of 640×480, the input image is down-sampled to a resolution of 22×16. Therefore, instead of 307300 pixels, Prototype 3 now converts only a total of 352 pixels. To maintain the texture and colour, the value of each pixel of the pixelated image is the average value of the 30×30 pixels of the original image. Because the colour image is composed of three channels (red, green, and blue), each channel is averaged individually.

Figure 3.16 shows an example image before and after pixelation. A comparison of the images on the left and right hand side reveals that the image after pixelation (right hand image) has a significantly lower number of pixels. As for the conversions, every pixel is mapped to each sound according to the conversion mapping. Each pixel contains its information, such as location and colour. The location of the pixel is expressed as the X and Y coordinates in the image. The colour of the pixels is determined by the HCM module, which takes the HSL value converted from the RGB value of the pixels. Because Prototype 3 swipes from left-to-right horizontally, the final result of this feature extraction is a list of pixels structured by arranging them column by column.

(a) Example image (b) Example image (pixelation)

Figure 3.16: Pixelation of example image

3.6.4 Conversion Mapping

There is a major difference in Prototype 3 in terms of the segmentation algorithm: it uses the pixelation technique instead of the blobbing technique that was used by Prototypes 1 and 2. As a result, the emphasis of this prototype is on the image texture rather than on the shape of objects. This also affects the conversion mapping: the location mapping is completely different from that of the earlier prototypes. Instead of the location of the blobs being sonified, the soundscape now has to encode the location of every pixel produced by the pixelation. In addition to the segmentation, the swiping direction implemented in Prototype 3 was changed. Prototype 3 swipes from left to right, column by column, instead of top-down as the previous two prototypes did.

Figure 3.17 shows the conversion process for an example image of a teapot. As compared to the previous conversions, illustrated in Figure 3.12 and Figure 3.13, the features extracted by Prototype 3 have been reduced. Prototype 3 simplifies the conversion mapping further. The supported features for conversion are the colour type and location (for both the X- and Y-axis). In summary, Prototype 3 focuses on image pixels instead of on the shape of the objects. Although it extracts

fewer features from the image (i.e., size), it sonifies the entire image pixel by pixel. Prototype 3 empowers users by allowing them to reconstruct the image entirely from the soundscape, so that they can interpret the information themselves by visualization using their mental map.

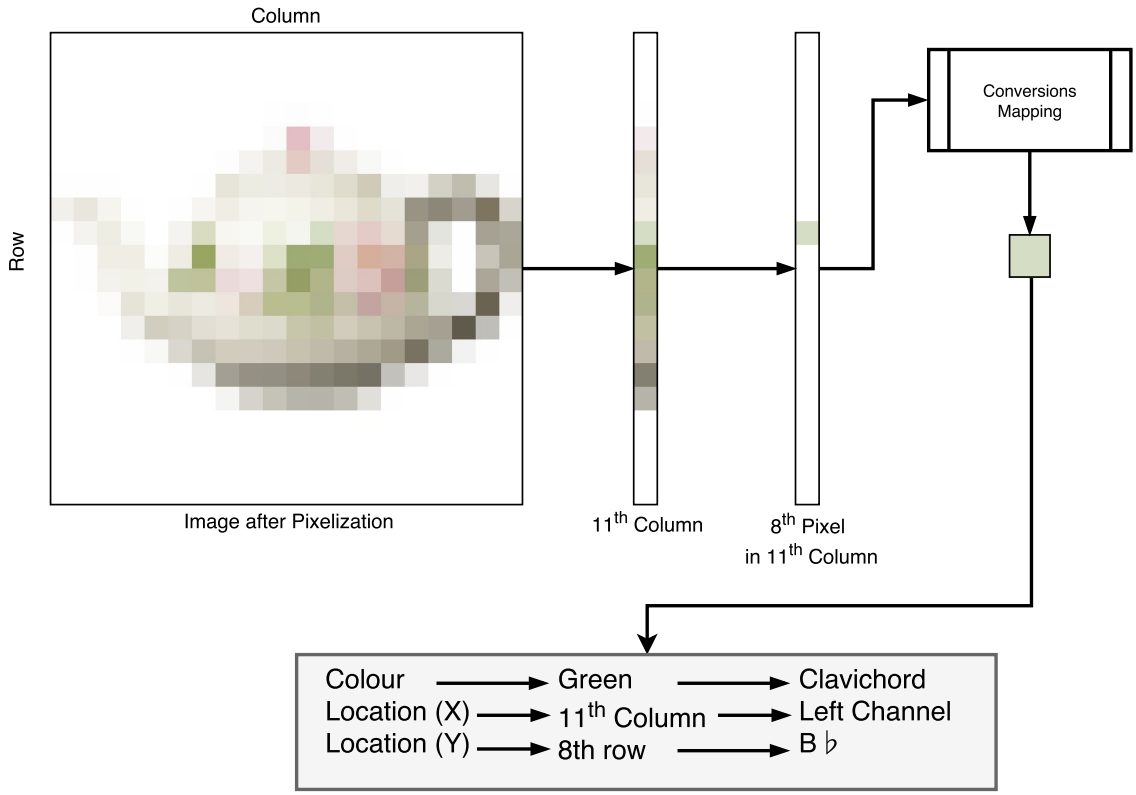


Figure 3.17: Prototype 3 conversion for example image

Colour Mapping

Prototype 3 follows exactly the same colour-to-instrument mapping as that previously optimized in Prototype 1. The colour conversion is discussed in detail in Subsection 3.4.5 according to the colour-to-instrument map shown in Table 3.2. In the example illustrated in Figure 3.15, which explains the conversion results pro-

duced by Prototype 3, the clavichord is used as the instrument to represent the specified pixel, because it has been assigned by the HCM to the green colour.

Location Mapping: X-axis

The horizontal location information of the pixels is converted to temporal information in the Prototype 3 soundscape because this prototype employs the left to right swiping direction. It is different from the previous two prototypes in that the swiping direction is changed from the top-down direction.

However, the core procedure that converts location into temporal form in a soundscape remains unchanged. By applying the equation presented in the previous section (Equation 3.2), a time delay can be applied to the sequence. Instead of applying the time delay to each row (as do Prototypes 1 and 2), Prototype 3 applies the time delay to each and every column from the leftmost column to the rightmost column. For instance, if the total time of a soundscape for an image with 20 columns is 2 s, a single column requires a time delay of precisely 100 ms. The sound is played according to the sequence of columns from left to right. They are not played simultaneously, but rather the subsequent column can be played if and only if the previous column has finished its allocated time. Let us refer to Figure 3.17. The green pixel is played after the 10^{th} column has finished playing, because the pixel is located in the 11^{th} column. Therefore, the sound of the green pixel can be heard only 1000 ms after Prototype 3 has started sonifying the image. When the swipe reaches the rightmost column, a small buzz is played for 100 ms, indicating that it has reached the end of the frame and the next frame will be starting soon.

Location Mapping: Y-axis

In order to represent the vertical location of the pixel, Prototype 3 uses different sound frequencies to represent the Y-axis information of the pixel. Previously, the pitch of the sound was used to represent the colour shade of a blob. However, decision was made to omit the support for encoding colour brightness and re-implemented

sound frequencies to represent the vertical location of the pixel.

Table 3.4 shows the mapping of the vertical location (Row Index) to its corresponding sound frequency. When representing colour shades, only one octave of the Dorian scale is used. However, because of the increased number of rows, three octaves of the Dorian scale are used to represent the vertical location. Using three octaves, Prototype 3 is able to convert an image having 22 rows. As shown in Table 3.4, the sound frequencies are arranged in descending order, where the top row has the highest frequency and the bottom row has the lowest frequency. Let us refer to Figure 3.17. The specified green pixel can be heard in a $B\flat_2$ tone, because it is located in the 8^{th} row.

Table 3.4: Dorian scale and vertical location map

Row Index	Note	Frequency (Hz)
21	C_0	65.406
20	D_0	73.416
19	$E\flat_0$	77.782
18	F_0	87.307
17	G_0	97.999
16	A_1	110.000
15	$B\flat_1$	116.541
14	C_1	130.813
13	D_1	146.832
12	$E\flat_1$	155.563
11	F_1	174.614
10	G_1	195.998
9	A_2	110.000
8	$B\flat_2$	233.082
7	C_2	261.626
6	D_2	293.665
5	$E\flat_2$	311.127
4	F_2	349.228
3	G_2	391.995
2	A_3	440.000
1	$B\flat_3$	466.164
0	C_3	523.251

3.6.5 Usage

In general, the operation of Prototype 3 is identical to that of the other prototypes. It receives visual data from a camera or a static image and translates them into the corresponding soundscape. The audio of the soundscape is then played in the earphones/headphones of the user. However, the conversion algorithm in Prototype 3 has been rewritten completely. In order to use Prototype 3, the user has to learn how to interpret the soundscape produced by the prototype.

In Prototype 3, not only is the conversion mapping different but also the concept of blobbing has been replaced in favour of the simplified pixelization approach. As described above, unlike Prototypes 1 and 2, Prototype 3 does not segment the image based on the contour-based image segmentation. The segmentation technique of Prototype 3 is considerably simpler, and is similar to that used in vOICe. Instead of using blobs, the soundscape now uses larger pixels arranged in a 2D array. Moreover, an additional major difference in Prototype 3 is the swiping direction. Previously, Prototypes 1 and 2 used the top-to-bottom swiping direction; However, Prototype 3 swipes from left to right. In order to interpret the soundscape, the user has to reconstruct the image from left to right, pixel by pixel. The leftmost column is heard first and subsequently one column after the other until the end of the image. The vertical position of the pixel is translated to the pitch of the sound.

Other than the image segmentation method and the swiping direction, Prototype 3 is similar to the previous prototypes. When using the prototype in live mode, similarly the user points the camera towards the intended direction. When the camera has captured the scene, a soundscape is generated. The user visualizes the scene by listening to and interpreting the soundscape.

3.7 Prototype 4

The 4^{th} prototype produced from this research studies utilizes a depth camera instead of the normal 2D camera used in the other prototypes. By using a depth

camera, this prototype is able to supply depth information in addition to the 2D colour image as the input. The main motivation behind the inclusion of depth is that depth information is an important element in the human vision system. In the first three prototypes, colour information was introduced, which led to some improvements in the performance in terms of accuracy and the user experience. Using their stereo vision, humans frequently use depth information to differentiate objects and gauge the distances between objects in an environment. It is hoped that the addition of depth information in VASS systems will improve the accuracy in some scenarios, such as navigation and scene recognition.

3.7.1 Overview

In addition to visual images, depth perception is always an important part of an individual's visual ability. Although there is no specialized organ for depth sensation, it always arises from a variety of depth cues, the main source of which is humans' stereo vision. Depth cues take different forms, depending on whether their source is monocular or binocular vision (e.g., motion parallax, visual perspective, and stereopsis). The extent to which humans rely on the depth information is not clear, but there are several purposes for which depth is an important factor, for example, the detection of the angle of the path of an object moving towards one or the avoidance of obstacles on the road during navigation. Essentially, for humans, the world comprises three dimensions, which explains why depth as additional information may help humans perform tasks in real-life situations such as navigation.

Currently, to a certain extent, VASS systems are capable of imparting depth perception to their users. Since most VASS systems translate 2D images into an audible soundscape one at a time, it is possible for users to perceive depth from the soundscape, in a manner similar to that in which they perceive depth by using monocular vision. Monocular vision provides depth perception through its own depth cues. Examples of monocular depth cues include motion parallax, the visual

perspective, the size of objects, and object occlusions. Most of these elicit a depth sensation through the viewer's comparison of an object's position in the scene. The user feels the sense of depth by gauging the relative position between objects and also the distance between him/her and the objects. Because of the construction of their binocular eyes, humans have an additional means of obtaining depth perception. Although usually, the depth cues elicited by monocular vision are adequate, in some situations binocular depth cues are preferable. Stereopsis is one the depth cues that result from binocular vision. By using two images (one seen by each eye) of the same scene that are positioned at slightly different angles, stereopsis creates a depth perception by triangulating the position of the same object residing in the overlapping region of the two images. Stereopsis is frequently more accurate than the depth cues elicited by monocular vision. In particular, in the case of a scene with minimal objects, it is even more effective than monocular depth cues, because gauging the relative distance from multiple objects is more difficult when there are only a few objects.

Therefore, in Prototype 4, a specialized depth sensor was applied on top of the existing 2D camera to supply depth information. It is hoped that using the data from the depth sensor, more accurate depth cues can be encoded into the soundscape to complement the existing ones.

3.7.2 Process Flow

Figure 3.18 shows the process flow implemented in Prototype 4. The flow is similar to that of Prototype 3, albeit a depth map is included as the additional visual input from the TOF camera. First, a 2D visual image and its corresponding 3D depth map are captured using the TOF camera. They are processed internally by the camera before being transferred to the computer through the attached USB port. The prototype then polls the camera and grabs both the 2D image and the depth map using the library provided by DepthSense's SDK. By default, as in Prototype 3

the software functions capture the entire image regardless of its depth. Alternatively, the user is provided with three additional options to adjust the depth level on which the system is focused. The three depth options are $< 3\,\text{m}$, $3\,\text{m} \leqslant \text{depth} < 5\,\text{m}$, and $5\,\text{m} \leqslant \text{depth}$. The input image that is to be sonified is filtered based on the depth option chosen by the user. If option 1 is chosen (depth $< 3\,\text{m}$), the soundscape includes only visual data within that depth and excludes all other visual data. Thus, the user can focus on the visual data only within the region.

The subsequent part of the process is identical to that in Prototype 3. The resolution of the visual data is reduced to 20×30 before passing through a process that extracts the visual information. The decision module then maps the visual properties, such as colour, location and size, to the auditory properties in order to generate a soundscape. Finally, the soundscape is synthesized using a sound synthesizer and passes to the audio driver to be played through the speaker/headphones.

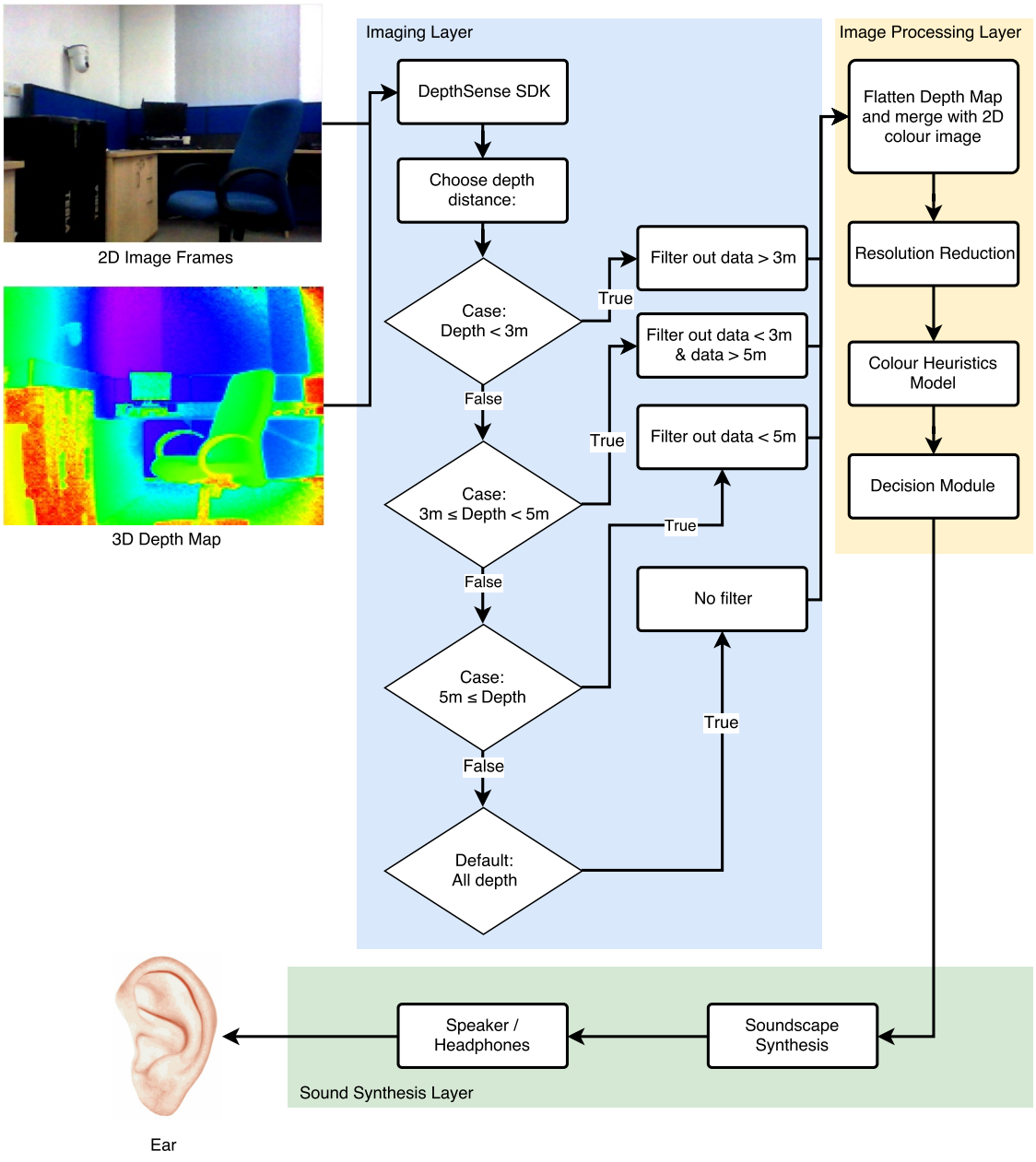


Figure 3.18: Prototype 4 conversion process

3.7.3 Software and Hardware

Hardware

A depth camera is used in Prototype 4 as the image capture device to supply 3D visual information in addition to the 2D visual images. A few different technologies exist that allow depth information to be captured from the surroundings electronically. However, the stereoscopic camera and TOF camera are among the widely used depth sensors because of their low cost and availability. A stereoscopic camera, commonly known as a stereo camera, uses two or more lenses in separate sensors, the distance between which is fixed to simulate human binocular vision. From the two images that are captured concurrently, a depth map can be interpolated through advanced image processing techniques. As opposed to the stereoscopic camera, the TOF camera uses a dedicated photon mixing device (PMD) sensor to capture a series of light signals from objects illuminated using an infra-red ray. From the infra-red ray reflected from the objects, the depth sensor is able to produce a depth map of the surroundings by calculating the distance from the intensity value of the pixels. Although each camera has its own advantages and disadvantages, the Luminophonics research group decided to use the technology of TOF to capture the depth information of the surroundings.

$$d = \frac{c}{2f_{mod}} \cdot \frac{\phi}{2\pi} \tag{3.5}$$

The depth camera used for Prototype 4 is a complementary metal–oxide semiconductor (CMOS) TOF depth camera called DepthSense® 311, developed by SoftKinetic. The main reason behind the decision to use a TOF rather than a stereoscopic camera is that the operation of a TOF camera is relatively less computationally intensive than that of other depth sensors. In principle, a TOF camera uses a very simple mathematical formula to calculate the depth information. The distances between the objects in the scene and the sensor are calculated from the incoherent light signal reflected from the objects (Schaller, Penne, and Hornegger, 2008).

Therefore, a separate light source that is intensity modulated by a cosine-shape signal of frequency, f_{mod}, has to constantly illuminate the objects of interest. The DepthSense depth sensor takes advantage of light in the infra-red spectral range so that the emitted light is invisible to the user. In normal conditions, the light travels through the medium at a constant speed, c. Hence, the object's distance, d, from the sensor can be calculated by estimating the phase shift, ϕ, between the emitted and reflected light signal as shown in in Equation 3.5 (Schaller, 2011).

Because of the relatively simple calculations involved, the TOF camera does not use a large amount of computational resources. As compared to other depth sensor technologies, such as laser range, structured light, coded aperture, and stereoscopic cameras, it is usually very fast, achieving a near real-time frame rate and high depth accuracy. Finally, because of its low computation resource requirement, it leaves more computing power for other computational intensive tasks that are often needed in the later stage of the prototype. The DepthSense® 311 (DS311) used in Prototype 4 (see Figure 3.19) features a 60 fps frame rate on a QQVGA (160×120) depth map resolution, as well as providing a colour image in VGA (640×480) resolution. Most importantly, the sensor is able to accurately gauge the distance of objects from 0.15 m up to 4.5 m. According to this specification, the DS311 sensor is suitable for application in a VASS system.

Figure 3.19: DepthSense® 311 by SoftKinetic

Software

Interaction with a DS311 sensor is not difficult, as most of the heavy lifting has been done by the manufacturer, SoftKinetic. The connectivity with all DepthSense cameras is through a single USB port in the back of the device coupled with an external power supply unit. For convenience, Prototype 4 utilizes a dual USB cable to connect the computer and the DS311 sensor, where one of the USBs is used for data communication and the second to provide the extra power needed to drive the depth sensor. Because of the dual-USB mode, the use of the DS311 sensor has significantly increased the portability of the prototype, because it can be powered by the additional USB port and does not have to draw the power from an external power brick connected to the power socket. A driver and a software development kit (SDK) are provided with the depth sensor, which were very helpful during the application development. Inside the SDK, there are a few sample applications and instructions that allow a first-time user to start using the device. So that the DS311 sensor could be used with the current software framework, a converter module was needed because the framework was designed to operate with a USB Webcam. To

convert the depth sensor as a drop-in replacement of the visual frame grabber, a converter module was coded to convert the image format captured from the DS311 sensor into the BGR format accepted by OpenCV's Image Container (Mat).

3.7.4 Image Segmentation

The image segmentation techniques implemented in Prototype 4 are identical to those used in Prototype 3, as discussed in Subsection 3.6.3.

3.7.5 Conversion Mapping

Because of time and resource constraints, instead of designing the Prototype 4 from the ground up, it was decided to create the conversion algorithms of Prototype 4 by basing them on one of the previous prototype. The base conversion process was selected based on the performance measurement conducted in one of the experiments. Because Prototype 3 scored fairly well in terms of information preservation and interpretability, the major parts of the conversion algorithms designed for Prototype 4 are based on Prototype 3. In addition, because of the addition of depth information, a simpler conversion algorithm was recommended. As compared to the previous two prototypes, the process of Prototype 3 is much simpler, because it does not involve an advanced image segmentation algorithm to extract the object blobs. In general, Prototype 4 was built on top of Prototype 3 by replacing the 2D camera with a TOF camera to sense the surrounding depth. This section explains the application of the conversion and the relationship between depth and other visual elements during the audio conversion.

Prototype 4 converts a total of three major visual elements, including the depth information that it uniquely applies. The three elements are colour, pixel location, and depth.

Depth

Depth information is rarely incorporated in a VASS system because the 2D cameras used by most of the systems are not capable of producing an accurate depth map similar to the stereopsis of normal binocular vision. Hence, not many depth implementations exist that can be used as a reference. Among these, See ColOr by Bologna, Deville, Pun, and Vinckenbosch (2007) is one of the few systems that are capable of including depth information in the visual-to-auditory conversion algorithm by means of using data from a stereoscopic camera. The application of depth in See ColOr is in the automated FOA algorithm. In the system, depth information is used as one of the guides in the visual saliency computation to search for the most relevant visual attentional field. In another implementation, Fristot et al. (2012) developed a method that encodes a depth map directly into the soundscape by utilizing a Microsoft Kinect® camera. The device resamples the depth map into several receptive fields, where each depth level is represented by a sound frequency in the musical scale. Other than these two VASS systems, there are several sensory substitution alternatives that encode depth in their implementation, such as the cane(Maidenbaum, Chebat, et al., 2014) and VTSS with ultrasonic sensors.

In Prototype 4, the encoding of the depth information is different from that in the previous two VASS systems that were discussed previously. They integrate the depth information within the conversions by either automatically calculating the visual attentional field using the depth information (Bologna, Deville, Pun, and Vinckenbosch, 2007) or encoding each depth level using different sound frequencies (Fristot et al., 2012). In contrast, Prototype 4 is designed to empower its user, providing him/her with more control of whether the depth information is included in the soundscape. The reasons behind this decision were two-fold: first, the effect of cacophony in the soundscape is reduced, and second, depth perception can be formed from a 2D image through its monocular depth cues. Therefore, in Prototype 4, users are given the option to specify the depth level from which the soundscape

is to be derived. Using this option, users are able both to listen to the soundscape from a specified depth level and to switch off the depth map, so that the soundscape is sonified from a 2D image. Furthermore, this implementation avoids the need to encode the depth information into one of the audio properties, which might overload the already crowded soundscape. If a soundscape is overloaded with information, it will create a cacophony effect, which will decrease the interpretability of the soundscape. Moreover, the user does not have to obtain depth information from an accurate source, such as the depth map; the depth perception can also be derived from the monocular depth cues that exist in a 2D image. Thus, the depth switch allows the user to switch on the function when it is needed.

The depth levels, d, that are available to the user are as follows.

- d / No filter (Default)
- $d < 3\,\mathrm{m}$
- $3\,\mathrm{m} \leqslant d < 5\,\mathrm{m}$
- $5\,\mathrm{m} \leqslant d$.

During the process of merging a depth map and a 2D image, the image pixels are filtered out based on the d value selected by the user. To illustrate, if the option $d < 3\,\mathrm{m}$ was selected, the pixels that were labelled with distance $3\,\mathrm{m}$ and above in the corresponding depth map are not merged, leaving only visual information located less than $3\,\mathrm{m}$ away from the user in the image.

Colour

Exactly as does Prototype 3, Prototype 4 uses the same colour-to-instrument map shown in Table 3.2, which was discussed in detail in Subsection 3.4.5.

Location

Prototype 4 uses the same swiping mechanism as Prototype 3. Therefore, the location conversion for every pixel follows the same procedure as described in Subsection 3.6.4 for the pixel's X-axis and in Subsection 3.6.4 for the pixel's X-axis.

3.7.6 Usage

As compared to that of the other three prototypes, the operation of Prototype 4 is slightly more complicated. However, all the prototypes are contained in the same executable that the user can switch according to the mode of the prototype he/she desires. The startup procedure of the prototypes is the same, as is the GUI. Initially, the Prototype 4 mode is disabled, but when a compatible depth sensor (e.g., DepthSense DS311) has been detected by the application, the functionalities of Prototype 4 are enabled immediately. From this point onwards, the user can choose to switch to the mode of Prototype 4 by choosing this mode in the selection menu. In short, the functionalities of Prototype 4 can run only if a compatible depth sensor is connected to the computer. The depth sensor is not exclusive to Prototype 4 because the other three prototypes can still use the 2D camera of the DepthSense's depth sensor to operate.

In order to use Prototype 4, the user needs to understand that it provides the option to choose the depth range to be sonified. The depth range can be switched on-the-fly during usage. When the user has chosen the depth range, the subsequent visual data are accordingly sonified. In comparison, the other three prototypes do not provide this option, because they operate using 2D visual data, which do not specify the depth information. As explained in the paragraph headed 'Depth Segment' in Subsection 3.7.5, four different depth ranges are provided in Prototype 4: entire depth, depth $< 3\,\mathrm{m}$, $3\,\mathrm{m} \leqslant$ depth $< 5\,\mathrm{m}$, and $5\,\mathrm{m} \leqslant$ depth. The default depth range is entire depth, where the entire image captured by the camera is sonified without any filtering. In this depth range, Prototype 4 operates like Prototype 3.

The other three options engage the information from the depth sensor to slice and filter the visual information that falls into the depth range. For example, if the option $3\,\text{m} \leqslant \text{Depth} < 5\,\text{m}$ is selected, the algorithm refers each pixel in the 2D image frame to its corresponding depth in the depth map. If the pixels do not reside in the depth range specified, which is between 3 m and 5 m, the pixels are not sonified. The resultant soundscape contains the sound of only the pixels that fall in the depth range specified.

With the inclusion of depth range, Prototype 4 presents an additional different approach for using the prototype. Usually (when the user uses Prototypes 1, 2, or 3), the camera is pointed to a specified region while the user listens to and interprets the soundscape. The depth is perceived either through the depth cues contained inside the soundscape (such as the size of the blobs) or by varying the position of the camera to compare soundscapes. However, this approach can be too simple. In certain situations, the depth perceived can be erroneous and misleading because of various factors, such as surrounding noise and the user's lack of use experience. With the depth range provided by Prototype 4, the user has the means to target a certain depth range, thus filtering out other unwanted scenes to obtain more accurate depth information of the surroundings. The prototype first sonifies the entire depth range, exactly as does Prototype 3. When the user wants to focus on a specific range, he/she chooses the range. For example, when the user is going to hit an object in front of him/her, the range $\text{Depth} < 3\,\text{m}$ can be chosen to sonify the nearby surroundings within the 3 m range. The soundscape sounds simpler and is more easily interpreted because the unwanted objects outside the specified range have been discarded. Furthermore, it helps the user to reconfirm the depth and the position in which he/she is located. Then, the depth range can be switched back to default mode to sonify the entire scene. Essentially, Prototype 4 introduces a new usage approach. It allows the user to simplify the scene by filtering based on depth range when in doubt.

3.8 Mobile Prototype

This section documents the work that was performed to create a new portable VASS prototype that would be more suitable for being carried by the user as a mobile vision aid. So far, the prototypes discussed were all developed for a personal computer (PC). This approach was used for almost all VASS systems in the past, because in general the hardware of a PC is much more powerful and highly customizable than the existing embedded systems. Furthermore, it is easier to develop a software application for the PC platform because of its better and considerably more complete OS; it also provides development tools that are better supported, as well as greater functionalities. As a result, the time between ideation and a completed prototype is shorter when developing a VASS system for a PC than for an embedded device.

Nevertheless, the PC-based VASS system is ideal for research purposes, where experiments are conducted in a controlled environment and the user has no intention of carrying the device around. However, it involves the major limitation that the prototype has to be constantly tethered to a PC. Because of its overall size, weight, and bulk, carrying a PC can be a strain on the user. Moreover, the VASS prototypes that are discussed in the previous sections are not suitable for running on a laptop computer powered by a battery. They not only consume a considerable amount of battery power but also generate more heat. A common observation was noticed in all the different experiments conducted with the prototypes is that it was possible to run the prototypes on average for only 1 hour on a fully charged laptop powered by an Intel Core processor with a 48 W h battery.

These limitations hamper the portability of a VASS system, rendering it unsuitable for daily usage. Therefore, it is important to solve these limitations so that the portability of VASS can be increased. An ideal VASS system must have a body that is light and small, a relatively long battery life, and a stable and efficient OS with a good set of development tools. Not many options meet these requirements easily. One of the possibilities is to build a customized embedded system for the purpose of

VASS system. However, with the availability of the smartphone platform in recent years, developing a VASS system as a portable system has become easier. It can be stated confidently that the smartphone opens up many more opportunities that allow VASS systems to thrive.

3.8.1 Overview

As the electronics industry continues to thrive, integrated circuit (IC) design has been advancing steadily, catering to the market demand. This has been fuelled by the increasing computing power and miniaturization of electronic devices. Finally, a smartphone revolution, spurred by the creation of the iPhone by Apple Inc., occurred in the late 2000s. A smartphone is a full featured mobile phone that is powered by an advanced mobile OS, such as iOS and Android, which are able to provide the features of a PC in a smaller package. With the development of the smartphone, a basic mobile phone became a powerful portable device that can be equipped with many functionalities only by installing mobile applications as on a computer. Moreover, with the introduction of the mobile application framework supported by a community of developers, developing applications for a smartphone has becoming easier. It is the advancement of both software and hardware that has led to today's powerful smartphone.

As mentioned in the research goals, the members of the Luminophonics project recognize that in order for a VASS system to be effective, the issue of portability, which affects most current VASS devices, must be addressed. Rather than requiring the user to be constantly tethered to a PC, a VASS has to be small and portable so that the user can carry it around everywhere, as people carry their spectacles. This is also one of the reasons why the use of the smartphone has become widespread recently. It has the power of a PC and yet is portable, which greatly enhances its usefulness. An opportunity was presented by the fact that every smartphone is equipped with at least a camera and a speaker. A small form factor that includes

an efficient mobile processor, a good camera, and a speaker constitutes the perfect combination for a portable VASS system. Therefore, a decision was made in order to harness the power of the smartphone and convert one of the prototypes into a smartphone application.

The goal was that one of the conversion algorithms would run as an application so that the user can turn his/her smartphone into a VASS device by installing the application. In this section, the details of the operation of the VASS application are discussed and also how it compares to the conventional VASS system built to be installed in a PC.

3.8.2 Process Flow

Overall, the process of this prototype does not differ significantly from that of the prototypes developed for PCs. Following the same method, they all first grab input images from a camera and pass them into a conversion function that converts the visual data into an auditory soundscape. However, because many of the components inside a smartphone are well integrated, the process of this mobile prototype is simpler than that of the other Luminophonics prototypes. Inside the PC prototypes, some customized adapters may be required, in particular for intermediary conversion. For example, in the rare situation where a camera uses an incompatible colour model, a colour model conversion step is needed to handle the conversion of the visual information into a colour model acceptable by the image container used in the software. By virtue of the tight hardware-software integration of modern smartphone architectures, such as Android and iOS, the hardware components (e.g., the camera and speaker) with which the smartphone is equipped frequently operate immediately on first use, requiring no adjustment. This also significantly reduces the overall development time, because the required functions mostly operate according to the specification using the API provided by the framework.

Figure 3.20 shows a flow chart of the general process of this prototype. It does not

differ significantly from the process used in Prototype 3 (Figure 3.15), because the mobile prototype was ported from this prototype. The only difference between the mobile prototype and Prototype 3 is the adoption of the Android framework in the development of the software. Using the API provided by the Android framework, the application is able to communicate with the hardware, i.e., the camera and speaker, directly, and an additional library or driver does not need to be installed. In the process, first an ImageReader class is used that reads out and buffers the image frames captured by the Android camera component. Then, the image frames are transferred into a separate thread that runs a native compiled code for further processing. The process is executed using the Android Native Development Kit (NDK), because processing image frames can be computationally intensive. The usage of NDK and the design decision are described in more detail in the following section. After the thread has completed the core processing, the Android SoundPool is used to synthesize the soundscape using the information from the previous process. The musical instrument models that were previously loaded into the SoundPool instance are used to generate an audio tune, which then passes into the speaker component to be played through the speaker/headphones.

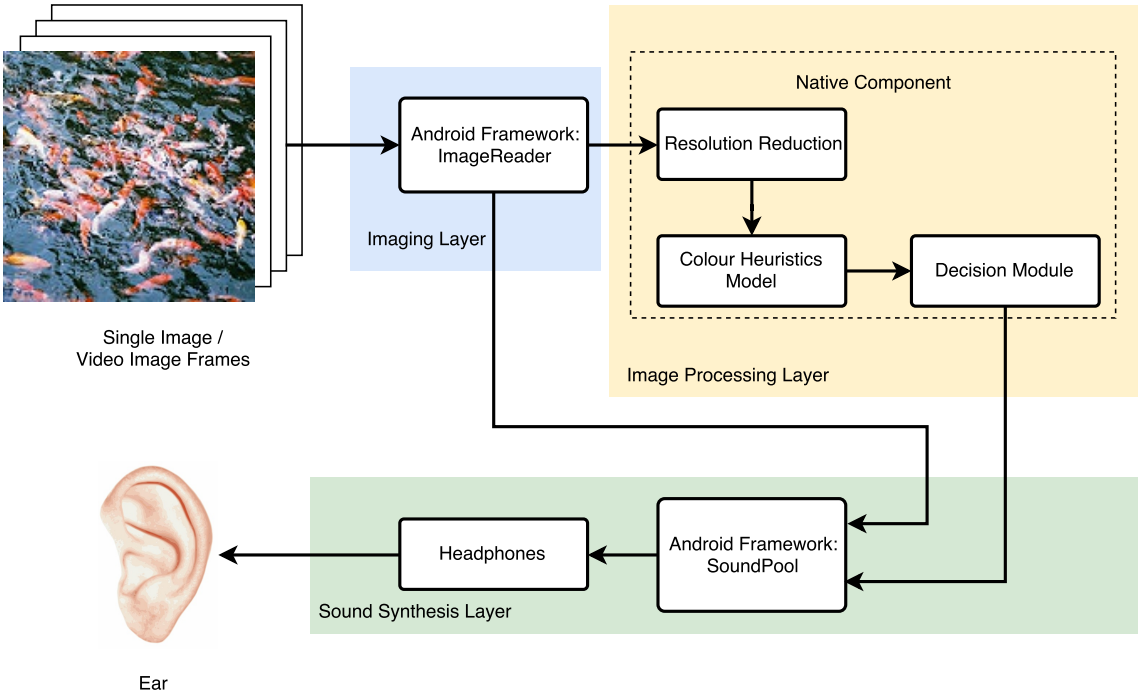


Figure 3.20: Prototype mobile conversion process

3.8.3 Software and Hardware

Software

Currently, there are two smartphone ecosystems that are widely used and well supported by the communities of both smartphone users and developers. They are Android, an open source mobile OS produced by Google, which is based on the Linux kernel, and iOS, a proprietary mobile operating system produced by Apple Inc., which is based on Darwin BSD. Although the OSs are equally powerful, because of the motivation behind their creation they target different users; however, they compete in the same market segment. iOS was created by Apple Inc. to be a proprietary mobile OS that is available exclusively for Apple products, whereas Android, a product of Google, was conceived as the open-source alternative that

could be used by different manufacturers to develop their own smartphones. An additional difference between Android and iOS is the main programming language supported by the SDK platform. The Android application was developed primarily using the Java programming language, whereas for iOS the approach of using the Objective-C programming language was used. Both have their advantages and disadvantages, and developing applications for the platform differs greatly depending on the chosen platform.

Google's Android was chosen as the primary smartphone platform for this mobile prototype for several reasons. First, Android smartphones are more widely available and less expensive in terms of overall cost. Because Android can be used by different manufacturers, it is installed in smartphones ranging from a low cost mobile phone to a more powerful fully equipped smartphone. The use of the Android platform allows this prototype to be installed on a much wider phone selection, which benefits users in poorer countries, because they can run the prototype on a cheaper smartphone. In addition, because it is an open source mobile OS, considerably more software libraries are readily available for use. Many open source audio and image processing libraries have been ported to support the Android platform by fellow open source contributors. Although there were no plans to apply the libraries in the prototype in the initial development phase, the availability of the option presented the opportunity to use customized functions that may be needed for special situations in the future.

One of the disadvantages of using the Android platform arises from its heavy reliance on the Java programming language, which may lead to problems for applications that require heavy computational resources. Android platforms are built almost entirely on Java, which uses a virtual machine (VM) to run its instructions. While the VM provides advantages, such as garbage collection (GC), an improved debugging mechanism, and abstraction, these features come at the expense of performance. In a normal situation, the performance loss is negligible, but it is a major problem in the case of computationally heavy tasks, such as image processing. If

the prototype were ported to run entirely on the Java programming language, there would be a significant lag between soundscapes, because it requires considerably more processing power to process each frame. The lagging effect has been proven to render the entire device unusable and degrades the overall usage experience. In order for the application to be usable, the soundscape must run at a minimum refresh rate of 1 Hz (one soundscape per second). Fortunately, this problem was overcome by the implementation of the computational chunk of code supplied with the Android native development kit (NDK)[9]. The NDK allows the application to leverage C/C++ compiled code for computationally intensive tasks. Therefore, as illustrated in Figure 3.20 in the dotted rectangle frame, processes including resolution reduction, the colour heuristics model, and the decision module were programmed entirely with C language specifically to take advantage of the native processor environment, bypassing the VM. While the application runs on the VM, it accesses the native components through using the Java native interface (JNI). When this mechanism is used, the overall lag is reduced substantially. The processing time, which was previously 5 s using the VM, is reduced to 200 ms for a standard ARM processor.

Hardware

It was considerably easier to select the appropriate hardware for a mobile VASS prototype than to customize a PC. Since a smartphone comes in a complete package, it does not involve a significant amount of hardware customization as compared to the previous prototypes. The work consisted only of choosing a smartphone that meets the minimum requirements for running the mobile prototype application. During the development of the prototype, the Sony Xperia S was the smartphone used. This smartphone was considered to be one of the higher-end Android smartphones at that time. However, because of the rapid smartphone development, its specification was quickly eclipsed by many Android smartphones that cost less. Currently,

[9]More information about the Android NDK can be found at https://developer.android.com/ndk/index.html

almost all smartphones on the market are capable of running the mobile prototype application effectively. Although the camera and speaker are the main components of a VASS system, the choice of the right processor and memory was the top priority, given that every smartphone is equipped with a camera and speaker that meet the basic requirements. To allow this mobile VASS prototype to run normally, the smartphone must be equipped with at least a dual-core 1 GHz processor (CPU) and a RAM amounting to 1 GB. These recommendations were proposed after trial and error tests of multiple different smartphones. As observed throughout multiple trials on different hardware, smartphones having a single-core processor or a memory of less than 1 GB must be deemed inadequate, because the processes of the VASS system are resource hungry, and they may cause the application to freeze or, even worse, cause the entire smartphone to reboot. When a multi-core processor is used, more threads can be generated for visual-to-auditory conversion and thus the main process that captures the video and plays the soundscape simultaneously is not disrupted. Conversely, the larger amount of RAM can be dedicated to buffering the incoming image frames while processing the current ones. In the future, a mobile graphics processing unit (GPU), such as PowerVR, NVIDIA Tegra, or Qualcomm Adreno, can be utilized to increase the performance by relieving the CPU of some of the computational workload. A GPU not only is able to reduce the time taken for image processing and information conversion, but also reduces the battery consumption by improving the efficiency of the entire processing procedure, taking advantage of the improved floating-point computation and parallel computation.

3.8.4 Conversions Mapping

Colour Mapping

Since the mobile prototype was ported directly from Prototype 3, its conversion mapping also follows the same procedure as Prototype 3. The same colour mapping is used for every prototype. The details of the colour mapping are discussed in

Subsection 3.4.5 following the same colour mapping, which is shown in Table 3.2.

Location

The swiping mechanism of the mobile prototype follows exactly the same procedure as Prototype 3. The location conversion for every pixel is presented in Subsection 3.6.4 for a pixel's X-axis and in Subsection 3.6.4 for a pixel's Y-axis.

3.8.5 Usage

To use this mobile prototype, the user must install the application in an Android smartphone. The application is initiated when the user starts the application in the OS. Then, the application displays a screen on which a video composed of the image frames captured by the phone's camera is shown. In the background, the application is performing the computation to convert the image frames into the corresponding soundscape. To hear the soundscape, the user can either listen through the speaker or plug headphones into the audiojack of the phone. It is recommended that the phone be aligned horizontally in landscape mode with the rear camera facing frontward. This stance is very natural, because it resembles that of a normal user taking a photograph with phone camera.

Normally, the application is used for two different primary purposes: object recognition and navigation. To visualize the object in front of him/her, the user must point the rear camera of the phone directly towards the object. The soundscape produced by the application contains the information about the object and the surroundings. However, to navigate the user may need to move the camera through multiple angles. Unlike the human eye, a phone camera has a limited field of view. By listening to the multiple soundscapes from multiple angles and consolidating them, the user can visualize a larger surrounding area. Taking the cues from the soundscape, a visually impaired user can navigate the area. Users can also combine the usage of the application with that of a cane to improve their navigation.

Chapter 4

Experiments and Results

4.1 Introduction

Experiments based on human subjects to obtain feedback and understand the interaction between a computer system and its user have always been an integral part of computer science research. This is especially true for computer systems or software applications with which human subjects constantly interact. A considerable amount of valuable information can be gained by observing a human subject using the systems while performing a set of carefully planned experimental tasks. Using the data gathered from these experiments, the system processes and other details of the prototypes can be fine-tuned. This process suits the prototyping process, where improvements are made incrementally using information supported by experimental observations.

In the past, each sensory substitution research study, including those on VASS systems, had its own unique set of experiments, where the simplest experimental method involved a feedback questionnaire or even a carefully planned scenario to measure the effect of the system on its users. In particular in the case of a VASS system, the tests and experiments using human subjects were designed mainly to measure its performance and to examine how the users perceived the soundscape it

produced. Moreover, a large part of the userbase of VASS systems comprises the visually impaired population. The feedback from this population of users is very valuable, because their experiences when using the systems with their limited sight are frequently unique. The results of the experiments were frequently published in scientific journals to demonstrate the capabilities of the systems and also to provide other researchers with a basis for comparison. Since the first VASS system was reported, various performance measures were introduced by researchers worldwide, such as the accuracy of object detection, the time taken by a user to learn the system, and the time required to interpret the soundscape. Different systems were measured differently under different sets of conditions. Thus, the evaluation of VASS through experiments based on human subjects is an important phase in VASS research.

As well as on the system's accuracy, emphasis should be placed on designing a system based on the needs of its intended users. This has been promoted strongly by the human-computer interaction (HCI) research community, because a system is ineffective if the user does not benefit from it. Moreover, VASS systems, like many other sensory substitution systems, are built mostly as an assistive technology device for the disabled for rehabilitation purposes. Because the intended users belong to a group of people with disabilities, the device must be designed to cater for their special needs. When addressing this population, the designers of the systems must take their concerns into account so that the systems answer these concerns, as well as helping them to accomplish their tasks. In order to measure the performance of the system, the experiments should also be designed such that they focus on the user experience, measuring the effectiveness of the features of the prototypes from the perspective of the user. In summary, the goal of experiments is not only to reveal the performance of the system but also to measure the user experience when using it. When the experimental procedures are embedded into the prototyping phase, the development of the prototypes can benefit from the data gathered from the experiments and at the same time they can be improved incrementally with each evolution.

4.2 Experiment 1

The first experiment was conducted after the completion of the Prototype 1. The main purpose of this experiment was to obtain an initial evaluation of the overall performance of Prototype 1 and also to verify some of the newly crafted ideas. The new ideas that are built into Prototype 1 consist of the inclusion of colour information in the soundscape, contour-based image segmentation, the swiping mechanism, and the usage of natural musical instrument models when synthesizing the soundscape. By evaluating the effectiveness of the prototype's performance, the relevance of the new ideas for the users were able to be gauged. This experiment played an integral role in the prototyping process, because the subsequent improvements depended on the information gathered from it. Using the results of this experiment, solutions to fine-tune Prototype 1 were designed and plans for the next iterations of the prototypes were made. The details of this experiment were reported in my first conference paper (Tan, Maul, N. R. Mennie, and Mitchell, 2010).

For measuring the performance of Prototype 1, a decision was made to take an approach similar to that used in previous similar research studies on VASS systems. As in other VASS system experiments, human subjects were asked to use the devices to perform carefully planned test activities. The main purpose of conducting the experiment in this fashion was to understand the human perception of engaging with the device through listening to the soundscape. The perception of the user plays an important role in determining the success of a VASS system. Therefore, the psychological aspect of the user when using the device should be monitored. To achieve this, the experiment was designed to include feature-based test activities that monitored the psychological aspect of the user when using the system. Basically, the experiment tested the perception and the reaction of the subject (user) when exposed to a physical stimulus (the soundscape). During the experiment, the participant was required to listen to the soundscape produced by the system while accompanied by a facilitator who recorded all his/her reactions, including the actions he/she performed

and the accuracy of his/her test performance.

The structure of this experiment comprised two phases. The results were recorded only after the participant had successfully completed both the phases. The first phase consisted of a training session, where the user was taught the basic operation of Prototype 1 and shown how visual information is converted into a soundscape. The objective of this phase was to equip the participants with general knowledge of the VASS system and the basic operation of Prototype 1. Upon completing the training, the participants were requested to complete the test cases in the next phase using the knowledge they gained from this session. Each participant attended one training session, which was limited in its scope, before performing the main experimental activities. In order to minimize the chance that the participants would gain uneven advantages from the training session, the training materials were standardized, with each participant receiving the same set of material. However, the participants were given the opportunity to ask questions when in doubt. The reason why the training sessions were as basic as possible was that one of objectives for this experiment was to evaluate the learnability of the prototype.

Immediately after completing the training session, each participant was instructed to complete the next phase, which was the main experimental phase. This phase consisted of a series of test cases designed to test each individual feature of Prototype 1. The test cases of this experiment were designed to focus on the conversion algorithm. During the tests, the participant was seated in front of a computer while being monitored by a facilitator. An application that implemented Prototype 1 was installed in the computer. It converted the test images into soundscapes and played them through a pair of headphones worn by the participant. The role of the facilitator was to observe the reactions of the participant closely and to record his/her judgements after listening to the soundscape. In total, six different experiments were conducted, each of which tested a different aspect of Prototype 1. They comprised a colour test, object test, shade test, location identification test, object location test, and finally an object counting test. So that the experiments would be suitable for a

beginner, the complexity and the difficulty of the test cases were set at a low level, with a maximum of two feature combinations appearing in each test. Each task had its own objective to which the participant was required to adhere. When the objective was completed, the accuracy of the participant in each experimental task was recorded. In summary, this experiment was designed to measure the extent to which Prototype 1 assists users in six different tasks through the features implemented in the conversion algorithm. The performance evaluation of the prototype depended on how well the participants executed the task as instructed for each experiment.

4.2.1 Phase 1: Training

Before the main experiment, each participant was provided with a training session individually facilitated by the prototype designer. The training session was designed to be limited to explaining the basic features implemented in Prototype 1. The facilitator explained the basic visual-to-auditory conversion mapping, including the conversion of colour, blob location, and blob size. After completing the training materials, the participants were expected to be able both to visualize the input image using the basic feature mapping by listening to the soundscape and to understand the combination of two or more basic feature mappings forming a complex visualization.

Although the training session was conducted only once for each participant, the duration of the training was determined by the participant. The participants were allowed to repeat the training material until they were satisfied with their newly acquired ability. Throughout the training session, opportunities were given to the participant to ask the facilitator any question when he/she was in doubt. The Q&A session between the facilitator and the participant significantly accelerated the learning process of the participants and improved their understanding of the VASS system.

The first element of the training session focused on differentiating the 10 colours encoded in the soundscape. The participant was given 10 different colours (black,

grey, white, red, orange, yellow, green, blue, indigo, and violet) and the corresponding sound timbre associated with each one. The participant was instructed to identify the 10 different musical instruments and memorize the associativity of their sound timbre and the colour. In addition, in the training session the participants were taught to differentiate the shades of each colour based on the pitch of the soundscape. Different shades of the colour were projected, while the participant was exposed to a soundscape in different pitches aligned with the shades of the colour.

The second element in the training session addressed the location of the object. Location is another important feature implemented in the prototype, allowing the user to locate an object based on the sound properties after forming a mental image of the soundscape. To train the participants to identify the location of an object, they were taught to relate time delays to vertical positions and stereo sound placement to horizontal position. In the training session, an image was separated into four quadrants, top-left, top-right, bottom-left, and bottom-right. Then, different images were placed in different quadrants each time repeatedly. The participant was shown how to differentiate the location by recognizing the effect on the soundscape when the objects were placed in a different quadrant.

The final feature that can be interpreted from the soundscape produced by Prototype 1 is the size of the blob. In the final training set, the participant was taught to describe the size of the blob (or object). In Prototype 1, the size of the blob was correlated with the volume of the sound it produced. Therefore, the participant was taught to recognize the volume and estimate the size of the blob relative to the other blobs around it. In the training material, multiple circular blobs were drawn in the input images and the algorithm generated a different volume according to the blob size using the same sound. The participant learned how to differentiate the volume and correlate it to the size of the blob.

The experiments in the next stage were designed to include both simple and complex images in the tests to examine the extent to which the user understood the system and also how well the system converted the images.

4.2.2 Phase 2: Experiment Activities

A total of six different experiments were conducted in the second phase, where each experiment focused on a different aspect of the soundscape produced by Prototype 1. The level of difficulty of these test cases was adjusted to a maximum complexity of two feature combinations per test case. The purpose of this was to lower the barrier for the participant, making the tests easier for them to complete. Moreover, at the time, Prototype 1 was still at the preliminary development stage and natural images captured from the surroundings were too complex for it to sonify. Therefore, it was decided that in this initial experiment, synthetic test images with simple colours and shapes would be used. The test images used in this experiment are included in Appendix B.

Colour Test

The first experiment consisted of the colour test, where the participants were required to identify a colour from the soundscape. The colour test was chosen as the first experiment because it was the simplest. A total of 10 test images, each of a different colour, were used in this experiment (see Appendix B.3). The participant was shown one test case, each time randomly selected from the pool, while listening to the soundscape generated by Prototype 1 through headphones. The process of selecting a test case and identifying the colour was repeated 20 times for each participant. The purpose of the test was to estimate approximately each participant's ability to identify colours.

Object Test

The objective of this experiment was to examine whether the prototype has the ability to convert sufficient visual information into an auditory soundscape to enable the user to identify an object through the soundscape. A total of four different object classes were prepared for this test, each having a different sound signature. During

the experiment, one object class was randomly selected from the four object classes. The objects were displayed and sonified in random order, which was different for each participant. Because each object had 5 different variations, the participants were required to identify 20 different test cases composed of 4 object classes. The accuracy of each participant's object recognition was taken as the average result of the 20 trials.

The test images of the object class were shown to the participants before they took part in this experiment. They needed to find and register any visual signatures that were unique to the object class to ease the process of identifying it using the soundscape. The four object classes (bee, house, stick-man, and tree) were pre-selected based partly on their distinctive features. The bee was chosen because of its combination of black and yellow colours; the participants were expected to hear the sounds of the black and yellow colours (see Appendix B.2). The images of the house were drawn based on two distinctive shapes in two different colours with a triangle at the top and a square underneath it (see Appendix B.4). It was hoped that the house could be determined from the features of these two shapes. To differentiate them from the other images, the stick-man images were specifically drawn using a single colour with a rounded head and a body composed of several straight lines (see Appendix B.7). Finally, the tree images were drawn with simplified green foliage and a brown trunk below it (refer to Appendix B.8). The simplistic test images were specifically designed with a maximum of two feature representations so that the beginners could recognize the objects based on their two distinctive sound patterns encoded in the soundscape.

Shade Test

In Prototype 1, not only is the type of colour represented by different musical instruments, but the lightness of the colour is represented by the pitch of the timbre as well. In this experiment, different colour shades were tested. However, the shade test implemented in this experiment was simple, the intention being to test whether

the concept of sound pitch augmentation based on colour lightness is effective. For each test case, two blobs of different shades of the same colour were drawn, one at each side of a rectangular image frame. The participant was instructed to specify the location (left or right) of the darker blob based on the pitch of the soundscape he/she heard (see B.5). The test case was repeated 20 times for each participant and the accuracy results were produced by calculating the number of correct guesses. This experiment tested the ability of the participant to discriminate the horizontal location and the shades of colour based on the soundscape produced by Prototype 1.

Identify the Object's Location

This experiment was designed to test whether the location information embedded in the soundscape could help the participant identify the location of an object. Using the same image object classes as in the previous test, the object test, four different objects were selected and redrawn in the four quadrants (top-left, top-right, bottom-left, and bottom-right) of an image frame. Figure B.9 shows four images, where in each a tree is located in a different quadrant. The participant was then instructed to identify the object's location, which was chosen by the facilitator. For example, if the facilitator requested the participant to locate the tree, the participant needed to listen carefully for the features of a tree and then pinpoint the quadrant that contained it. In this experiment, 24 different combinations were used and each participant was required to listen to 20 different test cases in random sequence. In order to correctly identify the location of the object, the participant had to be able to determine the location through the horizontal and vertical position, as well as to differentiate objects through complex features such as colour, shape, and texture.

Identify the Object in the Quadrant

This experiment was closely related to the previous experiment. Instead of identifying the quadrant in which the object was located, the participants were asked

by the facilitator to identify the object in the chosen quadrant. The same 24 combinations of images were used for this task as for the previous task. In order to correctly identify the object in the quadrant, the participant was required to differentiate multiple visual features in the soundscape, most importantly location (X- and Y-axis), colour, and blob. A correct answer was one of the four object classes.

Counting Test

The final test was the most difficult of all the tests used in this experiment. The participants (and the prototype) were tested on all three visual features, i.e., location, size, and colour. As shown in Appendix B.1, images that contained various circles in different colours and sizes were used in the test cases for this test. The participants were asked to count the number of blobs in the image by listening to the soundscape. The testing process was repeated 10 times for each participant. In order to correctly complete the task, the participant had to create a mental image containing the circles using the information in the soundscape.

4.2.3 Results

The bar chart in Figure 4.1 shows the recognition accuracy for each test in Experiment 1. Based on the fact that for most conditions the chance level was 25% (for the colour and shade tests the chance levels were 10% and 50%, respectively), the results suggest that Prototype 1 indeed was operating as expected. Although the participants were using the prototype for the first time, the results show that they were able to achieve a minimum accuracy of over 50%, which proved that Prototype 1 was promising. With better appropriate guidance and prolonged usage, the participants may even have achieved a higher accuracy using the prototype.

Among all the test cases, the accuracy for the 'find object' test was the lowest, being only 60%. Although the 'find object' and 'find location' tests were similar, the latter required a deeper understanding of the conversions and sharp ears to

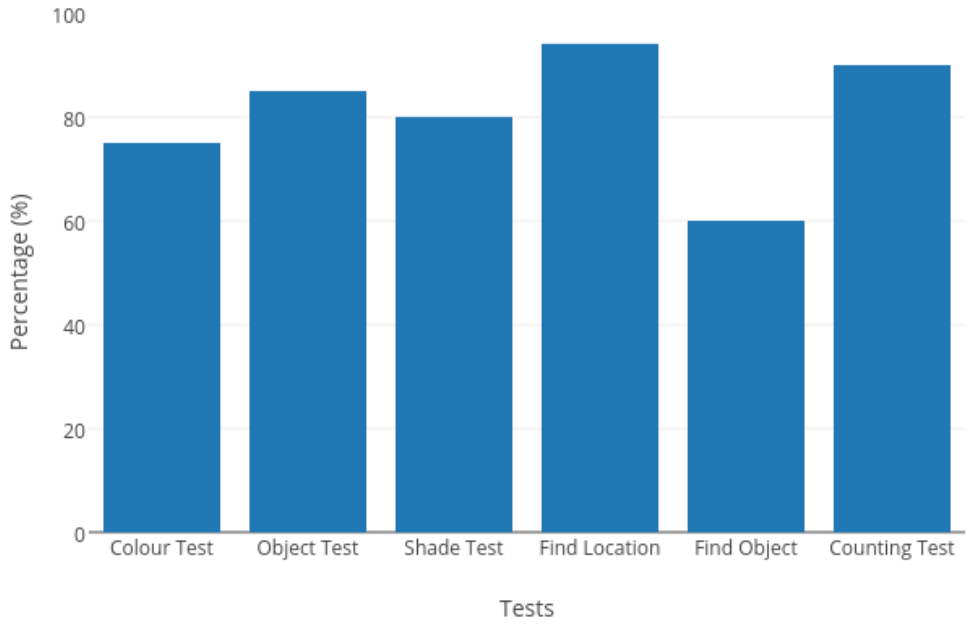


Figure 4.1: Experiment 1: Accuracy

differentiate the various elements of the soundscape in order to decode the visual information contained in it. The participants were required first to determine the specific region where the object was located and then identify the object, which the participants frequently failed to do. In the 'find location' test, the participant could listen to the soundscape as a whole and then identify the object through the sound signature of certain objects and then determine the location from the time delay. This suggests that the order of the information may affect the recognition accuracy, especially for beginners because their mental map did not endure as long when exposed to so many different sounds and noises at the same time. It is therefore important to take this effect into consideration when designing the training sessions

for VASS systems.

In contrast, 'counting test' which was expected to be the most difficult, yielded the best results. The participants obtained a high accuracy rate of 90% correct answers. In principle, the swiping mechanism the algorithm in Prototype 1 that is used to represent location information helps in such scenario. The results prove that the top-down swipe is useful for the users, helping them to create a mental map filled with blobs and their locations. The feedback from all the four participants later further confirmed that they were able to paint the blobs in their mind by listening to the swipe. By following closely the direction of the swipe from top to bottom, they were able to guess the number of blobs and their location in each region.

4.3 Experiment 2

In total, this research project has conducted two experiments, Experiment 1 and Experiment 2, to test the functionalities of the prototypes. Experiment 1 was conducted in order to evaluate new ideas. With the promising results of Experiment 1, the process of prototyping continued in order to further improve on the ideas. After the first experiments, no major experiment was conducted to evaluate the subsequent few iterations and the next major experiment was conducted only after the completion of Prototype 4. The second major experiment was conducted after the implementation of depth sensor that incorporate the depth element in addition to the usual 2D visual data. The purpose was to measure the research progress of this project. This experiment, Experiment 2 is crucial not only because of the depth concept and all the new adjustments, but also how well the prototypes perform and also how well they operate in a real-life scenario. Hence, this experment was designed mainly to evaluate the performance of each prototype. The details of the experiments were documented in my published article (Tan, Maul, and N. R. Mennie, 2015). The experiment has obtained the approval from the Ethics Committee of University of Nottingham Malaysia.

First, in this experiment the objective was to test the prototypes in a simulated real-life situation in a controlled environment using subjects other than personnel belonging to the project. After Prototype 1, the prototypes produced by the Luminophonics project were tested only internally by our own researchers and were never used by outsiders in a real-time situation. Therefore, in this experiment, participants from different categories of age and profession were invited to use all the prototypes. They were required to carry the prototype while completing the tasks designed for the experiment. The prototypes used in this experiment were implemented in physical form. The visual-to-auditory conversion algorithms for each prototype were fully developed and installed in a laptop that was equipped with the necessary input and output devices, the camera and the headphones. After using the device, the opinions of the participants of their perception and usage experience were collected through either questionnaires or verbal interviews.

Second, one of the prototypes was developed to incorporate depth information by using an additional depth sensor, as well as a normal camera. Using the depth sensor, the prototype is able to capture depth information from the surroundings and build a depth map. The combination of the depth map and the image from the normal 2D camera allows Prototype 4 to supply the additional depth information to the user that helps them in their judgement. In normal human vision, depth information is frequently used when making a decision in scenarios such as navigation and reaching towards an object. Because depth is relatively new subject in the area of VASS, Prototype 4 took an innovative approach in the implementation of depth information it its information conversions. The conversion algorithm was built such that it does not automatically convert the depth information, but the users are provided with the option to switch on the conversion of depth information into the soundscape whenever they need it. One of the objectives of this experiment was to determine the practicality of the depth implementation in Prototype 4. Because in the experiment the participants used all the four prototypes one after another, more data pertaining to the manner in which Prototype 4 processes depth information could be gathered.

The opportunity to compare Prototype 4 and the other prototypes, based on either observations of participants or their feedback, allowed us to collect more information about the effectiveness of the manner in which Prototype 4 represents depth and also the usefulness of depth information in the VASS system. Moreover, through their responses to the questionnaires, the participants shed some light on the effectiveness of the implementation and also possible means of improving Prototype 4.

Third, from the past trials and experiments, it was understood that it is very difficult to create a VASS system that completely replaces human vision and is able to perform perfectly in every situation. Thus, the efforts were refocused on finding the optimum features and settings for a VASS system that perform well in two major scenarios, navigation and object recognition, that involve the application of depth information. Therefore, the third part of this experiment consisted of using the prototypes in a controlled environment in the scenarios mentioned above. Navigation and object recognition tasks were combined into a single test course in a safe compound to evaluate the ability of the user to perform the tasks and the effectiveness of the prototypes' translation of visual information into a soundscape in terms of assisting users. In general, the users, who had been blindfolded, were required to navigate a series of obstacles in order to perform a task involving object recognition at the end of the course, and then return to the origin. The time they took to complete the tasks and their reactions during the course were recorded for performance analysis and comparison.

Finally, one of the features that the Luminophonics project hopes to improve is the learnability of the VASS system. When designing the prototypes, the aspect of learnability was seriously taken into consideration. The conversion algorithms were built to lower the barrier of entry so that new users can start using the device more quickly. In this experiment, the objective was to determine which type of algorithm a beginner can learn more easily and also the aspects that influence the learnability of a VASS system. In order to measure the learnability of the prototypes, the invitations to participate in the experiment were given only to people who had no

prior knowledge and experience of a VASS system. In contrast to Experiment 1, no training was provided for the participants before the actual experiment. These conditions were set so that, in the experiment, the time required by users with no experience to learn and adapt to the system in order to perform the task could be measured.

4.3.1 Course Design and Preparation

Specifically for this experiment, a fixed navigation course was designed and plotted in an indoor room. An indoor room was chosen for these experimental activities for a few reasons. The main reason was to provide a safe and controlled environment for the participants, because they were blindfolded for the greater part of the session. While an indoor environment is not as realistic as an outdoor one, safety is of utmost importance when conducting such experiments. Moreover, because the prototypes had not been fully tested in an outdoor environment, it was possible that they might not operate correctly in potentially harmful conditions, such as an environment with dangerous objects and situations that involve moving obstacles, for example, animals and vehicles. In these types of difficult and complex situations, the system is required to react sufficiently fast and provide accurate information to the user. The prototypes are still under improvements to meet these stringent requirements. Before the prototypes can meet the standard, it is safer to use them in controlled indoor environment conditions under supervision.

In addition to addressing safety concerns, an experiment conducted in an indoor environment has a second advantage. Inside a room, the surrounding condition can be easily controlled, in particular the lighting conditions, which might degrade the performance of the prototypes if not handled correctly. For example, in an environment with extreme lightning conditions, i.e., one that is too bright or dark, the colour of the objects changes. Moreover, the TOF depth sensor used by Prototype 4 does not effectively register depth information in a brightly lit environment. This is

because a high amount of infra-red rays from foreign sources degrades the infra-red pattern emitted by the sensor and hence increases the noise in the data received by the depth sensor. In order to obtain more accurate results, the surroundings in which the prototypes were operated needed to be consistent across all the participants. Therefore, an indoor environment was preferred, because it allowed us to control the light intensity in the surroundings, e.g., by controlling the number of fluorescent lamps that were switched on. In addition, an outdoor environment can be affected also by other uncontrollable conditions, such as weather. In summary, Experiment 2 was conducted in an indoor room to provide a consistent environment throughout the entire experiment, minimizing environmental and risk factors that might have affected the accuracy of the results.

When designing the course for Experiment 2, various research studies in which similar experiments involving a group of visually impaired people were being referred. After considering many issues and clues in the published studies, a navigation course that was inspired mainly by the Red Serpentine course used in the experiment of Bologna, Deville, and Pun (2008) was created for this experiment. The navigation course was a fixed course designed to allow the human subject to walk along a path, on the shoulders of which various obstacles were located. In Figure 4.2, the red line indicates the path that the participants were required to follow and the shaded boxes beside the red line are the obstacles that the participants were required to avoid. The path is a long wavy line from its origin to the goal (indicated in the figure by a flag). When the participant reached the goal, he/she was required to complete a task that involved object identification. Then, he/she was required to walk back to the origin.

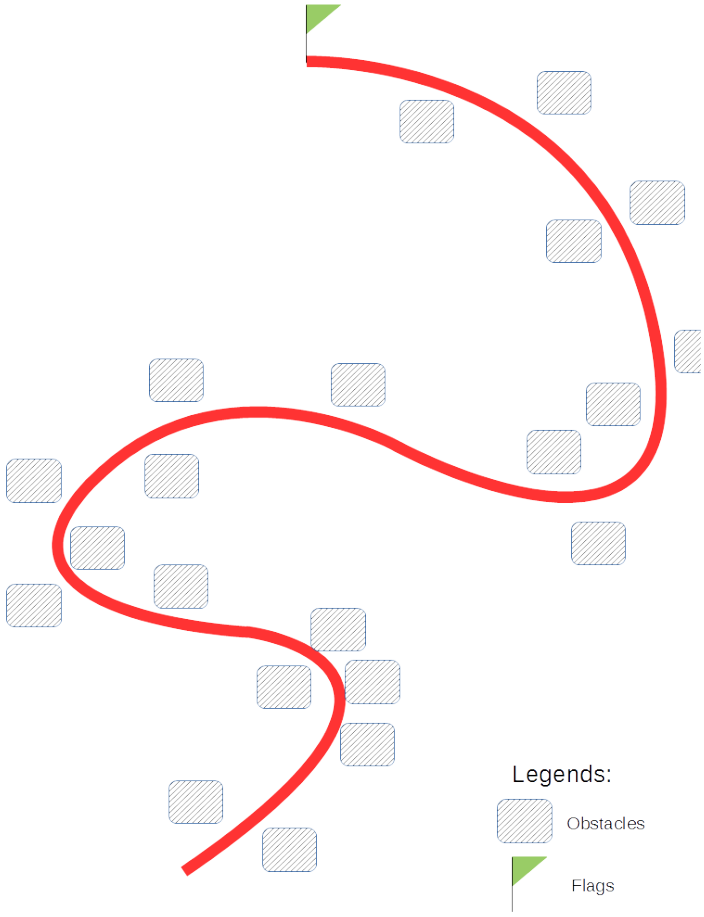


Figure 4.2: Experimental course design

Figure 4.3 shows the real experiment site set up in a carpeted indoor room with a level floor. As seen in the figure, a 90 m long path was marked on the floor using yellow floor tape. A bright yellow tape was chosen because it is more easily detected and recognized, especially by a beginner. Floor tape with alternating red and white stripes was stuck at both the left and right hand side of the yellow floor tape to act as a barricade. This tape was used as guidance lines: when the user heard the sound representation of the red and white stripes in the soundscape, they knew

that they were not focusing on the correct region. Because of space constraints, the wavy path drawn on the floor did not follow exactly the previously proposed design. Participants walked from the origin (shown in the middle of Figure 4.3a) following the yellow curved path to reach the table on which three balloons were located (shown in the middle of Figure 4.3b) and then back to the origin.

At the end of the course (as seen at Figure 4.3b), there was a table on which were placed three balloons of different colours (blue, yellow, and red). The table and the balloons were used in the object recognition task that the participants performed when they reached the end of the course. Upon reaching the table, the participants had to perform the task at the table before resuming their journey back to the start of the course. In both figures, multiple brown cardboard boxes laid randomly outside the barricade tape can be seen. There were a total of 10 cardboard boxes in the room, which acted as obstacles that the participants had to avoid. Empty cardboard boxes of various sizes were chosen as obstacles because of their characteristics of being soft and light, which meant that they would not harm a participant if he/she accidentally hit them. The role of these obstacles in the experiment was to simulate the static objects that should be avoided in a real-life situation, e.g., rocks, trees, and sign posts. Therefore, the number of boxes knocked by each participant was recorded as one of the indicators for the performance measurement.

(a) Front view (b) Rear view

Figure 4.3: Experiment room

4.3.2 Experiment Activities

A total of four prototypes produced by Luminophonics were tested in this experiment. It was important that the experiment would produce unbiased results for each prototype. Hence, a few preventive measures were introduced so that the results would be as fair as possible.

For instance, a unique sequence of prototypes for each participant was introduced. Given that every participant was required to navigate the same course four times using each prototype once, it was possible that a beginner would start to navigate better towards the end of the experiment. This is because, when the participant had navigated the course a few times, he/she might have developed a mental map of the surroundings or have become more familiar with the VASS system. Thus, it was unfair to conduct the experiment using the same sequence of prototypes for all the participants. This might have created the effect that the final prototype in the sequence appeared to perform better in the experiment when in fact the user navigated more easily because he/she was using the experience gained from prior runs. To minimize the effect of sequence bias via learning and memory, each participant was given a unique sequence so that each prototype was located an equal

number of times in each position in the sequence. Because there was a total of four prototypes, there was a total of 24 possible sequence combinations. For example, if his/her sequence was 4-2-1-3, the user would first use Prototype 4, then Prototype 2, Prototype 1, and Prototype 3, in that order.

In the object recognition component of the experiment, the position of the three balloons was also randomized for each run of each participant. Basically, at the end of the course the participant was required to select the correct balloon as instructed by the facilitator before returning to the origin. The balloons were not located in a fixed position, but rather the facilitator changed the position of the balloons randomly while the participant was navigating the course before he/she reached the table. By randomizing the position of the balloons, the effect of the participant memorizing their position was negated. Thus, the participant was required to depend on the soundscape produced by the prototype when selecting the balloon.

Because there were a total of 24 sequence combinations, the invitations to participate in the experiment were sent to more than 100 people in the hope that more than 24 people would accept. The initial plan was to allocate a session of about 1 h to each participant (15 m per prototype for a total of four prototypes). A schedule was drafted to accommodate at least 24 sessions, which comprised a total duration of five days for the entire experiment. A total of 28 people signed up for the experiment, but only 16 arrived. The background and age group of the 16 participants differed. However, they all had normal vision. At each experimental session, a facilitator was present to coordinate the experimental activities and provide guidance to the participants.

Before the start of each session, every participant was given a unique four- number sequence indicating the order of prototypes he/she would be using during the session. After receiving a short briefing presented by the facilitator on the operation of each prototype, the participant was blindfolded with a thick black sleep mask. Meanwhile, a laptop computer containing the prototype software was loaded and

started before being placed in a backpack that would be carried by the participant. Two options were given to the participant for carrying the prototype device: the handheld or head-mounted mode. The session started when the participant was prepared. Following the sequence given, the participant completed the course using the prototypes serially. During the entire session, the facilitator recorded all the data required for the performance evaluation, including the duration of the journey, number of obstacles knocked over, and time taken to select the correct balloon. At the end of the session, each participant was required to fill in a feedback form about the entire experiment. The feedback form was used to record information that could not be observed during the experiment, such as the opinion of the users and their experience while using the prototypes. All the information that was recorded during this experiment was important, because it would help us improve the systems in the future.

4.3.3 Results and Discussions

Participation

In an experimental session, each participant was allocated a total of 1 h to complete the navigation course four times using a different prototype in each run. All the participants were able to complete their session in nearly 1 h and not one exceeded the time allocated. However, 2 of the 16 participants ended their session prematurely. They were not able to complete the tasks as instructed. During their session, they both showed signs of discomfort when using the device. They were not able to interpret the soundscape or even comprehend the sound. One of them panicked when he was being blindfolded by the facilitator. When they were attempting to navigate the course using the provided prototypes, they were not only unable to follow the yellow guide path but also knocked over many cardboard boxes.

Although their sessions ended prematurely, the facilitator conducted an additional follow-up session consisting of a brief interview with these two participants.

The purpose of the interview was to investigate their behaviour further and understand the effect of the VASS system on them. In general, neither participant was able to distinguish the sound patterns encoded in the soundscape. In the interview, they clearly exhibited that they were uneasy navigating without being able to see. The fear of hurting themselves prevented them from focusing on the soundscape. They both insisted that they did not know how to decipher the soundscape after multiple trials. This could also have been caused by their poor hearing skills, which may have impeded their ability to understand the information the prototype encodes in the soundscape. These two participants were considered outliers, because thus far most of the people who used the prototypes did not complain about the systems or show any signs of an inability to start using them. To extrapolate, these incidents suggest that there may exist a group of people among the visually impaired who are unable to use this type of device on their first attempt. These people may perceive the sound in a manner that is deeply different from that in which the average person perceives it. This difference may contribute to the difficulty they experience decoding information in the soundscape and recreating mental images and spatial relations from the soundscape. The immediate solution to help this type of person is to design a special set of training material tailored for this group of people and encourage them to invest a sufficient amount of additional time in self-learning. However, if the further investigation can be continued on this group of people, it may benefit the development of VASS systems in the long term. Some hidden aspects may be discovered that can greatly improve the performance of the VASS system. Based on the knowledge, a set of tutorials can be designed to speeds up the system's learning process.

Navigation Results

Table 4.1 shows the average travel time (in seconds) for each prototype. The results are expressed as the average of the sum of the total distance travelled for all 16 participants. In the table, Prototype 1 is labelled 'P1', Prototype 2 'P2', Prototype

3 'P3', and Prototype 4, which is the prototype with a depth sensor, 'P4'. The navigation results are grouped into three separate values, the average total time, the average time it took to travel from the origin to the target, and finally the average time it took to travel in return to the origin. There was no statistically significant difference between groups as determined by one-way ANOVA ($F(3,52) = 1.9497$, $p = 0.1376$). However, according to the total average time taken for the entire navigation course, Prototype 1 was better than the other three prototypes. Prototype 1 was the only prototype that scored a total average time less than 300 s, whereas the results for the other prototypes were all above 300 s.

Table 4.1: Average travel time (s)

	P1	**P2**	**P3**	**P4**
Total	293.0714	337.9286	360.9286	366
From Origin	183.9286	199.2143	217.7857	254.3571
Back to Origin	109.1429	138.7143	143.1429	111.6429

That Prototypes 2, 3, and 4 required a significant 10% more time than Prototype 1 shows that the algorithm implemented in Prototype 1 was of help in a navigation situation. However, according to the results of the performance measurement studies (Tan, Maul, and N. R. Mennie, 2013), Prototype 3 scored best in terms of information preservation and also interpretability. Prototype 3 was expected to perform well in this experiment. As compared to the original expectations before the experiment, the results for Prototype 1 surprised us. From the opinions of the participants that were recorded post-experiment, the results were consistent with the views expressed by the participants. Most users felt that Prototype 1 generated the soundscape more quickly, which in turn facilitated fast decision making during the navigation. Prototype 1 takes 500 ms to generate the soundscape from a single frame, whereas Prototype 2 takes slightly less than 1 s to perform the same operation. Prototypes 3 and 4 take at least 2 s to generate the soundscape.

It is clear that the participants favoured the prototype that delivered a faster response. This probably reflects the fact that the role of the prototype is similar to that of vision when a human is navigating. According to Findlay and Gilchrist (2003), normal human performance relies on quick judgement rather than a long thought process before making the next decision when navigating, i.e., taking the next step. Humans naturally use gaze fixation to guide their movement to the appropriate target. This action is frequently quick and requires only minimal visual information. On average, a person fixates his/her gaze two steps ahead for a short time period (about 800 ms – 1000 ms) before taking a step (Patla and Vickers, 2003). During the experiment, the participants moved the camera in a unique manner to emulate gaze fixation. For example, when they used the prototype in the handheld mode, they moved the camera horizontally (left-right) and vertically (top-down) before taking the next step. These movements helped to increase the accuracy of their next action (N. Mennie, Hayhoe, and Sullivan, 2007). Hence, when a prototype is able to produce a soundscape in a shorter time frame, it increases the number of accurate judgements the user can make before he/she decides on the next movement.

The reason for separating the results into three different groups is that some interesting relationships exist between these three timings. The total average time, T, is the sum of the average time taken to navigate from the origin to the target, t_0, and the time taken to return to the origin, t_1.

$$T = t_0 + t_1 \tag{4.1}$$

We noticed that significant differences existed between t_0 and t_1, although ideally the two times should be approximately the same. The result for all the prototypes showed that t_1, the duration of the return journey, was significantly shorter than t_0, the duration of the journey from the origin to the target. Among all prototypes, Prototype 4 showed the most reduction between t_0 and t_1, with a 56.1% decrease in time on average. Prototype 1 was in second place with a 40.67% time reduction, while Prototypes 2 and 3 yielded similar results, with a 30.34% and 34.28% time

reduction, respectively. The results correlated with the number of obstacles knocked over in both the journey to the target and the return journey. The number of obstacles knocked over by the participants was significantly lower during the return journey than during the journey from the origin to the target.

The results for the difference in the journey duration and the number of obstacles knocked over showed the ability of the user to learn while using the provided prototypes for the first time. The improvements were apparent when the participants showed that they were able to traverse better during their return journey. The improvement for their return journey can be explained by two factors. First, the participants understood how to use the prototype better after they had completed the outward trip. Normally, before starting to use a VASS system, the user attends a training lesson. However, in this experiment no training was given except a short introduction to the operation of each prototype so that the participants would know what to expect when they put on the headphones. Therefore, the participants relied on their outward trip to learn, which in turn helped improve their return trip. According to the participants' responses and observations, they learned both how to operate the device and to interpret the soundscape. For example, it was observed that the participants operated the camera more efficiently; for example, they improved the angle at which they pointed the camera. Second, the results also suggest that the participants were able to create mental images/maps through listening to the soundscape on their outward trip. On their return trip from the target to the origin, they relied less on the soundscape because they were able to use the mental map in their mind as guidance. When they had a mental map of the navigation course, the users could focus their energy on detailed features in the soundscape, such as the boxes placed along the side of the path.

It is not clear why Prototype 4 showed the best improvement between the outward and the return journey of all the prototypes; however, the ease of mental image creation tends to be inversely correlated to the amount of information contained in the soundscape. For example, Prototype 1 converts less visual information and

therefore the user can create a mental image more easily. By using a computer to process the image frames, Prototypes 1 and 2 can discard many irrelevant features in the visual-to-auditory conversions, leaving important information in multiple blobs. With a simpler soundscape, the users can easily create a mental image in a shorter period. However, experienced VASS system users have the ability to create a detailed mental image when using a more complex soundscape. According to the feedback of the participants, the mental images/maps did not remain in their mind for the entire experiment. The map was retained only for each individual prototype period. The participants had to create a new mental image for each prototype. This may be because the users were inexperienced in using this type of device. Hence, with proper training and exposure, it is believed that users can be taught how to create rich mental images/maps from soundscapes with more information and how to prolong their retention of these mental images/maps.

Object Recognition

At the end of the navigation course, the participants were required to complete an object recognition task before they were allowed to return to the original location to end the test. The object recognition task was a simple one: the participants were asked to select a specific balloon based on the colour given by the facilitator who was monitoring them. A total of three balloons (red, blue, and yellow) were glued on the table. Before the participant reached the table and was asked to select the balloon, the facilitator randomized the location of the balloons. Each participant was given two chances to select the correct balloon and the time each participant took to complete the task and the choice(s) of each participant were recorded.

As in the previous table, in Table 4.2 Prototypes 1 to 4 are labelled from P1 to P4. The results presented are the average time taken by the participants to complete the experimental task sorted according to the prototype they used. The table shows that the results for this activity do not exhibit an obvious pattern, as do those for the navigational aspect of the experiment. This could be due to the

simplistic nature of the test, given that the entire experiment was focused mainly on the navigation task rather than on the object recognition task. In the future, to investigate the user's ability to recognize objects, improved object recognition tasks similar to those used in Experiment 1 can be included. Of the four prototypes, the users performed better using Prototype 3, with an average task duration of 72.71 s. The difference as compared to the worst performer, Prototype 4, was more than 10 s.

Table 4.2: Average balloon recognition time (s)

	P1	**P2**	**P3**	**P4**
Average Time	78.92857	80.78571	72.71429	83.28571

Because it was built on the concept of vOICe (Meijer, 1992), it is expected that Prototype 3 would perform better than the other prototypes. This was mainly because Prototype 3 converts relatively more visual information than the other prototypes, especially Prototypes 1 and 2 in which an image processing module is incorporated in the conversion process. A considerable amount of visual information is required to correctly complete a task such as this, because it requires the user to select the correct balloon together with determining its exact location based on the features of the object and its surroundings. A comparison of the results for both the navigational and object recognition task shows clearly that the different tasks required different aspects of visual information. Although Prototype 3 did not perform well in the navigational task, it performed better in the object recognition task because of its richer and more detailed soundscape.

Depth Implementation

One of the objectives of this experiment was to determine whether the depth implementation built into Prototype 4 contributed to a performance increment. Overall, according to the results of this experiment, Prototype 4 did not achieve any sig-

nificant improvement over the other prototypes. In fact, its performance was the worst for both the navigational and object recognition task. These results were not expected, especially those for the aspect of navigation, because humans naturally depend heavily on depth information when they are moving. It is clear that the manner in which depth information is implemented in Prototype 4 cannot be deemed suitable for application in VASS systems.

After the experiment, the participants were also interviewed about their opinion of Prototype 4. The participants unanimously agreed that the depth implementation did not help them visualize the surroundings better from the soundscape. As the depth information is implemented in Prototype 4 as an optional feature, most of the participants chose to switch off this option for the entire course. Some complained that the soundscape tended to be more confusing when they were navigating with the depth feature turned on.

There are two possible reasons for this poor result. The first is the particular method used to incorporate depth in the prototype. In Prototype 4, the depth map is used as an optional filter to slice the visual information so that the user can focus on a specific depth range. However, being able to choose the depth range during navigation not only caused the users confusion but also interrupted their flow of spatial reasoning. As mentioned previously, users rely on fast judgements and quick reflexes during navigation, which is why the prototype that provides a faster refresh rate frequently performs better. This kind of depth implementation necessitates an additional layer of thought process, hence slowing down the overall conversion if the user intends to switch to 3D mode. Furthermore, it adds confusion to the usual sound cacophony as the user switches back and forth between the 3D and 2D mode. The second reason is that the poor implementation may be caused by the factor of resolution provided by the TOF camera used in Prototype 4. The TOF camera provided only the QQVGA depth map resolution (160×120), which is significantly smaller than the standard resolution used for the 2D implementation. Thus, many of the visual features are not captured by the depth sensor, which lowers the overall

accuracy of the depth filter.

Although the results were discouraging in terms of depth implementation in the VASS devices, I still believe that depth information is essential for VASS systems. It has been reported that users perceive multiple depth cues even from a soundscape that is generated from 2D image frames. With the 2D depth cues, they were able to navigate through the navigation course despite the lack of accuracy of the cues. However, the VASS system can provide a more precise set of depth cues with a better depth sensor which will improve the performance of the system. Additional effort needs to be invested in researching new means of converting and integrating depth information automatically into the soundscape. It is possible that, with advanced technologies, a soundscape that maximizes the visual information while reducing sound confusion and user interaction can be generated.

4.4 Discussions

The results of the experiments provided many insights into the research. Not only did the feedback obtained from the experiments guide us to improve the prototypes, but also the results affected the direction of the Luminophonics project research. At the beginning, the aim of the project was to push the limit of visual-to-auditory cross-modality conversion in order to create a better VASS system that can completely replace the human eye by providing vision through a soundscape for the visually impaired. However, according to the results of this experiment, more work is required to achieve a device ideal for practical usage. Hence, it was decided that the research would take a slightly different route to provide the necessary infrastructure and support that would facilitate the development of future VASS systems based on the prototypes that were developed. In this section, the lessons learned from the experiments are summarized and presented.

4.4.1 Scenarios for VASS

The decision to combine both navigation and object recognition tasks in an experiment was a good one. The practice of testing a VASS system in a situation involving two different scenarios, i.e., navigation and object recognition, is not commonly implemented. However, this experiment design was created to find out which prototypes can perform well in both scenarios. The results of the search for the prototype that can perform best in both scenarios were unexpected. The results of Experiment 2 surprised us, given that some of them were unforeseen. For instance, Prototype 1, which was performed worst in terms of object recognition, outperformed other prototypes in terms of navigation. Most interestingly, the observations from the experiments revealed that the prototypes could not perform both tasks well. They all proved to be effective in one of the scenarios but performed poorly in the second. It is clear that the approach built in the prototypes is not adequate for producing a system that is sufficiently robust to perform in all scenarios.

The results of Experiment 2 showed that for the different scenarios users need to use a different aspect in order to complete the task optimally. However, very frequently the aspects contradict each other, which is the reason why it is difficult to create a perfect VASS system. In a scenario such as object recognition, the VASS should prioritize information content, because the user requires various visual features to correctly recognize the object in front of him/her. The soundscape produced to meet these requirements is slow and long because of the details of the information encoded into the soundscape. In contrast, in a scenario such as navigation, the amount of information required is minimal. In this type of scenario, speed is most important, because users need to make very fast judgements to help them move along a path while avoiding obstacles. Most of the time, a device that supplies only relevant information encoded in a short soundscape performs well in such cases. These two scenarios contradict each other: navigation requires speed but not the information richness that is important for object recognition. An intelligent

algorithm is needed to balance the two aspects so that the VASS system can be sufficiently robust to operate well in different scenarios. However, in the near future, this research project will be focused on improving the prototypes so that they excel with limited functions rather than on building a system that caters for all possible scenarios but performs poorly.

4.4.2 Disadvantages of Experiments

In addition to the two experiments presented in this chapter, several experiments were conducted on a minor scale internally to test the functionalities of the prototypes. Since the beginning of the development of VASS systems, experiments using human subjects have been the only tool available for measuring the performance of a system. Because humans are the end users of VASS systems, it is important to know how the system performs from the point of view of the user and also how the user perceives the soundscape output from the system. After reviewing the results of multiple experiment sessions, this evaluation method has some major weaknesses that might cause a bottleneck in the development of sensory substitution technology.

First, experiments using human subjects require a considerable amount of resources, including time, human effort, and money. Thus, it is inefficient to conduct multiple large scale experiments to measure different VASS systems. In order to obtain more accurate results, the sample size of the subjects must be sufficiently large to include humans of different gender, background, age, and experience. However, to recruit a sufficient number of human participants to meet the sampling requirements for experimental purposes costs money, as well as time and other resources. Moreover, time must be dedicated to conducting experimental sessions with each participant. Thus, the amount of resources needed to conduct a thorough experiment is very large. If the experiment fails, the entire experiment must be repeated, which further increases the costs incurred. Therefore, the amount of resources required has become a major hindrance for sensory substitution research, especially

for a small research group with limited funding.

Second, until now no standardized set of quality measurements has existed for VASS systems. The lack of a universal performance measurement for VASS is a concern, especially given the increasing number of similar systems that are being invented. As the number of new VASS systems being invented increases, so does the number of experiments for testing them. It has been the norm in the field of VASS that inventors create a different set of test cases to evaluate their system. This has led to many problems for the community, mainly because it is difficult to compare different systems without standardized benchmarking tools. It is important to compare the systems so that their strengths and weaknesses can be identified. As a result, the growth of VASS has been restricted because of the lack of common goals and direction coming from research, which is frequently conducted in silo. This problem is discussed in more detail in the next chapter, together with solutions proposed to overcome it.

4.4.3 Learnability

There was a slight difference between Experiments 1 and 2 in terms of allowing the participants to learn to use the prototypes prior to the actual experimental activities. The basic VASS training session in Experiment 1 that was compulsory for every participant was omitted in Experiment 2. The training phase was omitted in Experiment 2 so that the learnability of the prototypes could be examined. The results show very clearly that the prototypes developed by the Luminophonics project promote learning.

In both experiments, the participants demonstrated improvement in understanding the soundscape after their first exposure to the device. The situations of the two experiments were different. In Experiment 1, the training session was structured as a simple guided tutorial session in which the operation of the algorithm was explained to the participants. The users were blindfolded while they interpreted

the soundscape through the guidance of a facilitator. After the training session, all the users were able to start performing the experimental tasks and demonstrated the ability to interpret the soundscape in the case of a simple scenario. Although the accuracy was lower for complex scenarios that required the user to interpret different combinations of different features, this showed that even a short training session helps users succeed in learning to use the system quickly. This is probably because of the simplicity of the concept, which is easy for the user to comprehend and which they can learn intuitively.

Unlike Experiment 1, Experiment 2 did not include a training session before the actual activities. The participants were briefed only with a simple explanation on the mapping of visual-to-auditory features conversions for each prototype. According to the observations, most participants gradually learned how to interpret the soundscape while navigating the course. The feedback of the participants after the experiment showed that they mostly learned to interpret the soundscape by differentiating multiple soundscapes. The consistent differences and changes in the soundscape generated by the prototype when sonifying different environments helped the user to learn to interpret the soundscape. However, in Experiment 2, two persons were discovered among the participants were unable (or afraid) to use the device without prior training. They were considered outliers. This singular cases was very valuable for the research. It showed how valuable a training session can be.

From the experiments conducted, a few facts related to the learnability of the VASS system from the experiments were identified. First, intuitive and logical visual-to-auditory conversion algorithms facilitate the learning process of the users. When designing the conversion algorithms, it is important to consider human psychology in the basis of the cross-modality conversion. The feedback from the survey showed that this was beneficial, because the users felt that the soundscape was intuitive. The results of the experiments also indicated that the prototypes are easy to learn. Second, not all individuals have the capacity to start using VASS quickly.

For example, in Experiment 2, two participants were observed to unable to use the device on their first trial. The reason why these subjects in particular could not use the device is unclear. However, in the interview session, it was revealed that it could be related to psychological factors and personal experience. It is to be hope that this can be solved by providing an appropriate structured training session.

Finally, I strongly believe that a training session is very important for increasing the learnability of a VASS system. Because of their different personalities and experience, different individuals may react differently to the soundscape. Although the conversions can be designed such that they were as intuitive as possible, without proper guidance some individuals may interpret the soundscape differently. Exactly as for any new tool, a syllabus of training material need to be constructed so that every individual can fully grasp the concept of the conversion and interpret the soundscape more efficiently. An appropriately structured training session can both promote the learnability of all users and ensure that every user can learn at the same pace, regardless of their past experience.

4.4.4 Exterior and Hardware

In Experiment 2, the prototypes were presented in two forms, head-mounted and handheld. The source of the intuition that motivated us to develop these two forms was humans' natural navigation gait and stance. In the head-mounted mode, the level of the device emulates humans' natural eye level and the user can how move his/her head in the same manner as he/she would move it to control the angle of his/her vision. People move their head to control their field of view so that they can focus on the desired scene. When the prototype is used in handheld mode, the user holds the camera in the palm of his/her hand, waving it around in the same manner as a blind person waves a mobility cane. These two forms were tested in Experiment 2. The participants were given the option to use either of the forms and were allowed to change their preference at any point during their navigation.

As reported, all the participants finally chose to use the handheld mode.

The majority of the participants demonstrated difficulty using the prototype when they wore the camera on the top of their head. The handheld mode offers more flexibility and the user can move the camera into any position. This provides more degrees of freedom and a longer reach than the head-mounted mode, the field of view of which is limited by how far the user can turn his/her head. Although humans' eyes are located in their head, this may not be the optimal position for the camera in a VASS system. In addition to more degrees of freedom it offered, one of the reasons why the users preferred the handheld over the head-mounted mode may be the limitations of the camera used. As compared to human eyes, the camera used in the prototypes has a narrow field of view. Normal peripheral human vision has 180° of horizontal field of view and about 135° of vertical field of view (Strasburger and Pöppel, 1999). However, the DepthSense DS311 camera that was used in the prototypes has 50° of horizontal field of view and 40° of vertical field of view. Therefore, to compensate the limited field of view of the camera, the users preferred the handheld mode, which allowed them to expand the field of view of the camera. Furthermore, using reach a user can extend his/her hand forward to 'zoom' in on his/her field of view to capture more detail when needed.

In summary, these experiments confirmed that it is preferable to base the design of the VASS system on the handheld form, which provides users with more control of the visual information they hear. The outcome further confirmed the validity of the decision to incorporate the VASS system in a smartphone. As the smartphone is becoming the most widely used handheld mobile device, it is natural that an increasing number of VASS systems will be designed to be installed as a mobile application in a smartphone. Because of the size of the device and the camera placement (at the rear of the phone), it is indeed the perfect platform for VASS systems. Using a VASS system installed in a smartphone, the visually impaired user can sonify the scene towards which the smartphone is pointing. The fact that the use of smartphones is common allows the user to utilize the VASS system more freely

in public without appearing awkward. However, the investigation of the feasibility of using VASS in head-mounted mode should not end here. It may be that when a user becomes expert in using the VASS system (after long term usage), he/she may prefer the head-mounted mode. This is because when using the head-mounted mode the user's hands are free for performing other tasks, such as grasping and touching.

Chapter 5

Measurement and Optimization

5.1 Introduction

Presented in this chapter are the studies conducted that focus on the subject of improving the VASS system through the application of optimization techniques. According to the results of the experiments conducted after the prototyping phase, the performance of the prototypes is indeed better than that of some existing counterparts. However, the path of improving the visual-to-auditory conversions eventually led us to two main challenges. They are related to the questions of how the improved prototypes perform as compared to other systems and to what extent the performance of the proposed technology can be improved. Therefore, the idea of using an optimization process was conceived to meet these challenges and eventually to further elevate the performance of the prototypes.

The main objectives attempted to achieve were to determine which features constitute a good VASS system, and using this information, to determine through optimization a set of more effective feature pairs for the visual-to-auditory conversions mapping. However, in order to arrive at the process of optimization, the challenge was to obtain a cost function that quantitatively measures the performance of a VASS system and that can be optimized. Unfortunately, there are not many

tools for specifically measuring the performance of a system that converts visual to auditory information. Although there are many systems that translate visual information into the auditory form, there is no standardized method for correctly measuring the performance of a VASS system. Most VASS research groups conduct their own performance measurements in their own facilities, such as the experiments described in Chapter 4. As a result, activities were planned first to produce a set of performance measurement tools and then to use it to start the optimization process to determine the best visual-to-auditory conversions features.

In Chapter 5, the work that will be described is grouped into two main sections, addressing the automated measurement (Section 5.2) and subsequently the optimization (Section 5.3) of the visual-to-auditory conversion features. The purpose of developing the automated measurement was to assess the performance of a VASS system quantitatively, targeting certain elements of the system. For this purpose, an innovative approach for evaluating the performance of a VASS system from its soundscape by examining both the information preservation and the interpretability of the content is introduced. Then, the process of optimization is developed on the conversion parameters using the evaluation methods developed earlier as the cost functions. Finally, two sets of visual-to-auditory conversion feature mappings, which are optimized for two different scenarios are described in this chapter.

5.2 Automated Measurement

After the prototype development and conducted the experiments, the next step was to evaluate the performance of the prototypes in comparison with that of the other existing VASS systems. The purpose was to determine how the prototypes compare with the established systems and also to identify the improvements that are achieved in the prototypes. However, without standardized tools and guidelines, it is very difficult to compare the systems fairly. The main factor behind this problem is probably the fact that most VASS research is done in silo and there is minimal

collaboration between groups. This has resulted in a situation where different sets of experiments for performance evaluation have been created for each of the different systems. Their performance is measured according to different criteria and requirements, and thus, it is difficult to compare the measurement results.

Relatively frequently the evaluation methods are focused on the features of the prototype, which adds an additional layer of difficulty in terms of comparison because some other systems do not have the same feature package. For example, See ColOr (Bologna, Deville, Pun, and Vinckenbosch, 2007) performs very well, especially in terms of object manipulation and human navigation. The authors proposed testing their system by introducing experiments that measured the interaction of users with the system. It is easy to evaluate a system from the reported results of the experiments but it is not possible to compare it with other systems. As more systems are created, different kinds of tests are created together with them. In principle, internal experiments can be conducted to compare the prototypes with systems such as vOICe (Meijer, 1992) or the Raster Scanning method (Yeo and Berger, 2006), but the results would be skewed towards my prototypes. This is because the two systems do not share a common set of features with the prototypes and, most importantly, the systems emphasize sonifying the image texture rather than other information, such as colour. Since the experimental tests focused on the users' ability to interpret the colour information in the soundscape, the systems that do not incorporate colour in the conversions would be in an unfavourable position. Therefore, a common evaluation method that quantifies the performance of a system based on the content of its soundscape is preferable, because it provides a fairer platform on which to measure the performance regardless of the features of the VASS system.

Most of the current preferred measurements are based primarily on the approach of psychological experiments, such as the experiments described in Chapter 4. These experiments employed human subjects in order to gather their feedback on the sensation and perception of the systems of interest. Hence, by examining the interactions between the users and the systems through a series of carefully planned psycholog-

ical experimentation, it was possible to gather more intrinsic data that benefit the development of the systems. The team from the University of Trier also achieved good results from psychological experiments. They developed this type of experiment to evaluate the effectiveness of their visual-to-auditory conversion system on the perception of the representation of colour and also the effect of learning when people use these systems (Michael J. Proulx et al., 2008). From the data they collected, they were able to understand the learning effect and the human reactions to the system. Ultimately, psychological experiments are indispensable for evaluating an SSD because of their comprehensive results. It is therefore advisable to conduct such experiments before deploying an SSD. However, this remains a very expensive approach (in terms of both time and money) and thus is unsuitable for the early exploratory phase of a project. A more suitable approach for the testing, evaluation, and filtering of prototypes at the early developmental stage would have to rely on some type of mathematical measurement of the conversion process. If this measurement were, moreover, to avoid certain biases, it could even be instrumental as a universal method for comparing systems. Although it would never replace psychological experiments, it would allow cheap and effective prototype exploration and would add some objectivity to the comparison of systems.

The unavailability of a common performance measurement platform on which to evaluate the systems is a major concern for VASS researchers, as an increasing number of systems are being produced. In order to improve the image sonification technique, essentially systems comparison is inevitable. It is important to compare the systems so that their strengths and weaknesses can be identified. Without a standardized performance measurement, it is difficult to rank the systems fairly based only on the results generated by human test subjects.

5.2.1 Overview

In this section, a performance measurement is proposed. The objective of such measurement is to quantify the performance of a VASS by analysing both the input and output of the system. The evaluation method is aimed to be both fair to all systems and cheap to implement, and thus, it is ideal as a common performance measurement platform for VASS systems. The performance measurement addresses two main issues: the interpretability and the information preservation of a soundscape. Although, the quality of a soundscape is not limited to only these two components, they are valuable indicators of a good VASS system. The other features that contribute to the effectiveness of a visual-to-auditory conversion process include high learnability, a good listening experience, and the robustness of the system.

The measurements of interpretability and information preservation are independent of each other. The two measurements use a different set of visual and audio analysis algorithms to achieve their purpose. The interpretability of a soundscape answers the question of how feasible it is that a human user can interpret or learn how to interpret the generated soundscape. A highly interpretable soundscape is easy to learn and also perceive. It can be measured by analysing both the input visual information and the generated soundscape to determine the connection between the input and output. Frequently, the input and output information of a soundscape that is easy to interpret are highly correlated. Based on this premise, an interpretability measurement that uses components such as inter-image distance (IID) and inter-sound distance (ISD) is being proposed.

Second, the performance measurement includes the ability of a VASS system to preserve information during the process of visual-to-auditory conversion. As mentioned, a visual container is able to accommodate more information than an audio container of similar size. Therefore, the process of cross-modality conversion from a visual to an auditory domain usually results in information loss. However, different conversion methods cause different degrees of information loss, one higher

than another. By calculating the amount of information in the input image and the corresponding generated soundscape, the amount of information that is preserved during the conversion process can be gauged. The objective of this measurement is to search for a system that is able to encapsulate the most information in the soundscape.

5.2.2 Measuring Interpretability

As suggested previously (see Chapter 3.2.3), the process of optimizing the selection of a musical instrument set paved the way to the establishment of an interpretability measurement. Briefly, the goal of this optimization was to produce a set of 10 distinctive timbres. This was achieved by analysing the sound signature of each timbre and measuring the differences between these signatures to map the differences between all the timbres. The process of characterizing the timbres through MFCC and calculating the ISD based on their sound signatures was significantly instrumental for developing a method of measuring interpretability.

From the results of measuring the distance between audio signals, there is a strong connection between the distance between input images and their corresponding output soundscape and the interpretability of the soundscape. More intuitively, if the soundscapes generated by a system are to possess the property of interpretability, then, if two images are similar, their corresponding soundscapes should be similar, and conversely, if two images are different, then their corresponding soundscapes should be different. This property can be easily captured by a correlation measure. Based on that idea, the analysis of IID and ISD is introduced. IID is the similarity metric between two input images. Similarly, ISD measures the similarity between two soundscapes. It can be hypothesized that the correlation between IIDs and the corresponding ISDs measures to a significant degree the interpretability of the soundscapes generated by a VASS system. In other words, it is expected that a good VASS, i.e., one that allows easy interpretation of soundscapes, will exhibit a

relatively strong IID-ISD correlation.

Inter-image Distance

The algorithm for image similarity measures has advanced tremendously, aided by the growth of visual data and its applications, such as in image retrieval as used in most of the search engines available on the Internet. Although there are several readily available algorithms from which to choose, EMD was chosen in this study to measure the distance between two input images. EMD, which is also known as the Wasserstein metric, was first proposed by Peleg, Werman, and Rom (1989) to measure the distance between greyscale images. It describes the minimum cost of changing a probability distribution to another probability distribution. In this case, a representation scheme was created based on a pair of probability distributions for both images. EMD measures the lowest cost of transforming image A to image B using this representation scheme.

In the research studies, an advanced form of EMD is applied. It is commonly used for the purpose of image retrieval, developed by Rubner, Guibas, and Tomasi (1997). Instead of probability distributions, vector quantization was used as the basic representation scheme. This distance measurement is called EMD-KL , the name of which reflects that it is a combination of EMD and Kullback Leibler divergence. It improves the results by taking into account the perceptual similarity of images, and thus, the approach is considerably more robust and well suited for applications that involve colour and texture information. An additional reason why EMD-KL was used for calculating the IID is that it can also be used to calculate the differences between two audio signals (Logan and Salomon, 2001). By using the same algorithm for calculating both the input (image) differences and output (soundscape) differences, the need to normalize the results so that they were compatible with each other can be eliminated.

Inter-sound Distance

The task of computing the ISD between two soundscapes relied on an existing MATLAB toolbox proposed by Pampalk (2004), called the MA Toolbox, that computes music similarity from audio. The toolbox contains different approaches that focus on audio similarity measures, including content-based music analysis, feature extraction, and visualization. For ISD, the methods to extract the sound signature, which is similar to the approach used for the timbre selection, discussed in Chapter 3.2.3 is used. However, instead of using different timbre sound snippets, the extraction method was applied to the output soundscapes generated by the VASS system.

To calculate the ISD, a music similarity that combines the two approaches proposed by Aucouturier and Pachet (2004) and Logan and Salomon (2001) respectively was designed. Figure 5.1 shows the process of the ISD measurement for soundscapes. Basically, the ISD measurement method modifies the approach of Aucouturier and Pachet (2004) by replacing the Monte Carlo sampling with EMD-KL for the distance measurement. From this, the ISD measurement comprises a three-part process. First, the soundscape (in the form of audio signals) is chunked into multiple frames. Each frame is then transformed into MFCCs. An MFCC consists of a group of coefficients formed by the MFC, which is a representation of the sound spectrum on a non-linear mel scale frequency. MFCC is used because it records the features of the soundscapes; the sound analysis of the features is easier than that of a raw audio signal. In order to correctly model the audio frames, a Gaussian mixture model (GMM) is then used to cluster the audio frames (in MFCC) into one cluster model. Finally, the distance is computed by comparing two soundscape models using EMD-KL.

Process

The measure of interpretability is calculated based on the correlation between a list of IIDs and ISDs produced by a VASS system. So that a VASS system can produce

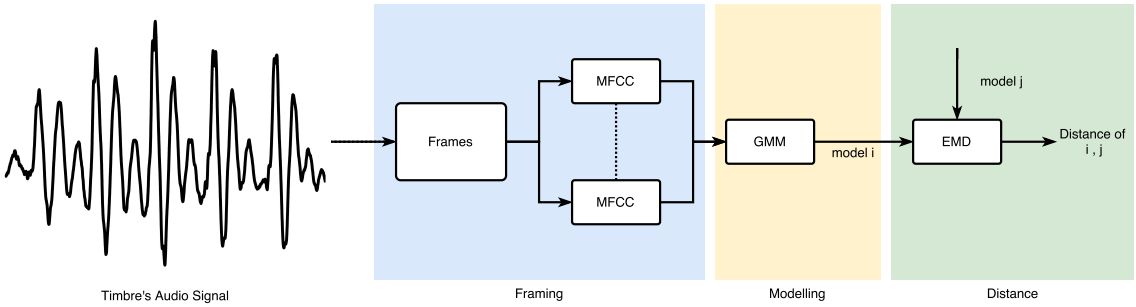


Figure 5.1: Inter-sound distance process

a highly interpretable soundscape, the distance between two input images must be highly correlated with the distance between the two corresponding generated soundscapes. Pearson correlation coefficient (PCC) is proposed to measure the correlation between the IID and the ISD using

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y}$$

where: (5.1)

$\bar{X}$ and $\bar{Y}$ are the sample means,

S_x and S_y are the standard scores

Numerous correlation coefficients are available, but PCC was chosen mainly because essentially the IID and ISD pairs are normally distributed bivariate data. However, the choice of correlation coefficients is not limited to PCC; other similar correlation coefficients, such as Spearman's rho coefficient and Kendall's tau coefficient, can also be considered for this purpose. Further investigation may show that they describe better the relationship between the IID and the ISD.

Results

To test the interpretability measurement, it was implemented on three prototypes and one external system, vOICe.

Table 5.1 and Figures 5.2 to 5.5 show the results when the interpretability measurements were applied to Prototypes 1 to 3 and vOICe. The blue dots in the scatter plots show the value of the IID vs the ISD and the orange line shows the PCC of the values for each plot. The scatter plots clearly show that the correlation of the prototypes is better that that of vOICe. Prototypes 1 and 2 scored a significant advantage over Prototype 3 and vOICe. This shows that, when image segmentation is applied, the generated soundscapes are considerably more correlated with the input image.

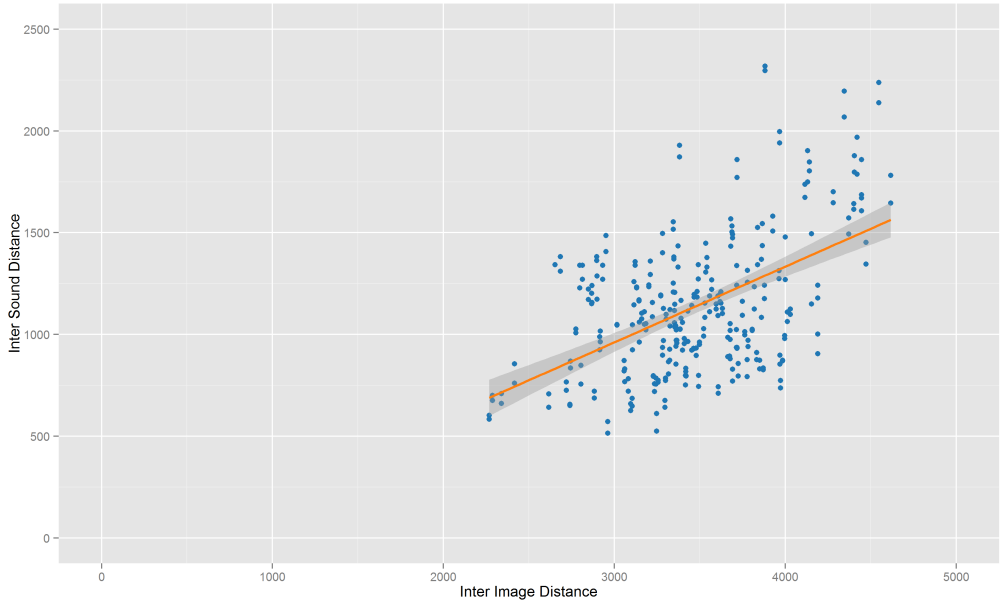


Figure 5.2: Correlation of Prototype 1 (value: 0.5209)

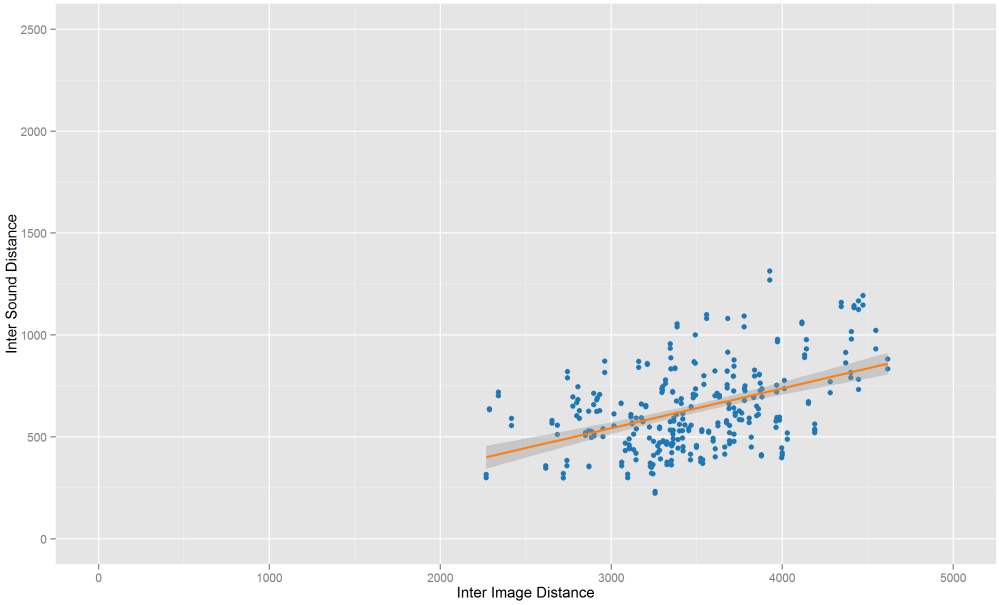


Figure 5.3: Correlation of Prototype 2 (value: 0.4546)

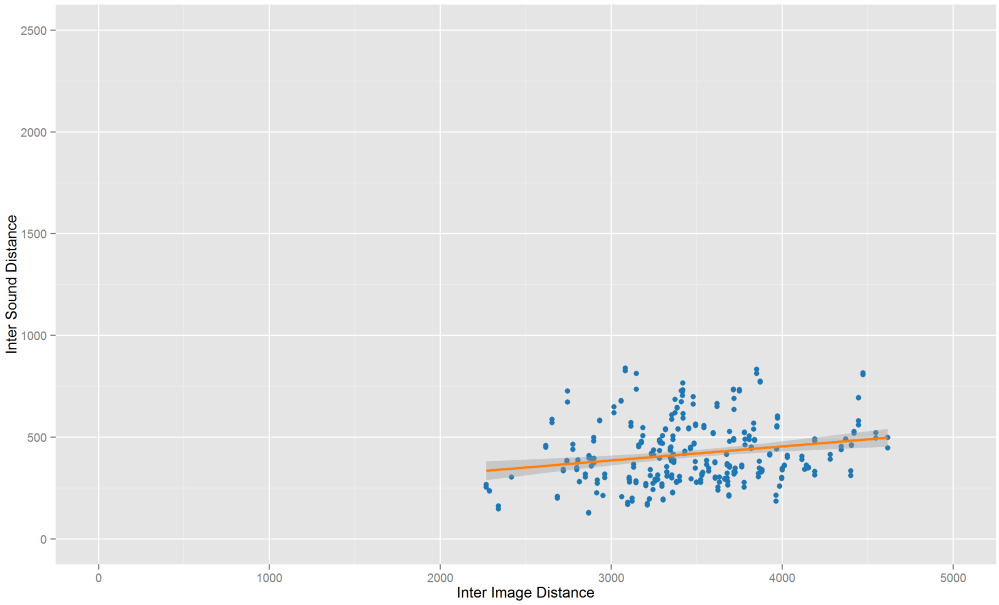


Figure 5.4: Correlation of Prototype 3 (value: 0.2142)

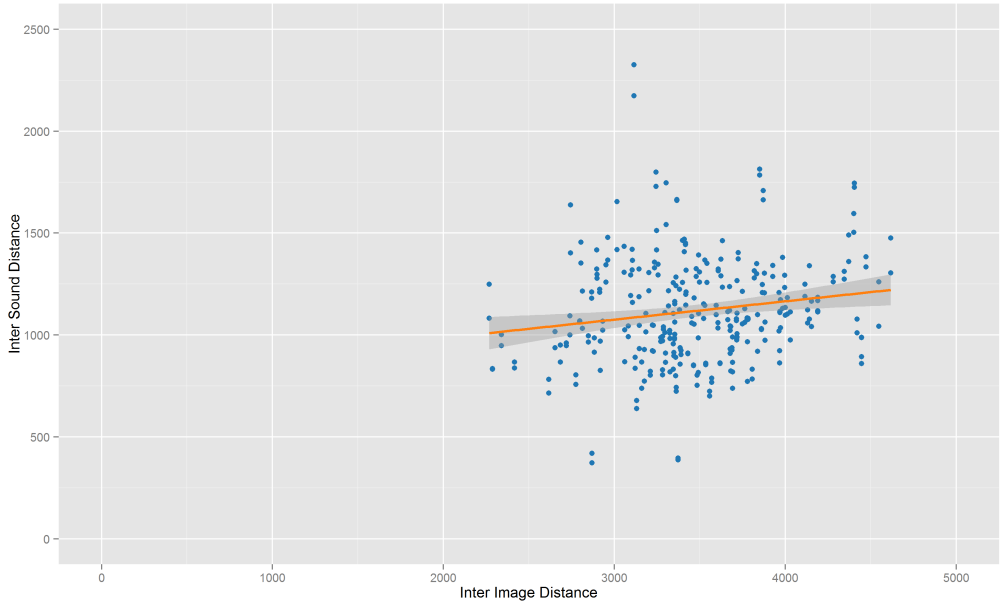


Figure 5.5: Correlation of vOICe (value: 0.1650)

Table 5.1: Pearson correlation coefficient of inter-image distance and inter-sound distance

	Correlation Value
Prototype 1	0.5209
Prototype 2	0.4546
Prototype 3	0.2142
vOICe	0.1650

One of the possible explanations for this situation is that the fundamental sonification processes of Prototype 3 and vOICe are similar: the conversions of both are based on the value of the pixel. Because they do not implement an additional image processing step before the sound synthesis step, all the information (including noise) is translated into the final soundscape. However, because of the blobbing implementation in Prototypes 1 and 2, the soundscape is considerably simpler. The

blobbing technique groups the pixels into multiple blobs, reducing the noise, and hence effectively filters most of the noise, resulting in a cleaner soundscape. Therefore, it can be confidently stated that the soundscapes of the prototypes in which blobbing is implemented as the feature extraction technique have a better correlation than those of the prototype in which it is not. Because it is easier to interpret a soundscape that correlates well with the input image, indirectly the interpretability of Prototypes 1 and 2 is higher than that of Prototype 3 and vOICe. Although they lost detailed information, they gained an advantage in terms of interpretability.

5.2.3 Measuring Information Preservation

Another aspect that is important for describing the performance of a good visual-to-auditory conversion is the extent to which the algorithm can preserve the information. Information preservation in visual-to-audio conversion is hotly debated and frequently difficult to achieve because, ultimately, converting information from a visual to an auditory form is a process of information reduction. Najjar (1996) suggested that spatial and recognition information is represented better by pictures. Similarly, Stoneman and Brody (1983) found that children subjected to visual or audiovisual commercial presentations could recognize advertised products more effectively than children subjected only to audio presentations. The papers reporting these two studies presented an important piece of information very relevant to the research studies in the context of information preservation. The studies showed that the amount of information contained inside an image allows very rich cognitive encoding that allows high recognition rates as compared to audio.

During the process of visual-to-audio conversion, information is lost primarily through dimensional reduction, i.e., the conversion of a 2D signal to a 1D signal. So that the users can make sense of the generated soundscapes, a sufficient amount of visual information must be preserved in them. In order to utilize the spatial information encoded in an auditory signal, users need to reconstruct mental images

by listening to the soundscapes. To ensure that the image sonification is effective, it is required that the reconstructed mental image be sufficiently similar to the real image. Ultimately, although the soundscapes generated by VASS systems experience severe information reduction, certain aspects of the visual signal, such as spatial information, needs to be retained. Over time, users can learn how to reconstruct the spatial information embedded in soundscapes.

The amount of information preserved needs to be moderately controlled. In most cases, the more visual information that is preserved in soundscapes, the better. When more information is preserved, there are more features that can be interpreted by users. For example, if sufficient information is preserved, the user may be able to interpret features such as spatial relationships, shape, colour, shade, texture, and motion. However, an excessive amount of information encoded in soundscapes may lead to the user being overwhelmed and/or confused. Conversely, systems may lose some of their usefulness if they do not preserve the appropriate amount and type of visual information. For instance, vOICe encodes only greyscale pixels into sound frequencies (Meijer, 1992). As a result, the colour information is lost during the conversion process, when there may be many situations where colour is essential for decision making. To conclude, in information preservation a balance needs to be maintained between sufficiency (too little leads to debilitated decision making) and excessiveness (too much leads to cacophony).

In this section, a method to measure the amount of information preservation by examining the input image and the corresponding generated soundscape is proposed. The main objective of this measurement is to identify the algorithms that are able to retain the most information during the conversion. Although this measurement cannot indicate a system that is able to retain the correct type of information, hopefully it can be the precursor in the search for the most efficient visual-to-auditory conversion.

Entropy

The degradation of information can be represented by calculating the difference between the entropy of the input images and that of the corresponding output soundscapes. An approximate measure of information preservation can be obtained by estimating the 'quantity of information' in images and their corresponding soundscapes and then calculating the difference between these quantities. For this purpose, a measure widely used in information theory that was introduced by Shannon (1948) is being utilized, that is, entropy. The basic equation for entropy is

$$H(X) = -\sum_{i=0}^{n} P(x_i) \log_2 P(x_i) \tag{5.2}$$

Entropy is used to measure the unpredictability and uncertainty of a random variable. In Equation 5.2, entropy, H of X is calculated on the number of bits needed to transmit the probability occurance of x. In essence, the easier a variable is to occur, the lower is its entropy. In the case of this research, an assumption is made that entropy is directly related to the information contained/encoded within the signal.

By computing the entropy of the input images and their corresponding soundscapes, pairs of values that overall represent the information preservation capabilities of a particular visual-to-auditory conversion is obtained. Because of the dimensionality reduction aspect of the conversion process, the entropy of a soundscape should in general be lower than that of its image. By averaging the differences for every matching input (image) and output (soundscape) entropy, the relative effectiveness of a system in terms of information preservation can be obtained. The lower the difference between the entropy of the output and its input, the higher the ability of the system to retain information during the conversion process. By using this method as a measurement standard, a comparison of systems in terms of information preservation is made possible.

Results

To test the information preservation measurement, three Luminophonics prototypes and an external system (vOICe) were tested. A set of 40 input images were prepared. The same process, in which the information preservation was computed for every image by comparing the information in the input image and the output soundscape, was applied to all the systems. Finally, the values were averaged to form the average information preserved in every system.

The results in Table 5.2 show the average amount of information preserved by each system. The table shows that Prototype 3 on average preserved the most information during the conversions, whereas the score of vOICe was the lowest among the four systems. Whereas Prototypes 1 and 2 lost considerably more information than Prototype 3, their soundscapes still contained more information than vOICe. As shown by the results presented in the previous section (Section 5.2.2), the soundscapes of Prototypes 1 and 2 are better in terms of interpretability than those of Prototype 3 and vOICe. The difference between Prototype 3 and Prototypes 1 and 2 is the implementation of the blobbing technique for feature extraction. While the soundscapes of Prototypes 1 and 2 do not contain as much information, the information retained is of a higher level. For example, the objects in the image are represented as shapes rather than pixels. Essentially, higher level information reduces the space needed to store the information and helps the user interpret the soundscape. However, the disadvantage is that, because the information is simplified in the case of Prototypes 1 and 2, they cannot describe the surroundings in as much detail as Prototype 3. Prototype 3 remains the preferred choice for tasks such as object recognition that require finer details.

Table 5.2: Average amount of information preserved during conversion

	Average Information Preserved (%)
Prototype 1	50.641
Prototype 2	48.966
Prototype 3	54.656
vOICe	42.447

An additional key fact shown in Table 5.2 is that all the prototypes scored better than vOICe in terms of information preservation. The main reason for this is that the prototypes were able to encode colour information, whereas the developers of vOICe decided to discard colour information. Without colour information, the information in the resultant soundscape is significantly lower. This is clear in a comparison of Prototype 3 and vOICe, because they both implemented a similar swiping mechanism and sonified the raw pixels instead of blobs. The difference is very large, with a 12.209% drop in information preservation. It is therefore preferable to sonify colour information in order to preserve the content, as well as because colour is frequently required in various tasks that humans undertake.

5.2.4 Discussion

Tools that measure information preservation and interpretability through the correlation of the IID and ISD were proposed above. Comprising only two measurements, they are not intended to be the definitive standard measurements for all VASS systems. These tools are a step towards the accurate and automated prediction of the effectiveness of a VASS system when used by human subjects. It is however recommended that these measurements be extended and improved to provide a more detailed description of the performance of a VASS system. Other measurements relevant to the prediction goal mentioned above should be developed. These include measures of the naturalness of generated soundscapes and estimations of the

learning complexity of different conversion processes.

Table 5.3: System ranking according to interpretability and information preservation

	Interpretability	Information Preservation	Average
Prototype 1	1	2	1
Prototype 2	2	3	3
Prototype 3	3	1	2
vOICe	4	4	4

For the purpose of improving the IID/ISD measurements, it would be useful to conduct a systematic study to determine which similarity measures are more adequate from the human perceptual point of view. In particular, it would be interesting to use human measurements of image or sound similarity and compare them with automated similarity measurements. In this context it would also be pertinent to further investigate the relative suitability of different pre-processing and feature extraction methods.

5.3 Optimization of the Visual-to-Auditory Conversions Features

Finding the best visual-to-auditory feature mapping is probably one of the most important processes in the development of a good VASS system. The performance of a system and the effectiveness of the soundscape are deeply connected to the conversion mapping. When the visual-to-auditory mapping is good, a VASS system is able to synthesize soundscapes more effectively, thus ensuring their higher quality. Frequently, it is easier for the user to interpret a soundscape that is converted using a good mapping. Moreover, a soundscape is able to describe the surrounding more vividly when more converted visual information is encoded in it. Furthermore, a slight error in the mapping may degrade the entire performance of the conversion.

An example is the problem that was encountered during the selection of a timbre set (as discussed in Subsection 3.2.3), where the soundscapes produced by the earlier prototypes introduced the effect of cacophony, so that they were more difficult to interpret. The reason was that a poor timbre set was used, causing the entire feature mapping to be ineffective, and the sounds that were generated were difficult to distinguish.

However, the creation of an optimized visual-to-auditory mapping can also be one of the greatest challenges in the development of a VASS system. The mapping of the correct visual properties to the corresponding auditory properties for cross-modality conversion is a complicated and difficult task. One of the reasons is the large amount of visual features that need to be converted. Because of the difference in the modality (visual and auditory), not every property in the output audio is compatible with the input visual features. This limits the scope of the visual features than can be converted. For instance, properties such as audio pitch and volume are more suitable for representing linear values than categorical information such as types of colour. Hence, a filtering process must be established so that only the most relevant visual information is selected for retention. The selection process can be based on the criteria and the objective of the system. However, the most difficult aspect of creating the mapping lies in the large number of combinations of configurations that are involved.

Let us take the case of mapping pixel intensity to the corresponding sound properties. There are a number of choices of features into which the pixel intensity can be mapped, such as sound pitch, sound volume, and sound amplitude. It is not reasonable to develop a separate prototype for each combination of pixel intensity and feature and then conduct an experiment for each of the prototypes. The search space became even larger after the inclusion of colour information into the conversion. When there are 10 different colours and many more musical instruments, the number of combinations can be extremely large. When every possible visual feature and the available audio properties are combined, the number of mappings is so large

that testing every mapping manually one by one is not feasible.

During the research, most of the time was spent looking for the most suitable mapping for the prototypes in order to elevate the performance of the soundscape. It was nearly impossible to test every mapping combination, because the search space was too large. However, after developing the automated measurements (discussed in Section 5.2), the idea of applying an optimization algorithm with the measurements as the cost function to automate the search for the optimized visual-to-auditory feature mapping was conceived.

5.3.1 Overview

In this section, a new method for searching the optimal visual-to-auditory feature mappings based on the automated measurements described in the previous section is examined. As described, the large number of visual features and many different variations of mappings resulted in an extremely large number of combinations, also known as a combinatorial explosion. Therefore, the best approach for finding the optimal mapping is to use computational optimization techniques. It was decided that an evolutionary algorithm is applied to address the problem, which required us to search for a high quality solution from a large number of combinations. For this case, an evolutionary algorithm called covariance matrix adaptation evolution strategy (CMA-ES) is used as the optimization algorithm.

Initially, the goal was to search for the optimal feature mapping that satisfied two objectives, interpretability and information preservation. However, from the results of Experiment 2 (see Section 4.3) a system that attempts to maximize both objectives, thus placing the solution at a Pareto front, essentially lowers its overall user performance. Let us take the example of the navigation scenario. It is not practical to maximize the amount of information to be encoded in a short soundscape. In order to include a higher level of detail, such as the texture of the objects, the soundscape has to be relatively long. However, long soundscapes do not allow the

user to make a quick decision as is required in scenarios such as navigation. For this reason, soundscapes having a higher level of interpretability, such as those produced by Prototypes 1 and 2, performed better in the navigation scenario in Experiment 2. The information delivered by the soundscapes is of a higher level, which is more easily understood by the user in a shorter time frame. Hence, it was foreseeable that mapping optimized for both interpretability and information preservation would result in a poorer system. Instead of attempting to use a multi-objective optimization approach, I decided to produce two different mappings, each optimized for a different objective, that is, one for information preservation and one for interpretability.

To use the algorithm, a set of parameters, better known as a genotype were established. A total of 18 parameters (including 10 different colours) were selected. They are discussed further in Subsection 5.3.1. For each generation, the algorithm yielded a series of mapping candidates according to the parameters. The candidates were then transferred to the VASS framework. Using the candidates for mapping, the framework was able to transform into a new VASS by changing the visual-to-auditory conversions according to the mappings of the candidates. The fitness of the candidates was evaluated using the soundscape they produced based on the measurements that were previously developed (Section 5.2). The optimization was run twice, once with the interpretability cost function and a second time with the information preservation cost function. To obtain the mappings optimized for interpretability, the interpretability measurement was used, that is, the correlation between the IID and the ISD. To obtain the mappings optimized for information preservation, the fitness of the candidates was calculated using the information difference between the input and output.

Many iterations of the process were executed, yielding variations of candidates in each generation. A soundscape based on the feature mapping proposed by the candidates was produced and then evaluated. The fitness of each generation was monitored until the fitness value was found plateaued at a range below the satisfactory threshold. The genotype of the final generation was then promoted as the

best candidate yielded by the process and consequently graduated as the feature mapping optimized for the objective.

Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

CMA-ES is an evolutionary optimization algorithm. It has been under continuous development for more than twenty years worldwide in many research laboratories since it was introduced by N. Hansen and Ostermeier (1995).

Algorithm 2 Basic CMA-ES Pseudocode

Input: m, σ ▷ Initialize mean and step-size

1: $C \leftarrow I$ ▷ Initialize a symmetric covariance matrix

2: **while** stopping condition not met **do**

3: **for all** $i \in \{1, ...\lambda\}$ **do** ▷ Sample and evaluate offspring

4: $x_i \leftarrow \mathcal{N}(m, \sigma^2, C)$ ▷ Sample multivariate normal distribution

5: **end for**

6: **sort** $\{x_i\}$ according to $f(x_i)$ ▷ Sort offspring according to its fitness

7: $x^{'} \leftarrow \sum_{i=1}^{u} w_i x_i$ ▷ Update mean

8: $p_s \leftarrow (1-c)_s \cdot p_s + \sqrt{c_s(2-c_s\mu_{eff})} \cdot C^{1/2}(m^{'} - m)$ ▷ Update isotropic evolution path

9: $p_c \leftarrow (1-c)_c \cdot p_c + \sqrt{c_c(2-c_c\mu_{eff})} \cdot (m^{'} - m)/\sigma$ ▷ Update anisotropic evolution path

10: $C \leftarrow (1 - c_1 - c_u) \cdot C + c_1 \cdot p_c p_c^{\intercal} + c_u \cdot \sum_{i=i}^{\mu} w_i((x_i - m)/\sigma)((x_i - m)/\sigma)^{\intercal}$ ▷ Update covariance matrix

11: $\sigma \leftarrow \cdot \exp(\frac{C_s}{D_\sigma} \cdot (\frac{\|p_s\|}{x_d} - 1))$ ▷ Update step-size using isotropic path length

12: $m \leftarrow m^{'}$

A basic version of CMA-ES is shown in the pseudocode above (See Algorithm 2) (Krause and Glasmachers, 2015). CMA-ES was chosen because it is effective for solving difficult non-linear and non-convex problems. The algorithm both oper-

ates at a reasonable speed and is very suitable for such problem, having a search space dimension of 18 parameters. To achieve the research objective, libcmaes was used. Libcmaes was proposed by Benazera (2015), a multithreaded high performance C++11 CMA-ES library that implements several flavours of optimization algorithms belonging to the CMA-ES family. For the use case, an algorithm called Active CMA-ES (aCMAES) from the library was applied. It appeared to be the most suitable algorithm for the purpose, because it implements a scaling strategy whereby it increases the number of offspring per generation if it fails to find a good minimum. In addition, aCMAES was observed to incur lower memory footprint than basic CMA-ESs (Arnold and Nikolaus Hansen, 2010). Because of its efficient memory management, the duration of the entire optimization process was reduced.

Parametrization

A total of 18 parameters were used in the process of optimization. Table 5.4 lists all the parameters that were applied in the CMA-ES algorithm. Because all the parameters were represented as a floating point number, the categorizations under each parameter were grouped according to the floating point value. For example, as there are two swiping directions, the top-down swipe was represented as any value between 0 and 0.4 and the left-right swipe as any value greater than 0.4 up to 1.0. The candidates yielded by the genetic algorithm bearing the values (in floating point) for each parameter were translated into the feature mapping accordingly.

Table 5.4: List of parameters for optimization

No	Parameter	Description
1 - 10	Colours	Map to musical instruments
11	Volume	Maximum volume level
12	Segmentation	0.0–0.4 for pixelation, 0.5–0.9 for blobbing
13	Stripe Width	Width of stripe in pixels (only for blobbing)
14	Blob Size	Minimum blob size (only for blobbing)
15	Pixel Height	Height of pixel (only for pixelation)
16	Pixel Width	Width of pixel (only for pixelation)
17	Swiping Direction	0.0–0.4 for top-down, 0.5–0.9 for left-right
18	Input Image Resolution	Video graphics array resolutions

1. **Colours**

   Parameters 1 to 10 were used to address colour types, one for each of a total of 10 colours. Each colour could be matched with 20 different musical instruments.

2. **Volume**

   Parameter 11 was used to determine the maximum volume that should be applied to the conversions. The range of the value was from 0 to 10. For instance, if the parameter value was 0.4, the maximum volume of the conversions was 4.

3. **Type of Segmentation**

   Parameter 12 was used to determine the type of segmentation the conversions should use. For value < 0.5, the segmentation mode was pixelation, as implemented in Prototype 3. For value $\geqslant 0.5$, the segmentation mode was blobbing,

as implemented in Prototypes 1 and 2.

4. **Stripe Size (Blobbing)**

   Parameter 13 determined the stripe size of the conversions. This parameter was used if the type of segmentation was set to blobbing. The stripe size was from a minimum of 10 pixels up to a maximum of 110 pixels. Its value determined the additional padding that should be added to form the final stripe size. For example, if the value was 0.5, the final stripe size was 60 pixels according to the calculation $S = 10 + 10x$, where x is the value of parameter 13.

5. **Blob Size (Blobbing)**

   Parameter 14 determined the blob size of the conversions. This parameter was also used only if the type of segmentation was set to blobbing. The blob size parameter determined the smallest size that could be determined as a blob. By setting a threshold for the minimum blob size, it was possible to eliminate other noise. The range of the value was 0 to 1000. If the value of this parameter was 0.5, the blob size was set at 500.

6. **Pixel Height (Pixelation)**

   Parameter 15 determined the pixel height for pixelation segmentation. For example, if the parameter determined that the pixelation should use 25 pixels as the height, the pixelation would group all 25 pixels in the same column as a single pixel.

7. **Pixel Width (Pixelation)**

   Parameter 16 determined the pixel width for pixelation segmentation. Combined with parameter 15, this parameter determined the size of the pixels that should be grouped together. Their values ranged from 0 to 50. For example, if parameter 15 and parameter 16 gave the value of 0.5 and 0.8, respectively, the pixel size would be a 25×40 pixel resolution, which means the pixels within the

same 25 pixels in a column and 40 pixels in a row would be grouped together.

8. **Swiping Direction**

   Parameter 17 determined the swiping direction of the conversion. For value < 0.5, the conversions would follow the top-down swiping direction. For value $\geqslant 0.5$, the swiping direction would be left-right.

9. **Input Image Resolution**

   Parameter 18 determined the resizing of the input image before it entered the conversion process. Please see Table 5.5 for all the resolutions together with their parameter values.

Table 5.5: List of video graphics array and its parameter value

No	Name	Width (px)	Height (px)	Parameter Value
1	QQVGA	160	120	$0.0 \leqslant x < 0.1$
2	HQVGA	240	160	$0.1 \leqslant x < 0.2$
3	QVGA	320	240	$0.2 \leqslant x < 0.3$
4	WQVGA	400	240	$0.3 \leqslant x < 0.4$
5	HVGA	480	320	$0.4 \leqslant x < 0.5$
6	VGA or SD	640	480	$0.5 \leqslant x < 0.6$
7	WVGA	768	480	$0.6 \leqslant x < 0.7$
8	FWVGA	854	480	$0.7 \leqslant x < 0.8$
9	SVGA	800	600	$0.8 \leqslant x < 0.9$
10	DVGA	960	640	$0.9 \leqslant x < 1.0$

Table 5.5 shows the 10 video graphics array (VGA) modes used for the optimization.

5.3.2 Discussion

Table 5.6: Optimization based on information preservation (at information preservation of 56.4873%)

No	Parameter	Result
1	White	BeeThree
2	Grey	BeeThree
3	Black	Piano
4	Red	Brass
5	Orange	Flute
6	Yellow	Moog
7	Blue	Saxophone
8	Green	Wurley
9	Indigo	Clarinet
10	Violet	Percussion Flute
11	Volume	5
12	Image Segmentation	Pixelation
13	Stripe Size	Not Applicable
14	Blob Size	Not Applicable
15	Pixel Height	8 pixel
16	Pixel Width	8 pixel
17	Swiping	Top Down
18	Image resolution	VGA

Table 5.7: Optimization based on interpretability (at correlation of 0.5728)

No	Parameter	Result
1	White	Bowed
2	Grey	BeeThree
3	Black	Brass
4	Red	Wurley
5	Orange	Flute
6	Yellow	Moog
7	Blue	Blowhole
8	Green	Flute
9	Indigo	Clarinet
10	Violet	Saxophone
11	Volume	8
12	Image Segmentation	Blobbing
13	Stripe Size	10 pixel
14	Blob Size	100
15	Pixel Height	Not Applicable
16	Pixel Width	Not Applicable
17	Swiping	Top Down
18	Image resolution	QVGA

Tables 5.6 and 5.7 show the final results of the optimization process. Presented in Table 5.6 is the best feature mapping candidate that was optimized for the measurement of information preservation. After 34 generations of evolution (in about 4 h), the candidate was obtained with a score of 56.4873% for average information preservation. The score was better than that of Prototype 3, the best prototype in the research studies in terms of information preservation, by a margin of 1.8277%. Table 5.6 shows the best feature mapping candidate that was optimized for the mea-

surement of interpretability. After 22 generations of evolution (in approximately 5 h), the candidate was obtained with a score of 0.5728 in the interpretability correlation. The score was a slight increase of 0.0519 over that of Prototype 1, the best candidate in the research studies in terms of interpretability.

Both the candidates presented here scored a marginal increase as compared to the best prototypes. Although the results are not spectacular, they prove that the optimization algorithm is able to generate feature mapping that can achieve a conversion that is equal to, if not better than, the best prototype that is engineered manually. After examining closely the parameters of the candidates, some interesting features were found. They were both reassuring and surprising. First and foremost, the results confirmed that blobbing performed better in terms of interpretability, whereas pixelation was better in terms of preserving information. This is shown by the fact that the best candidate optimized for information preservation bore a close resemblance to Prototype 3 and the best candidate optimized for interpretability functioned like Prototype 1.

Before the process, it was assumed that the candidate for information preservation would have the biggest input resolution available, which was that of DVGA (960×480 pixel resolution), because with a larger input image, the feature mapping is able to contain more information. It was interesting to observe that VGA was selected as the best resolution for information preservation. From this, it can be hypothesized that the soundscape generated from a large input image tends to have a larger difference in terms of information content. However, if the input image is too small, the information content will also be smaller. Therefore, the optimum the size of the input image hovers around a 640×480 pixel resolution to preserve the largest amount of information. Other than that, interpretability selected QVGA (320×240 pixel resolution) as the best resolution, probably because the minimum setting of the stripe size was 10 pixels. At QVGA, the system was able to capture most blobs while maintaining the interpretability of the soundscape.

An additional interesting feature was the maximum volume presented in both

candidates. The maximum volume for the candidate of information preservation was 5, whereas interpretability had a higher maximum volume of 8. This is interesting because information preservation tends to encode more information in the soundscape, which may frequently cause more noise. By capping the volume at 5, the optimization attempted to reduce the noise to balance the information contained inside. The choice of instruments mapped to the colours surprised us. Table 5.6 shows that BeeThree was chosen to represent both grey and white colours for the optimization of information preservation. This choice is aligned with the previous statement that the optimization attempted to balance noise and content. It may have considered that white and grey fill a similar role and therefore combined them as a single colour.

The biggest weakness of the optimization process was the listening experience of the soundscapes generated by the candidates: the listening experience of the soundscapes was not enjoyable. As compared to the current prototype, the soundscapes generated by the candidates were not as natural and pleasant to listen to. Thus, although the soundscapes were comparable with the soundscapes produced by the research prototype, they were lacking in terms of listening experience. This was probably due to the choices of musical instruments provided for the optimization. It was intended to provide as many timbres as were available so that there would be many variations that the algorithm could change. In the future, this situation can be improved by carefully curating the selection of timbres for the optimization or even better by replacing the current audio synthesizer with a more robust and commercial quality audio synthesizer that is able to produce audio that is much more natural.

In addition, a few timbres in both the timbre set selections sounded similar. Some of the timbre choices were the same for different colours. For example, the candidate for interpretability had the flute to represent both the green and orange colour. This could potentially cause cacophony in the soundscape. Previously, it was solved the issue by introducing some optimization steps by calculating the ISD

between two timbres in order to obtain a set of timbres that were as distinctive as possible. Hence, it is possible that the fitness function of the optimization can be improved to include the computation of the ISD between timbres so that the sound of the final timbre set is distinctive.

Chapter 6

Discussion

6.1 Overview

Chapter 1.3.1 introduces four main aims that the Luminophonics project is attempting to accomplish. The proposed goals were selected in order to increase the performance of VASS technology as well as to improve the practicality of such systems for the intended users, mainly people with visual impairment. It was hopeful that this research findings will facilitate the production of better VASS systems and allow more people to adopt this technology for the benefit of their livelihood, in particular for the purpose of rehabilitation.

The proposed goals were to generate better soundscapes that are more natural and more interpretable, increase the amount of information retention during the conversions, improve the interpretability of the soundscape, and finally develop a better evaluation method to measure the performance of VASS systems. In order to achieve these goals, several approaches are introduced that are relatively new to the field, including a top-down research approach through prototyping, the association of colours with musical instruments, and methods for optimizing the system. Although the research methodology may be unconventional, the findings from this studies meet the main goals of this research. In this chapter, the results of the work thus far are

consolidated and their contributions to the performance improvement of visual-to-auditory conversions.

6.2 Better Soundscape

It is unanimously agreed that the greatest effect of research in terms of increasing the overall performance of a VASS system is due to the continuous effort invested in the improvement of soundscapes. This research was also aimed to improve the overall quality of soundscapes produced by a VASS system. There are many factors that can be used to improve a soundscape, but the scope of this research was narrowed and focused on making the soundscape sound more natural, as well as on increasing its interpretability.

The encoding of visual information into soundscapes through audio frequency modulation has been the default approach for building a VASS system. The main problem raised by this implementation is that the soundscape can be unpleasant to listen to and cause fatigue after it has been used for a long time. The approach for solving this problem is to use the timbres of musical instruments together with a sound synthesizer to replace the frequency modulation method. Based on this core idea, all the conversion algorithms developed for the four prototypes were built to take advantage of the timbres of a set of musical instruments.

The main advantage of generating soundscapes using a sound synthesizer and the timbres of musical instruments as the core component is their sound quality. They sound more natural, because they are based on a type of sound that naturally exists in the environment. This approach takes cues from the performance of music by an orchestra, where combinations of instruments from different families, such as string, brass, and woodwind instruments, are played in an ensemble to perform a musical piece. The features of the visual data are mapped to the different characteristics of the musical instruments. In this case, different colours are represented by different timbres, as if a musical piece is composed of the sounds of colours. Table 3.2

shows the conversions for all the prototypes that map the 10 colours to 10 different instruments. For example, the colour red is represented by a saxophone and the colour violet by the ocarina. The decision to model the colours based on different instruments has its advantages. The main advantage is that it allows us to encode colour information in the soundscape. Since the introduction of including colour information in the visual-to-auditory conversion, the representation of colour by different timbres has been the most frequently used approach. Both See ColOr (Bologna, Deville, Pun, and Vinckenbosch, 2007) and EyeMusic (Hanassy et al., 2013) have demonstrated the effectiveness of this approach.

6.2.1 Timbre Set

The implementation was not, however, unimpeded. A major hurdle when using a set of different musical instruments lies in the control of its complexity in terms of the expected outcome of the soundscape. One of the problems was that it was difficult to select the appropriate combination of instruments to sonify the visual information. First, a set of 10 musical instruments available in a toolbox was randomy selected. However, initial tests indicated that the timbre set was not ideal for the purpose of sonification. Some of the timbre sound signatures were too similar, resulting in sound cacophony when similar timbres were played at the same time. To overcome the problem, a systematic approach was created for selecting an optimal timbre set in which there was minimum similarity between the timbres. The approach constitutes a process for optimizing the timbre set selection that searches and replaces very similar timbres (as described in Subsection 3.2.3). The final result (shown in Table 3.1) is a set of 10 timbres that provide the maximum disparity, which is used in all of the prototypes.

Our timbre set is far from perfect. More work can be done in the search for the optimal timbre set and to improve the optimization process. First, the total number of timbres should be increased from the current total of only 20 timbres in the pool

to encompass more musical instruments. An increase in the total quantity of timbres would increase the size of the search space, which would in turn result in a better set of timbres. It is apparent that the current timbre set suffers an inherent weakness caused by the limited number of timbres in the pool to begin with. In Figure 3.6, we can see that there are three pairs of timbres (violin and clavichord, guitar and flute, and saxophone and organ), which barely meet the minimum similarity requirement. However, a greater total number of musical instruments would allow more timbre choices for replacing the similar timbres.

There is also room for improvement in the cost function of the optimization process. The current implementation focuses on measuring the sound signature by means of the MFCC of audio samples. There are better audio similarity measures that are more robust and that maybe of better suit. Additional sound properties of musical instruments, such as perception, rhythm, and tone, should also be taken into account explicitly. For example, more recently created audio similarity measure algorithms, such as by OFAI, would be a good replacement for the current algorithm. The music similarity by Austrian Research Institute for Artificial Intelligence (OFAI) is able to compute the acoustic distance based on aspects such as timbral and rhythmic qualities, and thus, it is an ideal candidate for measuring timbre similarity (Pohle et al., 2009; Seyerlehner, Widmer, and Pohle, 2010). Further, an additional major problem observed in the current implementation is the audio samples. The audio samples used for the similarity measurements are all based on a fixed set of properties for the sake of simplicity. The audio sample was synthesized using the musical instruments model with the highest volume and La or A as the pitch. Because of this, the similarity measure algorithm did not consider every possible variation of the timbres. In certain circumstances, timbres may sound different in terms of pitch or volume. Therefore, timbres that are dissimilar in the same settings may collide because of a different pitch or volume. This weakness was not observed during the development of the selection process, but was discovered later during the development of the prototypes. Hence, it is recommended

that the selection process be improved such that every possible audio variation is included in the similarity measure. For instance, because the prototypes use the Dorian scale (see Table 3.3), at least eight different frequencies should be covered in the measurements. However, with the addition of eight frequencies for each musical instrument, the number of audio similarity measurements will also be multiplied. To accommodate the large number of combinations, the process should be handled by an automated (or semi-automated) optimization process.

6.2.2 Musical Instrument-based Soundscape

We initially examined musical instruments as the replacement for the current sound synthesizer for VASS systems for the sole purpose of generating a more soothing and pleasing soundscape. However, the attempt led to more discoveries. As compared to the synthesization of soundscapes using frequencies, the generation of musical instrument-based soundscapes is much more complex. However, the benefits it introduces outweigh the difficulties involved in its implementation. The main advantage of using musical instruments is evidently the more enjoyable listening experience it provides to the user. All the users who used the prototypes uniformly agreed that they preferred versions that use the timbres of musical instruments to those that use frequency modulation. Furthermore, the use of musical instruments allows more types of sound to be played concurrently. This provides more space in which the algorithm can accommodate more information in the soundscape, which can never be achieved using a frequency modulation method alone. This advantage in the prototype design was exploited to incorporate colour in the soundscape, which was not used in the first generation of VASS systems and is unprecedented.

It is important that a change in the means by which the soundscape is synthesized does not compromise the quality of the soundscape. The results of the experiments and the feedback received from the users show that the soundscapes generated by the prototypes perform at the minimum as well as, if not better than,

the traditional frequency-based soundscapes. The prototypes are able to preserve the quality of a good soundscape while the soundscape is also aesthetically pleasing to listen to. The results of both this major experiments (see Chapter 4) show that the prototypes perform better for every aspect, such as object recognition and location determination, than vOICe. Although the results may be premature, vOICe as the only frequency-based VASS system with which was being compared, this does not change the perspective that timbre-based VASS systems constitute a stronger alternative than frequency-based VASS systems.

Undoubtedly, the implementation of musical instruments to sonify the soundscape is becoming widely used among researchers of VASS systems. In newer systems, such as See ColOr and EyeMusic, this approach was also applied. The findings also confirmed that the application of musical instruments is the appropriate means of soundscape sonification. In the foreseeable future, the production of VASS systems for the public will be based on a similar approach, unless a better alternative is found. This is because normally humans prefer a more natural sound to an artificial sound when using the system in everyday life. The initial motivation when approaching musical instrument timbres was to emulate a musical orchestra as closely as possible. However, there is still a very wide gap that must be filled before a system is able to produce soundscape that is enjoyable and yet able to carry the relevant information. With newer technologies, such a user experience will not be implausible. However, effort should now be focused on the building of functionalities based on the proposed method that are beneficial to the visually impaired population.

6.3 Improving Information Retention

In a sense, building a visual-to-auditory conversions device resembles building a communication channel, where the input information is encoded and transferred through a certain medium and decoded by the receiver. Exactly as the conversions

implemented in a communication channel, cross-modality conversions also suffer from information loss during the transmission process. To exacerbate the situation, cross-modality from the visual to the auditory domain suffers from greater loss, because the bandwidth of the target domain (auditory) is much lower than that of the source domain (visual). Moreover, certain visual signals cannot be represented in auditory form unless they are manipulated. Solving the problem of information loss is another important step towards improving the performance of a VASS system. A good VASS system has to be able to deliver the information required by the user. If the system loses most of the information during the conversion processes, it is not beneficial to the user, because the soundscape does not contain sufficient information to enable the him/her to make sense of the surroundings.

Throughout the research, two different areas were being examined in which improvement can be made in order to build a better visual-to-auditory conversion that can more effectively control information loss, as well as preserve the information most relevant to the user. First, visual information from various digital images was examined so that the effort on improving the information retention in the conversion algorithms could be focused by selecting the appropriate features according to their relevancy to the intended users. Because of the limited channel capacity of soundscape signals, the preferred means of retaining the greatest amount of visual information helpful to the user is to convert only the most relevant features rather than packing every bit of information into a soundscape. Thus, an optimization process is also proposed to find the optimal feature set that minimizes information loss based on the cost function was formulated. Second, as well as focusing on feature selection, the incorporation of more visual information dimensions in the conversions was attempted. The additional information dimensions, colour and depth, were explored in the prototyping phase. It is anticipated that the use of more information will enable us to build systems that accommodate better the needs of the user.

However, it should be noted that a good VASS system is one that is able to deliver the information required by the user without the overall performance being

sacrificed. Therefore, in an attempt to improve the information retention, it is essential to balance the information loss for the sake of the interpretability of the soundscape.

6.3.1 Feature Mapping Optimization

Feature selection has always been the core problem for cross-modality conversion. It can be a complicated process and frequently causes a ripple effect that indirectly affects the overall performance of the system. From the results of the experiments, I learned that feature selection should be aligned with the objective of the task that the user is attempting to accomplish. In other words, in order to achieve an optimum user experience, the information retained has to support the user's objective.

The results of Experiment 2 (see Subsection 4.3.3) show that compromises are unavoidable when attempting to produce a system that maximizes both the interpretability and information content of the soundscape simultaneously. It is foreseeable that multi-objective optimization will give a solution at the Pareto front that may perform badly as compared to a solution of single objective optimization. Therefore, for situations that require detailed visual content for object recognition, it is preferable that the system emphasize stretching the soundscape to encode more visual information. The observation of all these research prototypes showed that an important feature that differentiates the prototypes in terms of the ability to preserve more information lies in the segmentation method. Prototypes 3 and 4 used pixelation as the segmentation technique, which produced the soundscapes that were able to preserve the most information. The results of Experiment 2 and the information preservation measurement were in agreement. They both indicated the soundscapes produced by Prototypes 3 and 4 were superior in terms of information content. Because of the pixelation, it was possible to capture more detailed information, such as texture, in the soundscape that could not be captured by using the blobbing method proposed. These minor details made it easier for the users to

recognize objects.

As described in Section 5.3, instead of optimizing the feature mapping based on two objectives together, it was decided that two sets of feature mappings were produced and optimized separately for information preservation and interpretability. Although the feature mappings that were obtained from the automatic optimization process achieved better results in terms of information preservation, the overall performance was lacking as compared to that of the prototype, the feature mapping of which was closest to the optimized feature mapping (Prototype 3). This was probably because, by using hand-engineered feature mapping, multiple minor tunings in different areas could be performed, which resulted in better overall results, in particular in terms of the listening experience. Manual tuning allows us (possibly unconsciously) to incorporate other subjective terms in the optimization. Therefore, in order to improve the optimization process, the fitness function should be examined carefully. First, and most importantly, the measurement of information preservation needs to be improved. The current measurement, which uses entropy, is excessively simplistic. The measurement does not take into account the information content for different modalities, but rather assumes that the levels of complexity of the input and the output content are similar. Further, the fitness function of the evolutionary algorithm must encompass additional different criteria, such as the distinctiveness of the timbres in the timbre set and the duration of the soundscape synthesis. In summary, an optimized feature mapping could be created if the fitness of the mapping was measured accurately without losing other important criteria that contribute to the performance of the system.

6.3.2 Additional Information Dimension

The existing VASS system focuses mainly on encoding texture information from images in the soundscape. The texture information is usually encoded by mapping the pixel intensity of each pixel of a greyscale image that has been converted from

a colour image to the corresponding sound property, such as frequency. Although the implementation of such a process is simple and straightforward, some other important visual information is discarded in the conversion (in this case the colour properties). By down-sampling to a greyscale image before the conversion, the system loses colour information that may be important for many tasks, such as object recognition and navigation, in the process.

From this, it was found out that one of the larger factors that affects the amount of information in the soundscape is the visual information that is converted. By limiting the property of the visual information input in the conversion mapping, the amount of information contained in the soundscape may be significantly reduced. Thus, an additional means of increasing the information retention of the conversion algorithm is to expand the richness of the content in soundscapes. This can be achieved by converting more visual properties into the soundscapes. Building on this concept, two means of preserving more information in soundscapes were tested. The first approach was to retain the colour information by mapping the colours to the timbres of musical instruments. Second, by utilizing a depth sensor, the depth information was attempted to be included explicitly in the soundscape as an additional visual property, as demonstrated in Prototype 4 in Chapter 3.7.

Colour

Colour information is one of the default features that the entire range of the research prototypes makes available. In this implementation, which was previously simplified, the colours are sonified based on the 10 different colour types (red, orange, yellow, green, blue, indigo, violet, white, grey, and black). The conversion process then maps the colours to the corresponding musical instruments following in Table 3.2. When they listen to the soundscape, users can recognize the colours by differentiating the timbres representing them.

To reduce the complexity of the cross-modality conversions and lessen the effect of cacophony in the soundscape, it was decided that colour should be down-sampled

to 10 different types. Although the implementation does not cover the entire range of colour depth, the conversion maintains a crucial piece of information, the representation of colour. Thus, users are able to differentiate colours by means of the soundscape, which they cannot do when using the traditional VASS system, which converts only texture information. When the method was used to measure the information preservation of a soundscape, it showed that all the prototypes demonstrated a respectable improvement in terms of limiting the information loss in the conversion after including colour information in the conversion process. This is most evident in a comparison of Prototype 3 and the external system most similar to it, vOICe. As shown in Table 5.2, Prototype 3 scored on average 12.209% better in terms of information preservation. In addition to improving the information preservation, the inclusion of colour information in the conversion provides an avenue for users to identify the colour of an object by listening to the soundscape. This could never be achieved if the conversion relied solely on greyscale images as the input, as exhibited by most traditional VASS systems.

However, the prototype implementation is not without flaws. Because the colour mapping is applied for only 10 different colours, the soundscape is not able to fully describe the entire range of colour depth. Moreover, information about colour shades is lost as a result of the colour reduction performed before the conversion. To mitigate this problem, Prototypes 1 and 2 utilize sound pitch to represent the different levels of colour brightness. Despite the efforts to expand the colour representation by sonifying their shades, the representation is sufficient to cover only a small part of the colour depth. For instance, the users of the prototype can approximately identify crimson as a lighter shade of red than mahogany, but not the precise type of colour shade, if the two shades appear side by side. To exacerbate the situation, the shade-to-pitch implementation can sometimes be counter-intuitive when the pitches of different timbres collide at the same time causing more cacophony in the soundscapes and thus making their interpretation more difficult. Therefore, in Prototypes 3 and 4 the mapping of shade to pitch is replaced in favour of mapping

location information. Prototypes 3 and 4 translate the vertical location value of a pixel into a sound pitch. More precisely, the higher the pitch, the higher is the pixel located, and the lower the pitch, the lower is the pixel located.

This shows that conversions that include colour information are able to produce soundscapes that contain more information than those that do not. This studies also showed that colour information can be sonified in a soundscape easily without losing any major disadvantages as compared to a soundscape with only texture information. Moreover, because of the importance of colour information for performing everyday tasks, it is undoubtable that colour will be one of the basic requirements of most future visual-to-audio conversion systems. However, other options should be explored beyond the scope of representing colour types with musical instruments in order to cover a wider range of colours. Alternatives such as an explicit description of the colour type by a human voice can be considered.

Depth

Prototype 4 is the only prototype that incorporates depth information into the visual-to-auditory conversion. The details of the implementation were described in the section headed 'Depth' in Chapter 3.7.5. Briefly, four levels of depth are provided in the prototype, which the user can select during use. When one of the options is selected, the prototype generates a soundscape that is narrowed according to the depth range specified. The usage of the toggles provides a means by which the soundscape can be modified to communicate the depth information such that the user does not need to relearn the interpretation. Most importantly, explicit mapping of depth information to an additional sound property may cause the already crowded soundscape to confuse the user.

However, the results for the depth implementation, as demonstrated in Prototype 4, were unsatisfactory. This conclusion is based on the observations of the participants and their feedback in Experiment 2 (Chapter 4.3). Although most participants agreed that the depth switch provides a means of receiving depth infor-

mation in the soundscape, the participants were observed to stop toggling the depth switch after they were accustomed to listening to the soundscape that sonifies the entire depth. The feedback indicated that almost all the participants agreed that the depth switch was unnecessary. From a summary of the data gathered from the observations, user feedback, and post-experiment interviews, it can conclude that the main reason for the participants' reluctance to use the depth switch was the implementation of the system itself. I strongly believe that depth information is important for VASS systems, because the users navigated the experimental course by interpreting the depth information from the subtle depth cues received from the soundscape.

There are two possible reasons why my implementation did not fully demonstrate its capability to supply depth information. The user may find that the addition of switches on top of the conversion process excessively complicates the use of the prototype. As the users focus on interpreting the soundscape, switching the depth level can disrupt the interpretation process. Thus, switching back and forth from one depth level to another degrades the user experience. Because the users gradually are able to comprehend the depth of the surrounding from the depth cues presented in a soundscape (full depth), the switch is of minimal benefit to them. Having realised that the disruptive nature of the switches negatively affects the user experience and that it provides a minimal benefit, it was not surprising that the users tended to avoid toggling the depth switches and preferred to focus on interpreting the soundscape.

The second reason may be that the TOF camera used in Prototype 4 is unsuitable for the purpose. In Prototype 4, a DepthSense DS311 was used as the depth sensor to provide the depth map of the surroundings. A TOF camera was chosen because of its low cost as compared to other more complex options, and specifically because this type of camera is the fastest depth sensor presently available. The main limitation that negatively affects the VASS system is that the TOF camera is very susceptible to lighting conditions. As a TOF camera relies on infra-red rays to measure depth, it cannot perform effectively in environments that are extremely brightly lit, because

bright light is frequently accompanied by a high infra-red ray that distorts the infra-red reception of the TOF camera, resulting in an unsuccessful overall depth map being generated by the camera. This is why the TOF camera frequently fails in outdoor scenes, in particular during the daytime. Moreover, the depth map resolution provided by the current TOF camera is usually very low. In the case of Prototype 4, it relied on the QQVGA resolution (160×120) supplied by DepthSense DS311. Because the original source image has a very low resolution, first it does not have sufficient information, and second, the noise to signal ratio of low resolution images is higher. During the segmentation and subsequent conversion process, noise is enlarged, and ultimately degrades the interpretability of the soundscape. Hence, it very significantly reduces the amount of information that can be encoded in the soundscape and this in turn degrades the overall performance of the soundscape as a result of limited information and more noise in the conversions.

Although Prototype 4 was not able to fully provide the depth information as it was designed to do, I still believe depth information is an additional important element for VASS systems that should be explored. This is because humans naturally rely heavily on depth perception for everyday spatial tasks, such as navigation and understanding scenes. Learning from the mistakes of this research, I suggest that an entirely new approach that focuses on a fluid user experience should be designed for implementing depth information. An automatic approach, similar to that used in See ColOr, where depth information is used to identify areas with high visual saliency for sonification (Bologna, Deville, Pun, and Vinckenbosch, 2007), may be the best method currently available. Furthermore, types of depth sensor other than a TOF camera should be explored. For example, it is recommended that sensors that use stereo triangulation, such as stereoscopic cameras, or sensors that rely on a coded aperture for depth sensing should be explored. These two types of depth sensor do not rely on structured light as the means of depth sensing and therefore they are less susceptible to light noise and thus more suitable for everyday scenarios, including outdoor operation. Moreover, these cameras can supply a larger image resolution

when they are provided with sufficient processing power.

6.4 Increasing Soundscape Interpretability

Its interpretability is by far the most important aspect that defines a good soundscape. An interpretable yet simple soundscape is probably more useful than a complicated soundscape that contains a large amount of information. For instance, a soundscape that is rich in information but difficult to interpret is of minimal benefit to users. This is because, if the soundscape is difficult to interpret, users will experience difficulty understanding its content let alone using the information to assist them in their everyday tasks. Hence, instead of focusing on introducing new content into a soundscape, the objective was to balance a useful amount of information with the level of interpretability of the soundscape.

During the prototyping phase, three major changes were introduced focusing on improving the interpretability of the soundscape. These changes were implemented in the prototypes in different phases of their development. They are feature extraction through image processing, cacophony reduction, and optimization of feature mapping.

6.4.1 Feature Extraction through Image Processing

The first approach that was used in the project to help the user interpret the soundscape better was to explore the area of applying feature extraction to the visual input prior to the visual-to-auditory conversions. The main purpose of deploying feature extraction in a VASS system is to be able delegate some interpretation work to a computer. The computer pre-processes the image, allowing a simple yet useful soundscape to be created for the user. Traditionally, most VASS systems directly map the intensity value of each image pixel to the corresponding sound properties to create a soundscape. The technique of direct mapping limits the visual information that is converted to the information at the lower level, neglecting the importance of

the information at the higher level. As a result, these systems frequently require the human user to perform more interpretation in order to fully understand the content of the soundscape. The approach is to exploit the processing power of a computer by using it to extract more features from an image using image processing algorithms. Thus, the prototype is able to shift the workload from the user to the computer, thereby reducing his/her effort required to interpret the lower level information but focusing on the bigger picture of the entire content. VASS systems that utilize feature extraction in their conversions belong to the semi-automated generation, as described in Subsection 2.1.2.

Feature extraction was used primarily in two of the earlier prototypes, Prototypes 1 and 2. These two prototypes apply image processing algorithms extensively throughout the processing, but primarily before the start of the conversion. A contour-based image segmentation technique is used prior to the conversion phase to extract the information about the blobs in the image. Through the application of the algorithm, the conversion process is able to use information, including the size, location, colour type, and shape of the blobs. This is a major step forwards as compared to the VASS systems of the manual generation, which convert only pixel-based visual information. Instead of encoding the raw pixel intensities, Prototypes 1 and 2 generate soundscapes based on the information of the blobs in the image. As a result, the users can reconstruct mental images based on blobs instead of raw pixels. Many benefits are reaped through using this approach. The main improvement can be seen in the results of Experiment 1, where Prototype 1 performed well even for tasks involving object detection. Moreover, the soundscape used in Experiment 1 was sonified at a rate of 2 s per frame, which is faster than the sonification rate of vOICe, which is more than 4 s per frame. This proved that using contour-based image segmentation, the prototype can achieve results that are similar to if not better than those of the traditional VASS systems belonging to the manual generation.

The key point of this approach is that users are able to interpret the scene at a much faster rate, because the soundscape helps them bypass the process in which

they need to reconstruct mental image from the pixel level. Instead, when using Prototypes 1 and 2, they can focus on the information at a higher level, such as blobs, shapes, and colour. Thus, the prototypes are able to produce a soundscape that is more easily interpreted, because the users are able to interpret the soundscape and understand the content more quickly. In addition, the approach also improves the learnability of the system. The users of the prototpyes were able to start an assigned task after a maximum of one training session. This can be seen in both Experiments 1 and 2.

The results of Experiment 2 enabled us to identify a major weakness of this approach. Even before the experiments, it was suspected that the possibility that the approach might in certain cases lead to oversimplification of the content of the soundscape. In general, an oversimplified soundscape may lead to a lower performance for specific tasks, the execution of which requires detailed information. Thus, Prototype 3 was developed based on pixel features rather than the blob features on which the development of Prototypes 1 and 2 was based. By comparing the two contrasting approaches, their strengths and weaknesses were discovered. In Experiment 2, the prototypes were tested in a real scenario composed of two major tasks, navigation and object recognition. The results of Experiment 2 further confirmed that the application of blobbing as the image segmentation technique (applied in Prototypes 1 and 2) helped the users navigate but did not effectively aid their object recognition. The main reason for these interesting results is that tasks such as navigation are more easily accomplished using systems that provide quickly delivered and concise information. The users are able to depend on this quickly delivered information to make instant judgements and choose the correct path. Although the information may not be detailed and precise, users are able to correct their course quickly by using the rapid and continuous stream of information supplied by the system. However, tasks that require detailed information, such as object recognition, are not very easily performed using systems that are tuned for speed. Users need considerably more information to correctly perform this type of task. Therefore,

Prototype 3, which is built based on pixels rather than blobs, performed better in object recognition because it provides more precise and detailed information. Prototype 3 is able to describe the scene better because the soundscape is considerably richer in details, providing more information, such as shape and texture.

The first purpose of implementing an image processing algorithm for feature extraction was to harness the processing power of a computer to reduce the time that a user takes to interpret a soundscape when low-level pixel information is used. It was anticipated that with the help of a computer the system can achieve a higher level of interpretability. However, the results of the experiments and the tests of the prototypes led us to an interesting finding. The interpretability of a soundscape depends heavily on the user's task. Tasks that require a faster response will always be accomplished more easily using short and fast soundscapes with clear and concise information, which in this case corresponds to Prototypes 1 and 2, which implement the blobbing technique as the primary image segmentation. However, a VASS system that is built based on image pixels has the advantage that more detailed information is encoded inside the soundscape, which favours tasks involving object recognition. Although there may come a time when VASS systems can be sufficiently general to cover all human visual functions, currently it is still best to build a VASS system that targets specific tasks to achieve maximum interpretability.

6.4.2 Reducing Cacophony

Sound cacophony is a major hurdle that is slowing down advances in VASS technology. It directly affects the interpretability of a soundscape, and each increase in the effect of cacophony makes the interpretation of the soundscape more difficult. Many factors may lead to a build-up of the effect of cacophony in the soundscape. Among them, the number of features encoded is considered the primary factor. The more visual information translated into auditory form, the greater is the effect of cacophony in the soundscape. Hence, it prevents us from packing more information

into a soundscape, because of the concern that this may cause more cacophony, which will then lower the interpretability of the soundscape and the overall performance of the system. Therefore, it should be noted that a balance between the amount of information and interpretability must be achieved in order to create a VASS system having a high performance.

The same problem was encountered during the development of the prototypes. The usage of musical instruments as a replacement for frequency-based soundscape synthesis was introduced to offer more variety and richness in the soundscape and to allow more information to be encoded in it. However, also this approach cannot avoid the effect of cacophony. If it is not appropriately handled, the chance of sound cacophony occurring is frequently higher than in the traditional approach. This is especially true when musical instruments with similar sound signatures are used in the sound synthesis. In order to mitigate the problem, a solution was proposed. An automated method was created to select the optimal set of timbres out of the pool of available musical instruments. The solution selects 10 timbres by computing the ISD for each timbre and filters out any timbre having characteristics and sound signatures that are too similar to those of another timbre. This creates a better overall timbre set with minimal sound collisions. In fact, the improvements in the quality of the soundscape were noticeable in terms of overall interpretability as compared with the soundscape that was produced when a randomized timbre set was used. Positive feedback was received from the users, the majority of whom indicated that the improved timbre set produced a clearer and more distinctive soundscape. The effect of cacophony was more obvious in the soundscape produced using the improved timbre set than in that produced using the initial timbre set, because in the latter some of the constituent timbres sounded too similar and were more difficult to distinguish.

Although the proposed method yielded some improvements in the soundscape, it did not completely eliminate the effect of cacophony. The interpretation of the soundscapes can become more difficult as the number of different musical instru-

ments playing at the same time increases. One of the factors that contribute to the problem may be the fact that the selection process did not cover the sound similarities of timbres at different pitches and volumes. Therefore, the measurement of sound similarity for each timbre should be expanded to cover a wider range of criteria, including different ranges of pitch and multiple levels of sound volume. To complement this, it is recommended that an optimization algorithm, such as an evolutionary algorithm, be incorporated to reduce the search time. This is because the search space will expand exponentially as the number of selection criteria increases.

When implementing a new feature in a system that translates visual information to auditory information, researchers should investigate further and find solutions that mitigate the effect of cacophony in the end result. I learned that the interpretability of a soundscape is strongly related to the sound cacophony. A high level of sound cacophony lowers the interpretability of a soundscape. It was confirmed that, when the proposed solutions to reduce the noise and sound cacophony were not applied, the soundscapes produced by the prototype are less interpretable than those produced by other existing VASS systems.

6.5 Better Evaluation Method for VASS

An additional contribution that Luminophonics has made to the development of VASS systems is a method for evaluating their performance. Previously, a standardized model that was able to measure the performance of VASS fairly across different systems, regardless of the algorithms implemented, did not exist. Historically, the default approach for evaluating a VASS system is through user-based experiments. Although evaluation through user trials is the best means of measuring the full potential of a VASS system, it is difficult to replicate trials across all systems. As a consequence, the results of these experiments are frequently not easy to use for comparing different VASS systems fairly. A method to complement the standard experimental methodology was developed, which consists of evaluating the VASS

quantitatively using the input images and the corresponding soundscapes. Chapter 5.2 describes such method has the potential to become a standardized evaluation framework for VASS systems because of its simplicity and straightforward implementation. Moreover, because the method relies only on input images and their corresponding soundscapes, which are common across all systems, the values in the results can be used to compare different VASS systems. By using such performance measurement tools, system that were previously developed in silo can be compared in a fairer manner.

As mentioned, the automated performance measurement for VASS systems that was introduced is a very good tool as a complement to the existing user-based experiment evaluation method. It does not only provide a fair overview of the performance of a VASS system (in terms of interpretability and information preservation); it has another benefit of which VASS developers can take advantage as part of their system development process. Because of its low implementation cost, it is an ideal tool for testing and for filtering out poor features. It can be quickly set up to evaluate newly implemented features without incurring the considerable cost of conducting an experiment. The results of the measurement can be a good indicator of whether the new features are affecting the overall performance of the system positively or negatively. Using the results, researchers and developers alike are able to make informed decisions at an early stage as to whether to proceed to further development without wasting many resources in the long run. Essentially, the automated measurement is a 'litmus test' for VASS systems in that it provides an early indication as to whether the performance of a system will be good or poor before the entire development has been completed. Hypothetically, if the process of automated performance measurement is incorporated in the development of every VASS system, the technology of visual-to-auditory conversion can advance more quickly. Because it has shortened the time it takes to create a system, which leads to a reduction in the total development cost, it gives more researchers (in particular those having a relatively small budget) the opportunity to be involved in contributing to the

technology of visual-to-auditory conversions.

However, the performance measurement requires considerable improvement before it can realize its full potential as the standardized performance measurements framework. So that the automated performance measurement can be accepted worldwide, it has to provide a more comprehensive evaluation of the system. Currently, the proposed measurements are limited to only two measurement criteria, interpretability and information preservation, which are not sufficient to completely describe the ability of a system. Improvements should be made to enhance the current measurements, and additional measurements need to be included. As discussed in Section 5.2.2, the measure of soundscape interpretability is based on the correlation between the IID and ISD. Because it is expected that the IID and ISD have a linear relationship, a PCC is used to measure the relationship between the two variables in order to describe a system that produces a highly interpretable soundscape. As suggested, other correlation measurements, such as distance correlation or rank correlation, should be explored in case the relationship between the IID and ISD is not linear in nature. Dimensionality reduction techniques can also be considered, as well as the values of the IID and ISD, to measure the interpretability of the soundscape. One suggested alternative is to reduce the dimensionality of both the images and the soundscapes to a common dimension using a technique such as principal component analysis (PCA) or singular value decomposition (SVD). The interpretability of a soundscape can be estimated by calculating the distance between the images and the corresponding soundscapes in the same dimensionality. A system that is highly interpretable should be able to produce soundscapes that are highly correlated with the input images. In other words, after dimensionality reduction, the distance between a visual input and its corresponding soundscape should be minimal. As does that for the measurement of interpretability, the approach for measuring the information preservation shows some weaknesses. The reliance on comparing the entropy of the input image and its corresponding soundscape may be too simplistic. Although it indicates which system preserves the information better

in the conversions, the comparison does not reveal the amount of information lost in the process. This is due to the inequalities of information entropy in different modalities. In order to mitigate the problem, methods for computing information gain, such as Kullback-Leibler divergence, which measures the relative entropy between two probability distributions, can be used to measure the difference in the amount of information before and after conversions.

The automated measurement should also be expanded to include more measurements in order to describe better the performance of a system. A good indicator in addition to interpretability and information preservation is the total time required to complete all the conversions in a system. In certain scenarios, users may be able to benefit from a system that is capable of quickly converting an image into a soundscape. The inclusion of this time in the measurements can help researchers whose objective is to filter out algorithms that do not meet the time requirement. Further, a set of audio analyses of the soundscape can be useful for finding a system that produces quality soundscapes. In general, a good soundscape should be pleasant to listen to and exhibit minimal cacophony. Using audio analysis, unpleasant sound quality can be detected and eliminated, such as high disparities between peak volumes and unnatural sound pitches.

Chapter 7

Future Works

7.1 Overview

In this dissertation, multiple improvements to advance visual-to-auditory cross-modality conversion systems are presented. The proposed improvements are aimed at different areas of the entire process. They include the audio quality of the soundscape, the ability to preserve information before and after conversions, the interpretability of the soundscape, and finally performance measurements for evaluation purposes. According to the results of the experiments and the feedback from all the testers, it can be stated that the proposed changes benefit visual-to-auditory conversion and its application as a whole. Although many possible solutions are presented to overcome the weaknesses of the current VASS systems, the problems are far from solved. Continuous improvements have to be made to the technology before it can be adopted widely by the public.

On a positive note, in the foreseeable future the growth momentum of VASS will be increased in tandem with the advances that have been achieved in numerous fields, such as computer science, neuroscience, and psychology. As discussed in Chapter 2, the rise of VASS was initiated by the advances in electronics, whereby the synthesis of a soundscape by translating visual signals electronically was made possible. As VASS

systems progress from the manual generation to the semi-automated generation and finally to the automated generation, an increasing number of techniques are being incorporated to process the visual information better in order to improve the interpretability of the system. The devices of the first stage, which produce soundscapes by directly converting the visual signal without passing them through additional image processing, are generally regarded as the VASS systems belonging to the manual generation. The second stage of VASS systems, the semi-automated generation, was advanced by the abundance of computing power. Currently, the power of computers can be capitalized to process more visual data, mapping them to generate a richer soundscape. Although electronics has played a crucial role in the application of the technology, the achievements of visual-to-auditory conversion in its current state were not reached without the help of our knowledge of neuroscience. As we continue to deepen our understanding of the human brain, much of the relevant knowledge, in particular that about brain plasticity, can facilitate the research of sensory substitution.

Although the second generation of VASS devices showed many of the capabilities of visual-to-auditory sensory substitution technology and its relevance to humans, in particular the visually impaired, they did not achieve any real positive results outside the experimental environment. The reasons for the poor public adoption of such devices are most probably its weaknesses and the limitation in terms of its robustness. The length of the learning period together with aesthetics and the interpretability of the soundscape are some of the major weaknesses that were identified in the second generation VASS systems. However, as the field of computer science and neuroscience have continued to mature, their advances have influenced the progress of VASS to a very great extent. This is particularly true given that in recent decades our understanding of neural networks has had a great effect on computer science, especially in the field of machine learning. This convergence of machine learning (in terms of neural networks) and neuroscience may be lead to the next breakthrough needed to advance VASS technology to the next stage. Signs of

this are slowly surfacing, as applications of deep neural networks are starting to be incorporated in some SSDs, usually in the form of visual recognition. The situation should improve further as VASS systems evolve into the next, fully automated, generation, driven primarily by machine learning and neuroscience.

Although most of the research efforts are channelled to improve the current generation of VASS, the next logical step for the research of visual-to-auditory conversions is to explore other opportunities to realise the coming generation of VASS systems, a fully automated one. In this chapter, other promising solutions are discussed that have potential to solve the limitations that VASS is currently facing. By redesigning the process of cross-modality conversion and coupling it with state-of-the-art machine learning techniques, it is possible that VASS systems can achieve a greater performance than that of the current generation. With high hope, SSDs, especially VASS systems, will be accepted by a wider public audience in the future and have a more significant effect as devices that provide access to individuals with visual impairment.

7.2 Visual Recognition

The integration of visual recognition in parallel with the soundscape will be the next great advance in VASS technology. Visual recognition allows the system to describe objects in the scene in a language understandable by its users. For instance, while playing the soundscape, the computer can pronounce the word 'tree' in the form of human speech if a tree is detected in the input image. The verbal audio description provided by the computer allows the user to be aware of the existence of a tree in front of him/her. This reduces the time and effort users need to invest in mentally reconstructing the visual information in the soundscape and subsequently interpreting the content.

Such an ability may have sounded implausible in the earlier phase of VASS development, but with the advancements in the field of machine learning, human-like

visual recognition by computers is achievable by using deep learning. A number of such systems currently exist, most of which are aided by convolutional neural networks (CNNs) trained with a large number of images labelled with a description in words. Of these, the most popular system that demonstrates the ability to describe visual data in words is NeuralTalk presented by Karpathy and Li (2015). To represent the image, they use a combination of region convolutional neural networks trained on ImageNet and MSCOCO datasets (Lin et al., 2014) to extract the relevant objects and plug the information into a recurrent neural network (RNN) , called the bidirectional recurrent neural network, to form a string of word representations. The results are exceptional: the model, in particular the most recent version (NeuralTalk2), is able to describe image frames using English sentences in a matter of seconds. Figure 7.1 shows a screenshot from a video of the NeuralTalk2 in real time produced by @kcimc[1] that describes a scene with the caption 'A man is eating a hot dog in a crowd'. The result is not perfect but it correctly describes part of the image. Another impressive image captioning algorithm was demonstrated by the Google Brain team. The model is able to describe images with high accuracy (Vinyals et al., 2017). The idea behind the model is that a transfer learning technique using a deeper CNN (Inception-ResNet v2) is applied on top of an RNN for caption generation. The model is very successful: it achieved first place in the MSCOCO Captioning Challenge 2015, outperforming other strong competitors, e.g., models developed by Microsoft Research and the University of Montreal, and NeuralTalk.

The element of visual recognition built into a VASS combines well with the core idea behind the second generation of VASS systems, which promotes using computational power to process the content prior to the delivery of the soundscape to the user in order to improve its interpretability. While the computing power in second generation VASS systems is used to describe the objects in terms of features, such as shapes and blobs, visual recognition provides a further improvement by

[1]Source: NeuralTalk and Walk (https://vimeo.com/146492001)

Figure 7.1: Screenshot from a video demonstrating NeuralTalk2 by @kcimc

describing the objects in the form of human language. As compared to the second generation, the effort that the user has to invest in soundscape interpretation in order to recognize the objects in the scene when using the third generation is minimal. Hence, this generation is called fully automated, as opposed to the second generation, which is called semi-automated. Ultimately, the obvious benefit of including visual recognition in a VASS system is the increase in the interpretability of the resulting soundscape. The fact that the interpretation of the soundscape is becoming easier creates many more positive additional effects. One direct effect in addition to the ease of interpretation is that soundscapes can now be shorter and faster. A few words formed into a sentence can fully describe the visual content in a shorter time frame. Because the user can more quickly perceive the content of the soundscape, he/she can make faster judgements. It was a major disadvantage of second generation VASS systems that they were unable to sonify the richness of the content in an

acceptable amount of time or to describe the content more quickly without sacrificing information. Hence, visual recognition may be a solution that provides the best of both worlds: it can both fully and quickly describe the content, because the richness of the visual input is translated semantically into human speech.

Because the soundscape is available in the form of human speech and easily interpretable, the learnability of the systems is also increased. Previously, users had to learn to decode the soundscape in order to understand its visual content. The duration of the training required to reach standard proficiency could be from a few weeks to months, depending on the users' experience. However, because the soundscape is translated into a language that users understand, they do not need to learn the feature mapping used in the conversions algorithm. Thus, the training period is reduced significantly by the fact that the VASS system and its users share a common set of communication protocols: the human language. Essentially, visual recognition uses language as the feature mapping/encoding to sonify the visual information for the soundscape. An additional benefit of using human language to describe visual information is that the users are already accustomed to human speech. This means that the systems do not have to invest the effort that is required to produce a natural soundscape, which was a result of the introduction of musical instrument in sound synthesis. If a VASS system can describe the visual information in human language, it can be treated as an agent that is constantly talking to the user describing what it sees.

Although visual recognition appears to be promising, the feature is still in the very early stages of development. More work is required before it can be incorporated into an actual VASS system. The most difficult problem that needs to be solved involves the attention mechanism. Currently, no control function that determines the segment that the neural network will describe is implemented. The caption of an image is generated based on the confidence level of the model trained on the dataset. An ideal visual recognition function has to mimic the saliency of human attention, describing the most important part of the image and filtering out the

other minor details. However, the contribution of deep learning (both modelling and datasets) is helping close the gaps in visual recognition. Recent work, such as Show, Attend, and Tell produced by the MILA Laboratory, shows good progress (Xu et al., 2015). They introduced a novel approach that is able to describe the content of images using an attention-based model that is capable of focusing on highly salient objects. The current state-of-the-art image captioning was developed by the researchers at the IBM Watson Research Center (Rennie et al., 2016). The approach uses a reinforcement learning technique called self-critical sequence training (SCST) and when trained on the MSCOCO dataset the model was able to achieve a score of 114.7 CIDEr in the MSCOCO evaluation. The advances in deep learning suggest that a comprehensive visual recognition function built into a VASS system may be possible in the not too distant future.

Meanwhile, until stable maturity in the field of visual recognition is achieved through image captioning, the best immediate approach is to develop a better image segmentation method, such as the blobbing technique that was proposed in this research. However, the use of neural networks to replace the traditional contouring techniques may allow the performance of the segmentation to be more robust and refined. Currently, there are many popular image segmentation approaches that use deep learning, such as R-CNN presented by Girshick et al. (2014), YOLO presented by Redmon and Farhadi (2016), and fully convolutional networks presented by Long, Shelhamer, and Darrell (2014) and Shelhamer, Long, and Darrell (2017). Using a high quality dataset such as MSCOCO, these segmentation methods are able to locate and outline the objects detected in a short time. The best semantic segmentation technique is Mask R-CNN developed by Facebook AI Research, which extends the idea behind R-CNN (He et al., 2017). By adding an additional branch to predict objects on top of bounding box recognition, the accuracy of Mask R-CNN was improved and it outperformed all existing methods. If the speed of these methods can be improved to be equal that of the current contouring-based segmentation methods (at least 30 fps), they potentially can be used as a drop-in

replacement for the segmentation layer. Using semantic segmentation performed by neural networks, the VASS system will be able to produce soundscapes based on actual objects rather than only on blobs.

7.3 Deep Learning and Audio

The future of visual-to-auditory cross-modality conversions will be strongly influenced by the many breakthroughs derived from machine learning. This includes improvements in the sound quality of the soundscape. Recently, many successes in the area of synthesizing high quality audio by using deep learning models were reported. By applying such methods, neural networks, trained with a high quality dataset, will definitely greatly facilitate the generation of better soundscapes. Hence, it is expected that the soundscape produced by the VASS system of the future will be of higher quality. As compared to the soundscapes from the current generation, the sound will be more natural and closer to real life.

One of the more prominent studies on using deep learning for audio synthesis was conducted by the researchers of the Magenta project at Google Brain. The core motivation of the Magenta project is the wish to develop high quality music and art generation using various machine learning techniques. Although their research does not affect the development of VASS research directly, the spillover effect from their work, especially that on audio synthesis, will benefit those working on creating a better soundscape. Many of their published results showed that, by using some of the state-of-the-art neural networks, such as long short term memory (LSTM) and generative adversarial networks (GAN), they were able to synthesize audio clips that closely resemble music composed by a professional artist.

The most recent results can be seen in NSynth, an audio synthesizer that is capable of generating raw audio samples using neural networks (Ramachandran et al., 2017). The core component of NSynth is WaveNet presented by Oord et al. (2015) of DeepMind, which basically is an autoencoder coupled with a fully

convolutional neural network (CNN) built for the purpose of learning the embeddings of audio samples. Trained on a relatively large set of audio and musical notes, NSynth is able to encode a sound pattern and then resynthesize it based on other parameters into a new variation of the original sound. The technology of NSynth will solve two major existing problems related to VASS systems. First, the learned embeddings from NSynth can be combined and mapped with the features of our visual input to create an automated visual-to-auditory conversion. Second, the technology offers more possibilities for soundscape synthesis, because researchers will not be limited to the musical instrument models that are currently supplied with a sound synthesizer. It is possible that its use will allow VASS to include many different sounds to sonify an image. For example, the sound of a bird chirping could represent a bird in the image and the sound of wind a clear blue sky. Soundscapes will be not only more natural but also more lively, creating a considerably more immersive experience for the listener.

An additional impressive result of the same group is the usage of LSTM to create a polyphonic music model, called Performance RNN (Simon and Oore, 2017). Trained on the Yamaha e-Piano Competition dataset[2], the model is able to modify the timing and dynamics of the music without interrupting the sound. The timing of the audio output has become one of the reasons why the soundscape produced by a VASS system does not resemble the music of an orchestra as closely as expected. Because of excessive audio modifications to match the mapping of the visual features, the soundscapes sound unnatural with many sudden intervals in between the sounds of instruments. It is hopeful that by applying the proposed Performance RNN, the timing in the soundscape can be modified without breaking the flow and dynamic of the music. Moreover, it provides an additional audio property, as well as the usual frequency and volume, that can be mapped to the visual features, that is, the temperature of the audio. Using this property, a VASS system can express the temperature of the visual input in the audio. For example, a blueish scene can be

[2]Source: Yamaha e-Piano Competition dataset (http://www.piano-e-competition.com/)

expressed by a slower tempo in the soundscape, while a soundscape with a faster tempo can describe a reddish image.

There is no doubt that the availability of such models will, with little effort, allow the sound of soundscapes to be considerably improved in the future. It is possible that with a larger dataset and a sophisticated model, a neural network-based audio synthesizer will make considerably more audio properties available that can be changed. With more parameters, the soundscape can be more flexible, providing considerably more avenues for mapping visual information. To conclude, the soundscape in the future will both sound more natural and be considerably richer as a result of the flexibility and capacity provided by a better synthesizer.

7.4 Cross-Modality Mapping

An additional important feature that is expected to be introduced in the next generation of VASS systems is automated visual-to-auditory mapping through optimization. Currently, the building of a visual-to-auditory mapping that is able to translate the visual information into a soundscape that the user can intuitively understand and that is able to assist the user in many tasks presents a major problem. In addition, at the present the process of creating a new mapping is frequently resource intensive, because creating the mapping currently involves manual feature engineering by means of a considerable amount of trial and error. Although the expected performance is achieved, the resulting soundscapes are far from optimal. There is room for the algorithms to be developed further, so that they can produce soundscapes of higher quality. However, because it is a slow and expensive process, we may not be able to finally achieve a good visual-to-auditory mapping for a long time if the path of manual engineering approach is continued. Hence, it is recommended that a more efficient method to build the mapping automatically through the application of an optimization algorithm should be explored. Chapter 5 describes a technique that can be considered a first step towards such an implementation. To

the best of my knowledge, this is the first time an optimization process has been used to explicitly evolve VASS systems. CMA-ES as the evolutionary algorithm was applied to form two sets of mappings optimized for interpretability and information preservation, respectively. Fortunately, many more optimization techniques are currently being developed that are suitable for this purpose. In this section, explore other state-of-the-art methods that have great potential are explored They have the potential to help us create a better visual-to-auditory cross-modality conversion.

Again, deep learning plays a very significant role in advancing VASS and it too will help to create a better translation mapping for visual-to-auditory cross-modality conversions. In this case, unsupervised learning neural networks, such as autoencoders, are useful. Autoencoders are a type of neural network that is comprised of two parts, an encoder and a decoder. The objective of such a network is to apply a backpropagation algorithm to learn the embeddings/representation of the input from the encoder element and attempt to recreate the output such that it is as close as possible to the input from the embeddings. Currently, state-of-the-art autoencoders are implemented in many applications, such as those for pretraining other classifiers, data compression, and information retrieval. Because autoencoders excel in dimensionality reduction, they are very suitable for the purpose of cross-modality translation. Many studies have been reported that involved the usage of autoencoders in cross-modality applications, in particular to retrieve information based on different modalities (Ngiam et al., 2011; Wang et al., 2016; Vukotić, Raymond, and Gravier, 2016). One particularly interesting study is that of Fried and Fiebrink (2013), who attempted to achieve a goal that is very similar to ours. They proposed training a stacked autoencoder to find the mapping between visual input and the corresponding audio output. Using the mapping, they were able to link a piece of music to a slideshow and identify the audio by means of the corresponding gesture image.

Although the study of Fried and Fiebrink (2013) provided some glimpses of the application of an autoencoder to create visual and auditory cross-modality mapping,

their proposed technique is still far from the goal, because options for synthesizing audio directly from the learned embeddings are lacking. While currently, no methods exist that are able to provide these options, two probable solutions using an autoencoder exist that may lead us to tools that can create automated visual-to-auditory features mapping for soundscape synthesis. A direct approach is to build an autoencoder using the image in the encoder to capture the information, whereas the decoder is formed to create a synthesizer that generates the corresponding soundscape. This approach is simple and more easily implemented, but it has a major weakness in that the network may generate a soundscape that is related well to the embeddings but is incomprehensible to the user. An additional indirect approach is to use an autoencoder, such as a stacked autoencoder or variational autoencoder, to learn the embeddings of the image dataset first. Then, another generative model using networks, such as GAN or LSTM, is connected to synthesize the soundscape from these embeddings. The second approach is more robust because it provides the option to guide the sound of the soundscape by giving a higher score to a sound that is deemed suitable for the user. Thus, a more natural sound can be created using musical instruments in the sound synthesizer.

However, it is by no means a trivial task to create an automated cross-modality mapping. The major reason for the difficulty is that the construction of an ideal cost function that measures the content of the input image and the resulting soundscape presents many challenges. The problem was also faced in this research when building the automated optimization feature mapping using CMA-ES. Therefore, the effort should be focused on creating a better cost function that correctly defines the qualities that characterize a good soundscape before applying such an optimization algorithm. An additional approach is to collect a vast number of datasets comprising images and soundscapes to train the algorithm to learn the 'good' feature mapping between the image and soundscape pair. Thus, we may achieve a system that incorporates an automated process for synthesizing soundscapes that is capable of describing much visual information in the form of audio in near future.

Chapter 8

Conclusion

To conclude, our research project achieved a number of improvements in multiple different areas of visual-to-auditory cross-modality conversion. First, the possibility for a VASS system to sonify the soundscape using the timbres from musical instruments through the implementation of a sound synthesizer is being demonstrated. The objective of using musical instruments is to make the soundscapes sound more natural than the traditional soundscape that is based on frequency modulation. Our results show that not only do users prefer to listen to a soundscape constructed from the timbres, but also the performance of the soundscapes is comparable to, if not better than, that of frequency-based soundscapes.

However, the idea of using musical instruments was not conceived solely for the purpose of improving the user's listening experience. It also made available possibilities to encode more information into a soundscape. Thus, the inclusion of colour information in visual-to-auditory conversions was introduced. The mapping of 10 different musical instruments to 10 different colours allows our soundscapes to sonify the colour information of the input image. In order to reduce the effect of cacophony in the soundscape, a timbre set selection process was developed. It is based on comparing the sound signature of a series of two different timbres to obtain a set of 10 different timbres having sounds that can easily be distinguished. Our efforts to increase the amount of information retention achieved by the conversions algorithm

did not include only the introduction of colour information; the implementation of an image processing algorithm that is applied mainly in the image segmentation process was proposed to further increase the feature extraction for the input images. The resulting soundscape is able to sonify additional information, such as blob size, which is mapped to different audio properties, such as volume and pitch. In addition, depth information was attempted in Prototype 4, collected using a TOF depth sensor, as an addition to the 2D visual information supplied by images captured using a normal camera.

In this dissertation, several approaches for improving the user's experience of soundscape interpretation were suggested. As demonstrated in Prototypes 1 and 2, a contour-based image segmentation algorithm is used for the purpose of feature extraction. The main purpose of including an image processing algorithm in the conversions is to allow a computer to handle this part of the processing task so that the user can focus on the interpretation. The results are encouraging because this feature allows users to interpret the soundscape more quickly, making the systems suitable for tasks that require rapid decision making, such as navigation. To complement our effort to increase the interpretability of the soundscape, a measurement was developed based on the correlation between the input and output of the system as a basic tool to measure the performance of a VASS system in terms of interpretability.

Central to our approach are several innovative solutions that were introduced in addition to the successes achieved with our prototypes. The purpose of these contributions was to solve some of the difficult problems that hinder us from achieving our goals. One of the opportunities for improvement that was identified was the lack of an evaluation framework for VASS systems. Without a standardized evaluation method, it is difficult to compare performance across different VASS systems. Two new measurements were developed to measure the performance of a VASS system, interpretability and information preservation. These two measurements form the basis of an evaluation method to rank the performance of a VASS system consider-

ably more quickly than the current user-based experimental methods. It is hopeful that with further development and refinements by various parties worldwide, the measurement methods can become a standardized evaluation platform for VASS. Building on the measurements as the cost function, an optimization method was developed to search for the best visual-to-auditory mapping for the conversions. The proposed method is based on the usage of an evolutionary algorithm such as CMA-ES to produce two sets of feature mapping optimized respectively for two different purposes, namely, interpretability and information preservation. The optimization process reduced the duration of the prototyping and the testing processes used to search for an optimized visual-to-auditory feature mapping.

Overall, the results achieved thus far in this research studies are satisfactory. I believe that the contributions in this research will advance the field of sensory substitution, in particular in the domain of visual-to-auditory cross-modality conversion. Towards the end of the research phase, observing the recent resurgence of machine learning, in particular in deep learning, I realized that the future of VASS systems is going to be relatively bright. If efforts are invested in combining the state-of-the-art neural networks and visual-to-auditory cross-modality conversion, many more interesting VASS systems can be developed. They are expected to excel in terms of features such as visual recognition, listening experience, and visual-to-auditory conversion mapping. The next generation of VASS systems is slowly emerging as advances in machine learning are achieved, and I hope that this will lead to the greater public adoption of VASS technology. When VASS technology is improved and its capabilities enhanced, more people will understand its benefits. Ultimately, the livelihood of people with visual impairment will be improved with the usage of these new visual-to-auditory sensory substitution systems.

Appendices

Appendix A

Kodak Lossless True Color Image Suite

The pictures from Kodak Lossless True Color Image Suite was used extensively in Project Luminophonics. They were released by the Eastman Kodak Company for unrestricted usage. Kodak Lossless True Color Image Suite hosted by Franzen, 1999 can be found at http://r0k.us/graphics/kodak/.

Figure A.1: Kodak Lossless True Color Image Suite

Appendix B

Test Images for Experiment 1

These are the images used in Experiment 1. They can be categorized into 7 different categories, including: Ball, Bee, Colour, House, Shape, Stick-man, and Tree.

B.1 Ball

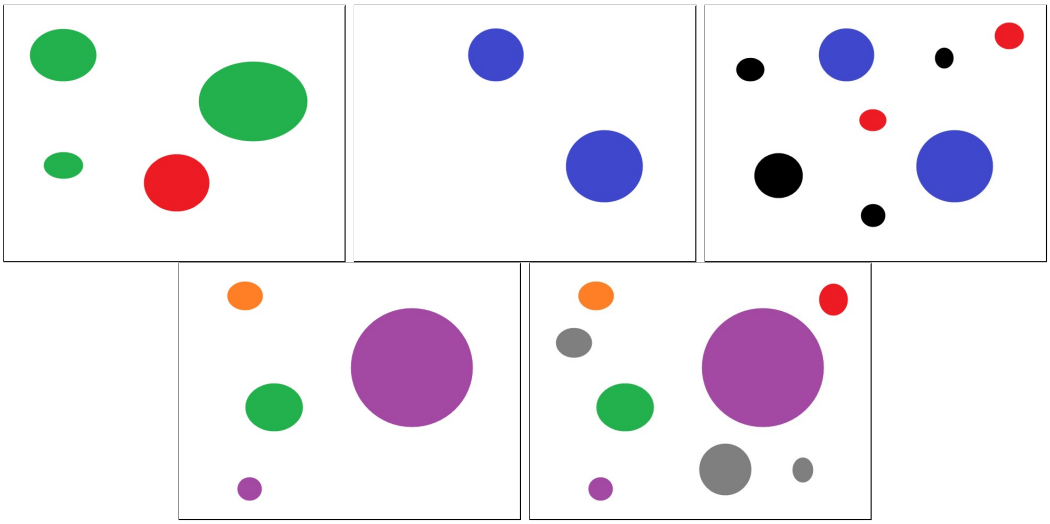


Figure B.1: Ball Test Images for Experiment 1

B.2 Bee

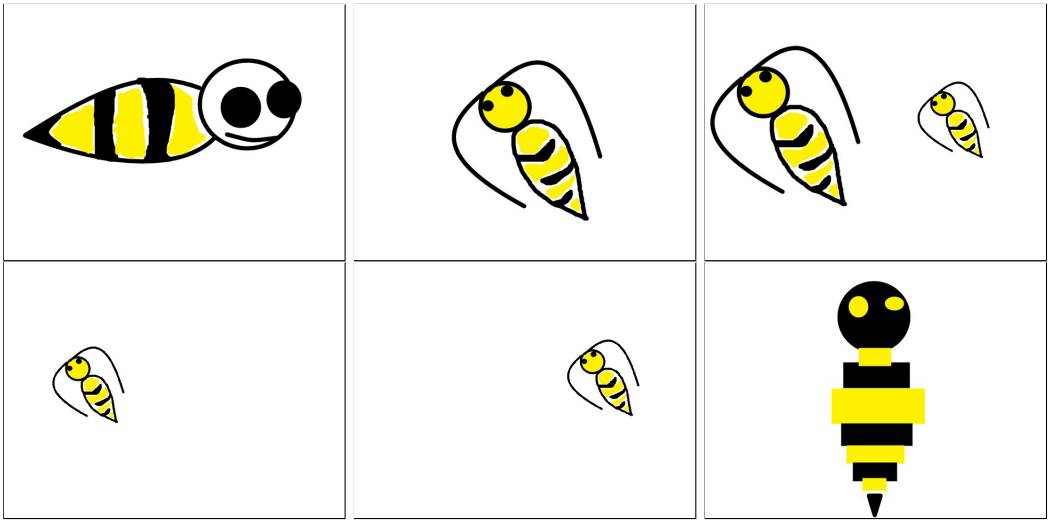


Figure B.2: Bee Test Images for Experiment 1

B.3 Colour

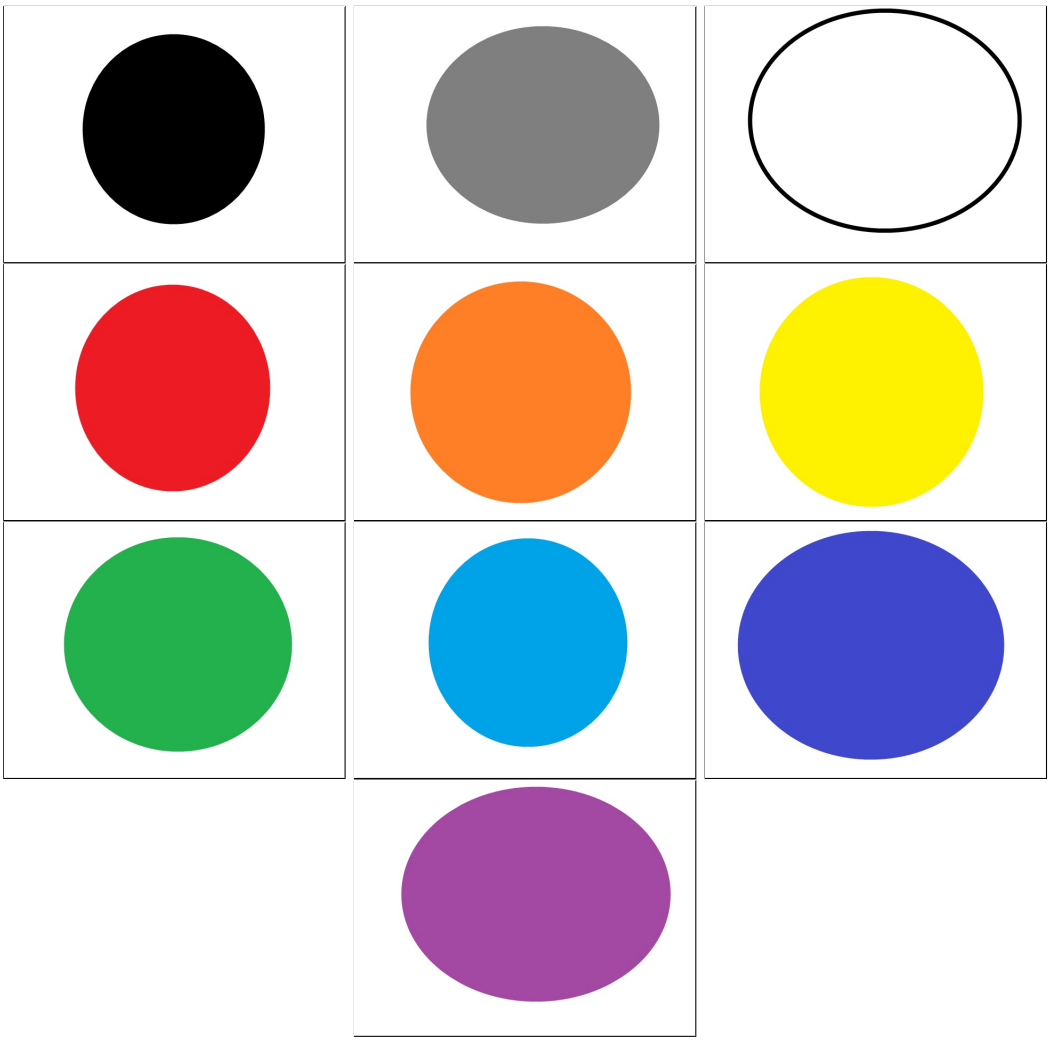


Figure B.3: Colour Test Images for Experiment 1

B.4 House

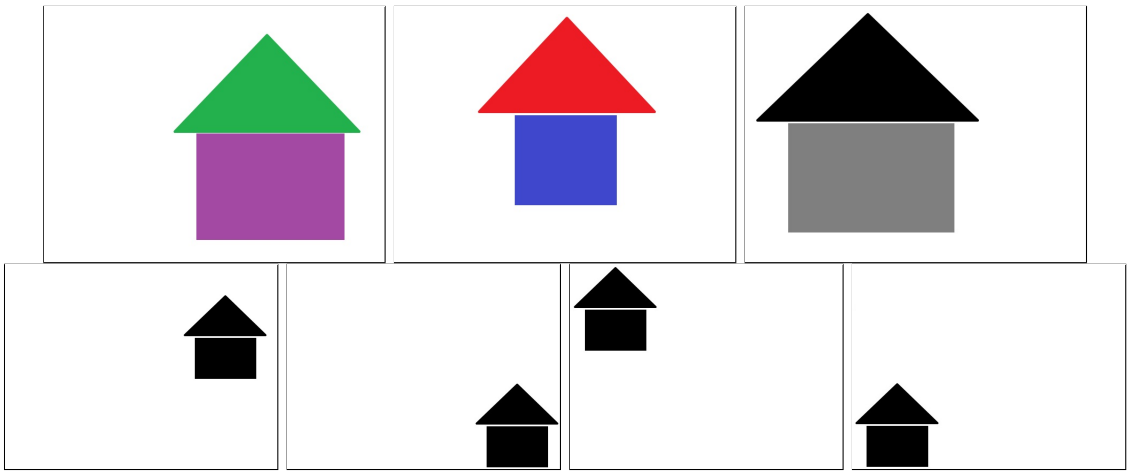


Figure B.4: House Test Images for Experiment 1

B.5 Shape (Shade)

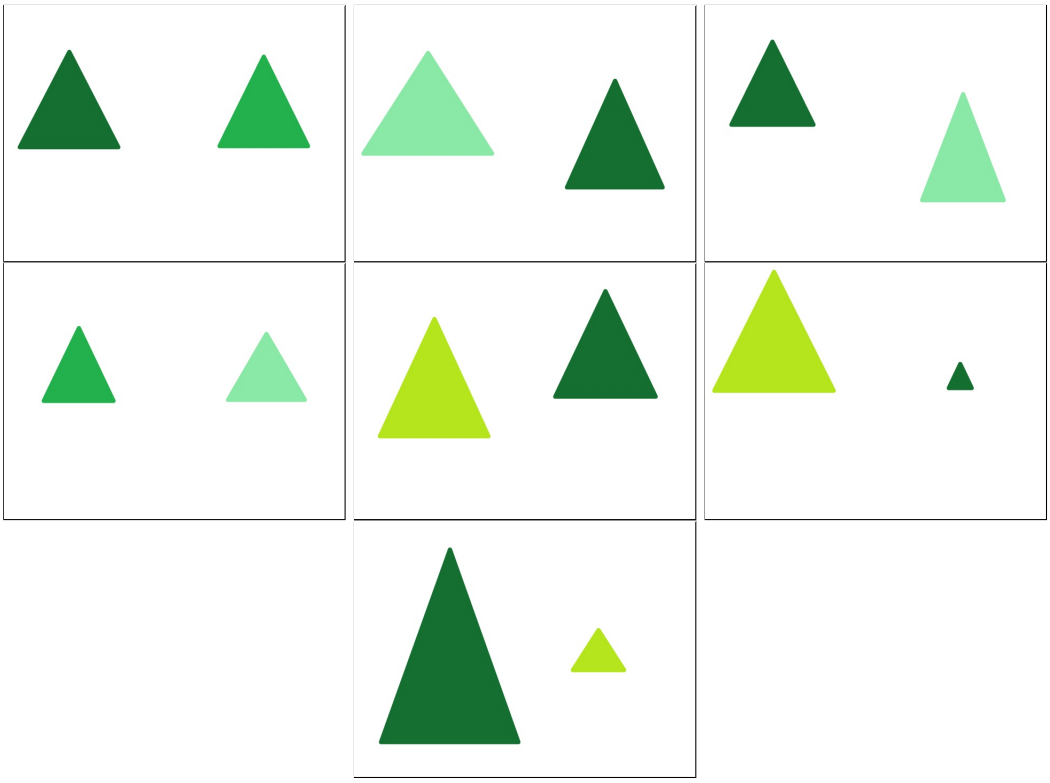


Figure B.5: Shape Test Images for Experiment 1 (Different Shade)

B.6 Shape (Size)

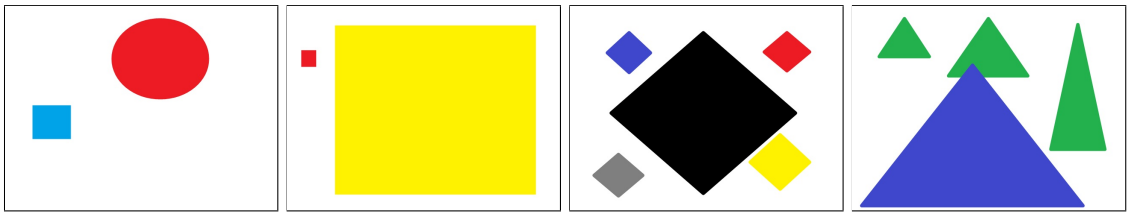


Figure B.6: Shape Test Images for Experiment 1 (Different Size)

B.7 Stick-man

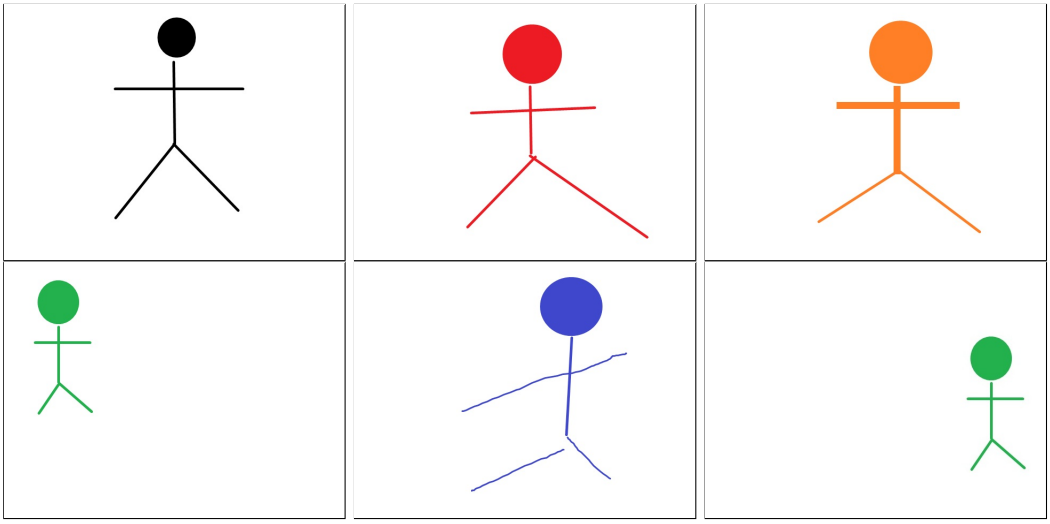


Figure B.7: Stick-man Test Images for Experiment 1

B.8 Tree

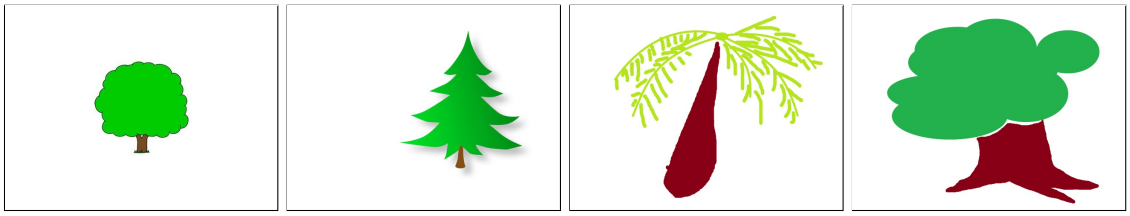


Figure B.8: Tree Test Images for Experiment 1

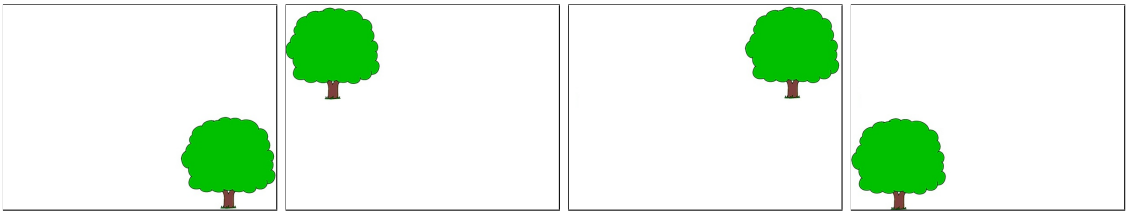


Figure B.9: Tree Images in Different Quadrant for Experiment 1

Bibliography

Abboud, Sami, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi (Jan. 2014). "EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution." In: *Restorative neurology and neuroscience* 32.2, pp. 247–57.

Agarwal, A.K., K. Nammi, Kurt A Kaczmarek, Mitchell Tyler, and D.J. Beebe (2002). "A hybrid natural/artificial electrostatic actuator for tactile stimulation". English. In: *2nd Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Proceedings (Cat. No.02EX578)*. IEEE, pp. 341–345.

Arnold, Dirk V and Nikolaus Hansen (2010). "Active Covariance Matrix Adaptation for the (1+1)-CMA-ES". In: *Genetic And Evolutionary Computation Conference*. Portland, United States, pp. 385–392.

Aucouturier, J.J. and Francois Pachet (2004). "Improving Timbre Similarity : How high's the sky?" In: *Journal of Negative Results in Speech and Audio Sciences* 1.

Bach-y-Rita, Paul (1972). *Brain Mechanisms in Sensory Substitution*. Academic Press Inc.

— (Jan. 1990). "Brain plasticity as a basis for recovery of function in humans". In: *Neuropsychologia* 28.6, pp. 547–554.

— (1995). *Nonsynaptic diffusion neurotransmission and late brain reorganization*. Demos, p. 215.

Bach-Y-Rita, Paul, Carter C Collins, Frank A Saunders, Benjamin White, and Lawrence Scadden (1969). "Vision Substitution by Tactile Image Projection". In: *Nature* 221.5184, pp. 963–964.

Bach-y-Rita, Paul, Yuri Danilov, Mitchell Tyler, and Robert J. Grimm (2005). "Late human brain plasticity: vestibular substitution with a tongue BrainPort human-machine interface". In: *Intellectica* 1, pp. 115–22.

Bach-y-Rita, Paul, Mitchell Tyler, and Kurt A Kaczmarek (Apr. 2003). "Seeing with the Brain". en. In: *International Journal of Human-Computer Interaction* 15.2, pp. 285–295.

Bach-y-Rita, Paul and Stephen W Kercel (Dec. 2003). "Sensory substitution and the human-machine interface." In: *Trends in cognitive sciences* 7.12, pp. 541–6.

Balakrishnan, G., G. Sainarayanan, R. Nagarajan, and Sazali Yaacob (Nov. 2005). "Stereo Image to Stereo Sound Methods for Vision Based ETA". In: *1st International Conference on Computers, Communications, & Signal Processing with Special Track on Biomedical Engineering*. Ieee, pp. 193–196.

Balasuriya, L. S. and J. P. Siebert (2003). "An artificial retina with a self-organised retinal receptive field tessellation. Biologically-inspired Machine Vision, Theory and Application symposium". In: *Biologically-inspired Machine Vision, Theory and Application symposium*, pp. 34–42.

Bedny, Marina, Alvaro Pascual-Leone, David Dodell-Feder, Evelina Fedorenko, and Rebecca Saxe (Mar. 2011). "Language processing in the occipital cortex of congenitally blind adults." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.11, pp. 4429–34.

Benazera, Emmanuel (2015). *libcmaes*.

Bologna, Guido, Benoît Deville, and Thierry Pun (2008). "PAIRING COLORED SOCKS AND FOLLOWING A RED SERPENTINE WITH SOUNDS OF MUSICAL INSTRUMENTS". In: *14th International Conference on Auditory Display*. Paris, France, pp. 2–7.

Bologna, Guido, Benoît Deville, Thierry Pun, and Michel Vinckenbosch (Aug. 2007). "Transforming 3D coloured pixels into musical instrument notes for vision substitution applications". In: *EURASIP Journal on Image and Video Processing* 2007.2, p. 8.

Bologna, Guido and Michel Vinckenbosch (2005). "Eye Tracking in Coloured Image Scenes Represented by Ambisonic Fields of Musical Instrument Sounds". In: *1st International Work-conference on the Interplay between Natural and Artificial Computation*. Canary Islands, Spain, pp. 327–337.

Bradski, Gary (2000). "The opencv library". In: *Dr. Dobb's Journal of Software Tools*.

Bramão, Inês, Alexandra Reis, Karl Magnus Petersson, and Luís Faísca (Sept. 2011). "The role of color information on object recognition: a review and meta-analysis." In: *Acta psychologica* 138.1, pp. 244–53.

Brown, David, Tom Macpherson, and Jamie Ward (2011). "Seeing with sound? exploring different characteristics of a visual-to-auditory sensory substitution device". In: *Perception* 40.9, pp. 1120–1135.

Capelle, C, C Faik, C Trullemans, and C Veraart (1994). "Real time experimental visual prosthesis using sensory substitution of vision by audition". English. In: *Proceedings of 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 255–256.

Capelle, C, C Trullemans, P Arno, and C Veraart (Oct. 1998). "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution." In: *IEEE Transactions on Biomedical Engineering* 45.10, pp. 1279–93.

Cook, Perry R and Gary Scavone (1999). "The Synthesis Toolkit (stk)". In: *Proceedings of the International Computer Music Conference*, pp. 164–166.

Cronly-Dillon, J, K Persaud, and R P Gregory (Dec. 1999). "The perception of visual images encoded in musical form: a study in cross-modality information

transfer." In: *Proceedings of Biological sciences / The Royal Society* 266.1436, pp. 2427–33.

Danilov, Yuri and Mitchell Tyler (Dec. 2005). "Brainport: an alternative input to the brain." In: *Journal of integrative neuroscience* 4.4, pp. 537–50.

Danilov, Yuri, Mitchell Tyler, K L Skinner, R A Hogle, and Paul Bach-y-Rita (Jan. 2007). "Efficacy of electrotactile vestibular substitution in patients with peripheral and central vestibular loss." In: *Journal of vestibular research : equilibrium & orientation* 17.2-3, pp. 119–30.

DeValois, K. K. and M. A. Webster (2011). "Color vision". In: *Scholarpedia* 6.4, p. 3073.

Deville, Benoît, Guido Bologna, Michel Vinckenbosch, and Thierry Pun (2008). "Guiding the focus of attention of blind people with visual saliency". In: *Workshop on Computer Vision Applications for the Visually Impaired*. Marseille, France, pp. 1–13.

Eagleman, David M. (2015). *VEST: A Sensory Substitution Neuroscience Project*.

Ella, Striem-Amit and Miriam A N D Amedi Amir Guendelman (Sept. 2012). "'Visual' Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution". In: *PLOS ONE* 7.3, pp. 1–6.

Findlay, John M and Iain D Gilchrist (Aug. 2003). *Active Vision*. Oxford University Press, p. 236.

Franzen, Rich (1999). *Kodak Lossless True Color Image Suite*.

Fried, Ohad and Rebecca Fiebrink (2013). "Cross-modal Sound Mapping Using Deep Learning". In: *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 531–534.

Fristot, Vincent, Jérémy Boucheteil, Lionel Granjon, Denis Pellerin, and David Alleysson (2012). "Depth Melody substitution". In: *European Signal Processing Conference*. Elsevier, pp. 1990–1994.

Fu, Chang, Chun-Jen Chen, and Chi-Jen Lu (Feb. 2004). "A linear-time component-labeling algorithm using contour tracing technique". In: *Computer Vision and Image Understanding* 93.2, pp. 206–220.

Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Goffaux, Valérie, Corentin Jacques, André Mouraux, Aude Oliva, Philippe G. Schyns, and Bruno Rossion (Aug. 2005). "Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence". en. In: *Visual Cognition* 12.6, pp. 878–892.

Gougoux, Frédéric, Robert J Zatorre, Maryse Lassonde, Patrice Voss, and Franco Lepore (Feb. 2005). "A functional neuroimaging study of sound localization: visual cortex activity predicts performance in early-blind individuals." In: *PLoS biology* 3.2, e27.

Hanassy, Shlomi, Shachar Maidenbaum, Dina Tauber, Amir Amedi, Sami Abboud, and Shelly Levy-Tzedek (May 2013). "EyeMusic: A colorful experience for the blind". In: *Multisensory Research* 26, pp. 116–116.

Hansen, N. and A. Ostermeier (1995). "Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation". In: *Proceedings of IEEE International Conference on Evolutionary Computation*. IEEE, pp. 312–317.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (Mar. 2017). "Mask R-CNN". In:

Hebb, Donald O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, p. 378.

Hubel, D H and T N Wiesel (Feb. 1970). "The period of susceptibility to the physiological effects of unilateral eye closure in kittens." In: *The Journal of physiology* 206.2, pp. 419–36.

Hubel, D H and T N Wiesel (1977). "Functional architecture of macaque monkey visual cortex". In: *Proceedings of the Royal Society B: Biological Sciences* 198.1130, pp. 1–59.

Jenkins, W. M., M. M. Merzenich, M. T. Ochs, T. Allard, and E. Guic-Robles (Jan. 1990). "Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation". In: *J Neurophysiol* 63.1, pp. 82–104.

Jones, Willie D (2004). *Sight for Sore Ears*.

Kaczmarek, Kurt A (Dec. 2011). "The tongue display unit (TDU) for electrotactile spatiotemporal pattern presentation". In: *Scientia Iranica* 18.6, pp. 1476–1485.

Kadir, Timor and Michael Brady (2001). "Saliency, Scale and Image Description". en. In: *International Journal of Computer Vision* 45.2, pp. 83–105.

Karpathy, Andrej and Fei Fei Li (2015). "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

Krantz, John (2012). *Experiencing Sensation and Perception*. Pearson Education, Limited.

Krause, Oswin and Tobias Glasmachers (2015). *A CMA-ES with Multiplicative Covariance Matrix Updates*. Tech. rep.

Kupers, Ron and Maurice Ptito (Aug. 2004). ""Seeing" through the tongue: cross-modal plasticity in the congenitally blind". In: *International Congress Series* 1270, pp. 79–84.

Landragin, Frédéric (Dec. 2004). "Saillance physique et saillance cognitive". Fr. In: *Corela* 2-2.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár (May 2014). "Microsoft COCO: Common Objects in Context". In:

Logan, Beth and Ariel Salomon (2001). "A music similarity function based on signal analysis". In: *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.* 00.C, pp. 745–748.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell (Nov. 2014). "Fully Convolutional Networks for Semantic Segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Maidenbaum, Shachar, Sami Abboud, and Amir Amedi (Apr. 2014). "Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation." In: *Neuroscience and biobehavioral reviews* 41, pp. 3–15.

Maidenbaum, Shachar, Daniel Robert Chebat, Shelly Levy-Tzedek, and Amir Amedi (2014). "Depth-to-audio sensory substitution for increasing the accessibility of virtual environments". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8513 LNCS. PART 1. Springer International Publishing, pp. 398–406.

Meijer, Peter B. L. (Feb. 1992). "An experimental system for auditory image representations". In: *IEEE Transactions on Biomedical Engineering* 39.2, pp. 112–21.

Mennie, Neil, Mary Hayhoe, and Brian Sullivan (May 2007). "Look-ahead fixations: Anticipatory eye movements in natural tasks". In: *Experimental Brain Research* 179.3, pp. 427–442.

Mermelstein, Paul (1976). *Distance measures for speech recognition, psychological and instrumental*.

Merzenich, M. M., J H Kaas, J Wall, R J Nelson, M Sur, and D Felleman (Jan. 1983). "Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation." In: *Neuroscience* 8.1, pp. 33–55.

Najjar, Lawrence J (1996). "Multimedia information and learning". In: *Journal of Educational Multimedia and Hypermedia*.

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng (2011). "Multimodal Deep Learning". In: *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696.

Novich, Scott D. and David M. Eagleman (Oct. 2015). "Using space and time to encode vibrotactile information: toward an estimate of the skin's achievable throughput." In: *Experimental brain research* 233.10, pp. 2777–88.

Oord, Aäron van der, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2015). "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO".

Pampalk, Elias (2004). "A Matlab toolbox to compute music similarity from audio". In: *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, pp. 254–257.

Patla, Aftab E. and Joan N. Vickers (Jan. 2003). "How far ahead do we look when required to step on specific locations in the travel path during locomotion?" In: *Experimental Brain Research* 148.1, pp. 133–138.

Peleg, S., M. Werman, and H. Rom (July 1989). "A unified approach to the change of resolution: space and gray-level". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7, pp. 739–742.

Pohle, Tim, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer (2009). "On rhythm and general music similarity". In: *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. Kobe, Japan, pp. 525–530.

Proulx, Michael J., Petra Stoerig, Eva Ludowig, Inna Knoll, and A Schnitzler (Mar. 2008). "Seeing 'Where' through the Ears: Effects of Learning-by-Doing and Long-Term Sensory Deprivation on Localization Based on Image-to-Sound Substitution". In: *PLoS ONE* 3.3. Ed. by Malika Auvray, e1840.

Proulx, Michael J and Arne Harder (2008). "Sensory Substitution. Visual-to- auditory sensory substitution devices for the blind". In: *Dutch Journal of Ergonomics/Tijdschrift voor Ergonomie*, pp. 20–22.

Ptito, Maurice, Solvej M. Moesgaard, Albert Gjedde, and Ron Kupers (Mar. 2005). "Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind". In: *Brain* 128.3, pp. 606–614.

Puckette, Miller (1996). "Pure Data: another integrated computer music environment". In: *Proceedings of the Second Intercollege Computer Music Concerts (1996)*, pp. 37–41.

Ramachandran, Prajit, Tom Le Paine, Pooya Khorrami, Mohammad Babaeizadeh, Shiyu Chang, Yang Zhang, Mark Hasegawa-Johnson, Roy Campbell, and Thomas Huang (2017). "Fast Generation For Convolutional Autoregressive Models". In: *arXiv preprint*.

Redmon, Joseph and Ali Farhadi (Dec. 2016). "YOLO9000: Better, Faster, Stronger". In:

Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel (Dec. 2016). "Self-critical Sequence Training for Image Captioning". In:

Röder, Brigitte and Frank Rösler (Oct. 2003). "Memory for environmental sounds in sighted, congenitally blind and late blind adults: evidence for cross-modal compensation". In: *International Journal of Psychophysiology* 50.1-2, pp. 27–39.

Rubner, Yossi, Leonidas Guibas, and Carlo Tomasi (1997). "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval". In: *Proceedings of the ARPA Image Understanding Workshop*, pp. 661–668.

Sampaio, Eliana, Stéphane Maris, and Paul Bach-y-Rita (July 2001). "Brain plasticity: 'visual' acuity of blind persons via the tongue". In: *Brain Research* 908.2, pp. 204–207.

Sathian, K. and Simon Lacey (2007). "Journeying beyond classical somatosensory cortex". In: *Canadian Journal of Experimental Psychology* 61.3, pp. 254–264.

Scavone, Gary P and Perry R Cook (2004). "RtMidi, RtAudio, and a Synthesis Toolkit (STK) Update". In: *In Proceedings of the 2005 International Computer Music Conference*. Citeseer.

Schaller, Christian (2011). "Time-of-Flight - A New Modality for Radiotherapy". In: pp. 1–125.

Schaller, Christian, Jochen Penne, and Joachim Hornegger (2008). "Time-of-flight sensor for respiratory motion gating". In: *Medical physics* 35.7, pp. 3090–3093.

Seung, Sebastian (2012). *Connectome: How the Brain's Wiring Makes Us who We are*. Houghton Mifflin Harcourt, p. 359.

Seyerlehner, Klaus, Gerhard Widmer, and Tim Pohle (2010). "Fusing Block-Level Features for Music Similarity Estimation". In: *13th Int. Conference on Digital Audio Effects (DAFx-2010)*. Graz, Austria, pp. 1–8.

Shannon, Claude E (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.July 1928, pp. 379–423.

Shelhamer, Evan, Jonathan Long, and Trevor Darrell (2017). "Fully Convolutional Networks for Semantic Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Simon, Ian and Sageev Oore (2017). *Performance RNN: Generating Music with Expressive Timing and Dynamics*.

Stoneman, Zolinda and Gene H Brody (1983). "Immediate and long-term recognition and generalization of advertised products as a function of age and presentation mode". In: *Developmental Psychology* 19.1, pp. 56–61.

Strasburger, Hans and Ernst Pöppel (1999). "Visual field". In: *Encyclopedia of Neuroscience. 2nd ed. Elsevier, Amsterdam*, pp. 2127–2129.

Tan, Shern Shiou, Tomás Henrique Bode Maul, and Neil Russell Mennie (Jan. 2013). "Measuring the performance of visual to auditory information conversion." In: *PloS one* 8.5, e63042.

— (2015). "Luminophonics experiment: A user study on visual sensory substitution device". In: *PeerJ Preprints*.

Tan, Shern Shiou, Tomás Henrique Bode Maul, Neil Russell Mennie, and Peter Mitchell (2010). "Swiping with Luminophonics". In: *4th IEEE Cybernetics and Intelligent Systems (CIS)*. Singapore, pp. 52–57.

Thomas, B., C. Sage, M. Eyssen, S. Kovacs, R. Peeters, and S. Sunaert (2007). "Brain Plasticity and fMRI". In: *Clinical Functional MRI: Presurgical Functional Neuroimaging*. Ed. by C. Stippich. Medical Radiology. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 209–226.

Thulborn, K. R., P. A. Carpenter, and M. A. Just (Apr. 1999). "Plasticity of Language-Related Brain Function During Recovery From Stroke". In: *Stroke* 30.4, pp. 749–754.

Tyler, Mitchell, Yuri Danilov, and Paul Bach-y-Rita (Dec. 2003). "CLOSING AN OPEN-LOOP CONTROL SYSTEM: VESTIBULAR SUBSTITUTION THROUGH THE TONGUE". In: *Journal of Integrative Neuroscience* 02.02, pp. 159–164.

Tyler, Mitchell, S. Haase, Kurt A Kaczmarek, and Paul Bach-y-Rita (2002). "Development of an electrotactile glove for display of graphics for the blind: preliminary results". English. In: *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*. Vol. 3. IEEE, pp. 2439–2440.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan (Apr. 2017). "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4, pp. 652–663.

Vukotić, Vedran, Christian Raymond, and Guillaume Gravier (2016). "Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking". In: *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion - iV&L-MM '16*. New York, New York, USA: ACM Press, pp. 37–44.

Wang, Kaiye, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang (July 2016). "A Comprehensive Survey on Cross-modal Retrieval". In: *arXiv preprint*.

Ward, Jamie and Peter B. L. Meijer (Mar. 2010). "Visual experiences in the blind induced by an auditory sensory substitution device." In: *Consciousness and cognition* 19.1, pp. 492–500.

Weiller, C, C Isensee, M Rijntjes, W Huber, S Müller, D Bier, K Dutschka, R P Woods, J Noth, and H C Diener (June 1995). "Recovery from Wernicke's aphasia: a positron emission tomographic study." In: *Annals of neurology* 37.6, pp. 723–32.

Wong, F., R. Nagarajan, Sazali Yaacob, A. Chekima, and N.-E. Belkhamza (2000). "A stereo auditory display for visually impaired". In: *TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium*. Ieee, pp. 377–382.

World Health Organization (2003). *WHO — Up to 45 million blind people globally - and growing*.

— (2014). *Visual impairment and blindness*.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio (Feb. 2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *International Conference on Machine Learning*.

Yeo, Woon Seung and Jonathan Berger (2006). "Application of raster scanning method to image sonification, sound visualization, sound analysis and synthesis". In: *International Conference on Digital Audio Effects*. Montreal, Canada, pp. 309–314.

Yeo, Woon Seung, Jonathan Berger, and Zune Lee (2004). "SonART: A framework for data sonification, visualization and networked multimedia applications". In: *Proceedings of the 2004 International Computer Music Conference*, pp. 180–184.

Yip, Andrew W and Pawan Sinha (Jan. 2002). "Contribution of color to face recognition". en. In: *Perception* 31.8, pp. 995–1003.