Towards a better understanding of sensory substitution

The theory and practice of developing visual-to-auditory sensory substitution devices.

By **Thomas D. Wright**

Submitted for the degree of

PhD in Psychology

To the

University of Sussex in **September 2013**

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signed: .....

Date: .....

Thesis summary

Visual impairment is a global and potentially devastating affliction. Sensory substitution devices have the potential to lessen the impact of blindness by presenting vision via another modality. The chief motivation behind each of the chapters that follow is the production of more useful sensory substitution devices. The first empirical chapter (chapter two) demonstrates the use of interactive genetic algorithms to determine an optimal set of parameters for a sensory substitution device based on an auditory encoding of vision ("the vOICe"). In doing so, it introduces the first version of a novel sensory substitution device which is configurable at run-time. It also presents data from three interactive genetic algorithm based experiments that use this new sensory substitution device. Chapter three radically expands on this theme by introducing a general purpose, modular framework for developing visual-to-auditory sensory substitution devices ("Polyglot"). This framework is the fuller realisation of the Polyglot device introduced in the first chapter and is based on the principle of End-User Development (EUD). In chapter four, a novel method of evaluating sensory substitution devices using eye-tracking is introduced. The data shows both that the co-presentation of visual stimuli assists localisation and that gaze predicted an auditory target location more reliably than the behavioural responses. Chapter five explores the relationship between sensory substitution devices and other tools that are used to acquire real-time sensory information ("sensory tools"). This taxonomy unites a range of technology from telescopes and cochlear implants to attempts to create a magnetic sense that can guide further research. Finally, in chapter six, the possibility of representing colour through sound is explored. The existence of a crossmodal correspondence between (equi-luminant) hue and pitch is documented that may reflect a relationship between pitch and the geometry of visible colour space.

Acknowledgements

My PhD represents the single most challenging period of my life to date, but with this challenge has come enormous opportunity for personal growth. I am deeply grateful to all those who have helped and supported me on my journey. I feel very privileged to have met so many wonderful people along the way. (And I'm very pleased that those who knew me beforehand have stuck around!)

A particular debt of gratitude is owed to my supervisor Prof. Jamie Ward. As well as providing the support to see my ideas realised, Jamie supported me when I felt out of place and encouraged me to persevere. Most of all, Jamie has been a formidable intellectual sparring partner and an enormously constructive collaborator.

This gratitude extends, of course, to the members of all the lab groups I've been associated with over the last few years. The synaesthesia group in particular have become like a second family to me, but I am also very grateful for the support of the Human Centred Technology group and the members of the Sackler Centre for Consciousness Science. Thanks also to the Psychology department as a whole. They have all been so friendly and helpful – my PhD has felt much more manageable because of the warmth I've experienced in the corridors of Pevensey I and II.

In particular, there are three members of faculty who have consistently gone out of their way to help me out. Thanks to Sam Hutton for his support with the eye-tracker, to Ryan Scott for his help with Matlab, and finally thanks to Anna Franklin who helped me with the colour-related aspects of my research. All three of them are also owed thanks for their continued encouragement.

Special thanks are due to my project students over the years, who have helped me develop my ideas and who have piloted experimental designs. Thanks to Sarah Simonon and Aaron Margolis for their efforts in developing the "Where's Wally" paradigm (even if we did drop the name!) Thanks also to Stefanie Maurer for her assistance with the interactive genetic algorithm experiments.

I also want to thank my friends and family for their support over the last few years. The final stages of a PhD have made me into an inattentive friend / son / brother in recent months, so huge thanks for tolerating me! A special mention here must go to Martha Casey and Nikki Luke and Clare Jonas who kept me sane during whilst I was on campus! They kept me fuelled with caffeine and even managed to stimulate me with non-academic conversation. Hopefully my

post-PhD self will have a little more free time with which to perfect my Bowie karaoke routine. Thanks also to my mum (Jackie) and dad (Nick) for their sound advice and hot meals, and to my siblings Joe, Sam and Megan for their humour, which has managed to cheer me up even at the lowest points of my PhD journey.

Finally, I want to thank my long-suffering girlfriend Lisa. Over the last four years she has regularly been put into competition with my research (for my time, if not for my affection). She, more than any other, has listened to my ramblings and helped me shape them into ideas worth committing to paper.

For knowing you all, I am a very lucky man. Thank you.

Contents

Chapter 1: Introduction and general discussion

We live in a highly visual world. For many sighted people, living without vision is difficult to comprehend. Such a great proportion of our daily life is oriented around, and dependent on, our visual experience.

But seeing is not a purely functional thing. Vision is central to the human experience. Together with audition, vision may be considered a "major" sense. Indeed, in the artificial facsimiles of reality – our plays, operas, movies and television shows – we are largely content with the reproduction of just these two senses. Though we move inexorably towards deeper multisensory entertainment experiences, sight and sound remain central. Huxley's "feelies" have yet to materialise and "Smell-O-Vision" never really caught on.

Yet for hundreds of millions of people around the world, the visual experience is diminished or inaccessible. Visual impairment, in one form or another, is thought to affect around 285 million people (World Health Organisation, 2012). Worldwide, it is estimated that 5.7 people in every thousand are "blind" and another 20 people in every thousand have "low vision". The vast majority (over 80%) of those affected are aged 50 or older, reflecting the degenerative nature of many of the conditions that result in visual impairment (Resnikoff et al., 2004). That said, it is estimated that in some developing countries, incidence of visual impairment in children (younger than 16) is as high as 1.1 per thousand (Gilbert, Anderton, Dandona, & Foster, 1999). In these regions, a large proportion of these instances are either preventable or reversible.

At this point, it is important to note that the terms visual impairment, blindness and low vision are often used without strict adherence to any particular set of definitions. In this context, the definitions of blindness and low vision are taken from the World Health Organisation's "Family of International Classifications" (WHO-FIC), which publishes the International Classification of Diseases (ICD). In the 2010 version of the ICD, visual impairment (including both low vision and blindness) occupies section H54. The ICD distinguishes between low vision and blindness primarily using the visual acuity of the better eye using the best possible correction. In imperial measures, a person may be classified as having low vision if their acuity is worse than 20/70 (i.e. they are able to see at 20 feet what a typical person could see at 70), or as being blind if their acuity is worse than 20/400 (World Health Organization, 2010). Since many forms of visual impairment affect the eye in a non-uniform manner, the ICD also states that those with a visual field with a radius of less than 10° should also be classified as blind.

Ninety percent of blind people live in a developing country (Thylefors, 1998). Moreover, 80% of visual impairment is “avoidable” and could either be prevented or treated. These avoidable cases of visual impairment include infectious agents, cataracts and even uncorrected refractive error. Indeed, just over half of worldwide incidence of visual impairment can be attributed to uncorrected refractive error despite the ease with which this can be treated (Resnikoff, Pascolini, Mariotti, & Pokharel, 2008). That this situation exists is deeply disturbing, but the solution must be driven by policy rather than by science. Programmes that use existing technologies to correct refractive errors and remove cataracts remain the most effective means of reducing visual impairment in developing countries (Fotouhi, Hashemi, Mohammad, & Jalali, 2004).

In more developed countries, the much lower rates of visual impairment can be attributed largely to the prevention and treatment of avoidable causes. The result of this is a greater proportion of visual impairment is caused age-related and degenerative. The authors of a recent meta-analysis estimate that, due solely to age-related macular degeneration (a degenerative condition of the retina), 3.5% of people over 75 are visually impaired (Owen, Fletcher, Donoghue, & Rudnicka, 2003).

That isn’t to say that blindness in young people is absent from more developed countries – in the UK, for example, 0.59 children (younger than 16) per thousand are visually impaired. Of these, the majority (61%) were born with their condition, often (33%) due to an hereditary condition (Rahi & Cable, 2003). People born blind are sometimes referred to as congenitally, or “early” blind, compared to the “late” blind, whom have lost their vision later in life. Making matters worse, these numbers are almost certainly underestimates: slightly less than half of those British people eligible are registered as visually impaired (R. Robinson et al., 1994).

In between children and the elderly, adults in more developed countries are most likely to become visually impaired as a result of trauma (Parver, 1986). There are many ways in which an injury can lead to visual impairment. These include bleeding, dislocation of the retina or lens, and rupturing of the globe (Tumulty & Resler, 1984). Trauma is often sustained during sport or at work and efforts to improve safety in these environments has been successful in reducing accident rates (Parver, 1986). Though most cases result in the monocular loss of sight, instances of bilateral trauma can be particular disruptive to the individual due to the suddenness of their onset. At the societal level, the fact that those affected by ocular trauma tend to be of working age means that the economic effects are particularly worrisome (Congdon, Friedman, & Lietman, 2003).

Alongside efforts to reduce the number of people affected by visual impairment, the potentially debilitating consequences make it vital to support those whose blindness we weren't able to prevent. The book "Touching the rock" makes clear the profound impact blindness has upon everyday life (Hull, 1992). From the practical (locating conversational partners; requiring chicken to be cut from the bone) to the personal (powerlessness; being treated as a child), the author meticulously catalogues the struggles (and triumphs) of modern life as a blind adult. Strikingly, many of the issues the author describes are societal or interpersonal. This is echoed in "The making of blind men", the central tenet of which is that many of the problems experienced by blind people are caused by societal beliefs about the nature of blindness and the institutional responses to which these beliefs give rise (Scott, 1969).

It is clear then, that science and technology are not a panacea when it comes to visual impairment. In developing countries in particular, scientific innovations are still not being exploited to the fullest possible extent; the need for policy to apply existing technology is therefore greater than the need for new innovations. Furthermore, many problems faced by those who are visually impaired will not be alleviated by the application of technology. With that proviso, it is important to note that visual impairment has been (and will likely continue to be) the target of a wide range of technological innovations.

1.1 Assistive technologies

Assistive technology is a term used to describe tools which aid adaptation or rehabilitation in people with disabilities. Examples of assistive technology include wheelchairs, prostheses, hearing aids and personal alarms. Some of the best recognised assistive technology is designed for people with visual impairments. These assistive technologies are almost all designed to assist with either mobility or communication (Scherer, 1996). In the first category, the best known examples are the long cane and the guide dog. Guide dogs facilitate locomotion at speeds closer to an individual's preferred walking speed than do canes (Clark-Carter, Heyes, & Howarth, 1986) and may provide other, non-mobility related, benefits (Lane, McNicholas, & Collis, 1998), but will not be suitable for all visually impaired people. In addition to these relatively prosaic aids, mobility has been a fertile ground for high-tech developments, such as personal navigation systems (Ivanov, 2012; Loomis, Golledge, & Klatzky, 2006) and obstacle detectors (Farcy & Damaschini, 2001; Shoval, Ulrich, & Borenstein, 2003). Significantly, despite the availability so many types of mobility aid, it is estimated that approximately 30% of visually

impaired people do not journey out of their homes independently (as cited in Clark-Carter et al., 1986).

Although traditionally Braille was the primary mode of stored communication for visually impaired people, this is now in competition with synthesised speech systems. As communication has become increasingly mediated by digital technology, Braille has moved away from printed forms to refreshable displays. This same shift however, has reduced the significance of Braille, as recorded and synthesised speech systems become more practical and accessible. Why spend money on a refreshable Braille display when your computer can be loaded with text-to-speech (TTS) software (Earl & Leventhal, 1999)? Why would a refreshable Braille display be preferable to a Daisy digital book player that can also be used to play music CDs (Leventhal & Holborow, 2001)? Nor is the speed of speech a limiting factor, as visually impaired people have been shown to be capable of accessing speech at much faster than typical speeds (Asakawa, Takagi, Ino, & Ifukube, 2003; Stent, Syrdal, & Mishra, 2011).

Additionally, there are some assistive technologies for visually impaired people that are designed to make the best use of any remaining vision. These systems are called “Low Vision Aids” and typically operate by magnifying a portion of the visual field. These can be as simple as a magnifying glass or as complex as a closed-circuit television system (den Brikner & Beek, 1996).

With the exception of low vision aids, most assistive technologies for visually impaired people are designed with a particular task or behaviour in mind. Beyond the examples given above, some visually impaired people make use of:

- **symbol canes:** to alert sighted people to the presence of a visually impaired person.
- **liquid level indicators:** to assist with the filling of containers (e.g. making tea)
- **writing frames:** to keep writing within a box (e.g. when signing a cheque)
- **speaking clocks:** which announce the time on request
- **adapted phones:** that make it easier to dial

- Examples described by Dudley (1990)

Despite this variety in assistive technologies however, a systematic review found visually impaired people to be much less likely to be adequately supported than people with other forms of disability (Alper & Raharinirina, 2006). The authors of that review speculate that this is due primarily to the complexity (and associated costs) of assistive technologies for visually impaired people. It is unclear how the aforementioned levels of under-reporting of visual

impairment interact with this finding. Given the impact blindness can have on a person's life, it seems probable that the bulk of those under-served by assistive technology fall into the low vision category.

Fundamentally though, even those visually impaired people who are able to access and use assistive technologies are only given the ability to perform formerly-visual tasks. None of the assistive technologies described above attempt to convey any of the *quality* of sight. It's easy to see why functional assistance has been the priority over the years; in a world built for the sighted, so many tasks become extraordinarily difficult when vision is lost. That said, not all facets of the visual experience can be anticipated by a dedicated form of assistive technology. Accordingly, there exists a much more general form of sensory assistive technology, which aims not to aid the user with a particular task, but instead attempts to present the raw visual information using a functioning sensory modality: the sensory substitution device.

1.2 Sensory substitution devices

In general terms, a sensory substitution device is one which takes sensory information from one modality, transforms it, and presents it in a different modality. In theory, sensory substitution devices can exist for any pairing of the sense, but in practice it is vision which has enjoyed the primary focus of sensory substitution researchers.

1.2.1 Tactile devices

As a field, the study of sensory substitution is relatively young. The first sensory substitution device was described as recently at 1969 by Paul Bach-y-Rita. This device – which Bach-y-Rita called the Tactile Vision Sensory Substitution (TVSS) device – converted the visual signal from a camera to tactile information delivered onto the skin on the user's back (Bach-y-Rita, Collins, Saunders, White, & Scadden, 1969). The process was achieved using a television studio camera and 400 vibrating solenoids set into the back of a repurposed dentist's chair. The solenoids were arranged in a square 20 to a side, such that each corresponded to a

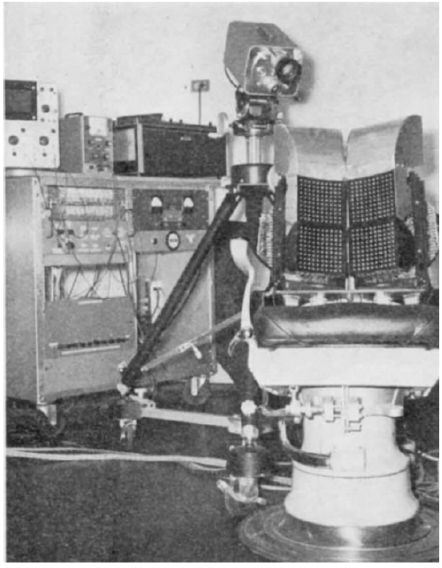


Figure 1: The original TVSS apparatus (Bach-y-Rita *et al.*, 1969)

pixel from the camera. The intensity of the vibration of each solenoid was linked to the brightness of the relevant pixel, so that the pattern of vibration on the user's skin would form a tactile picture of sorts.

Despite the relatively simple setup (which, by modern standards, looks almost medieval), the TVSS gave rise to some revolutionary findings. After very short interactions with the system, participants were able to use it to identify shapes and make accurate judgements related to acuity and orientation (White, Saunders, Scadden, Bach-y-Rita, & Collins, 1970). Beyond these cognitive, functional results, anecdotal evidence suggests a more visceral connection between sensory substitution device and participant. For instance, an experimenter increased the zoom of the camera lens without previously informing the participant, who promptly lurched backwards, away from where the camera was pointed (Bach-y-Rita, 1972). This suggests that the participant was – at a fundamental level – re-localising the sensation from their back to somewhere in front of them; a property of vision, not touch.

Bach-y-Rita's lab has continued to develop tactile-visual sensory substitution devices. In the early '90s, they began to explore the use of electrodes and electro-tactile stimulation instead of the solenoids and vibro-tactile stimulation employed by the TVSS. Early devices to use this technique acted on the fingertip (Kaczmarek & Haase, 2003), but later efforts concentrated on the tongue as a receptive surface because of its greater conductivity (Bach-y-Rita, Kaczmarek, Tyler, & Garcia-Lara, 1998).The principle underlying these electro-tactile systems is identical to that of the TVSS: stimulators are arranged in a two-dimensional array and are paired with a pixel in a corresponding location within the source image; the intensity with which the stimulator is dependent on the brightness of this pixel. Although this tongue-based system originally only used 49 points of stimulation (in a seven-by-seven grid), participants were shown to be able to discriminate between simple shapes with a high degree of accuracy (Bach-y-Rita et al., 1998). Later versions would see the number of electrodes increase to 144 (Bach-y-Rita, 2004; Kaczmarek, 2011).

This device, which was later named the Tongue Display Unit (TDU), is also significant because it was the first sensory substitution device designed to substitute more than just vision (Bach-y-Rita & Kercel, 2003). In the early 2000s, Bach-y-Rita's lab (which by this point had morphed into Wicab Inc.) demonstrated that the TDU could be used to provide feedback about balance to people with bilateral vestibular loss (Danilov, Tyler, Skinner, & Bach-y-Rita, 2006; Tyler, Danilov, & Bach-y-Rita, 2003). Consequently, users demonstrated increased postural stability

and reported feeling less “wobbly”. Furthermore, these effects persisted even after removal of the device.

1.2.2 Auditory devices

In the same vein, it was during this period that the field first explored the use of a modality other than touch to do the substituting. In 1992, a Dutch engineer called Peter Meijer launched a sensory substitution device that he called the vOICe (the capitalised letters spelling “Oh, I see!”). The vOICe was novel because it conveyed the visual information using sounds. Operating like an inverse spectrograph, the vOICe scans over the image (from left to right) once a second and sonifies only a single column of pixels at any given time. Time therefore represents horizontal position. Additionally, cues to horizontal location are given by stereo panning from the left ear to the right ear. The vertical axis of the image is represented by pitch, so that each frequency represents a row of pixels in the image. As in the tactile devices, the brightness of each pixel determines the amplitude of each point of stimulation (though this is occasionally inverted, so that amplitude decreases with lightness). Unlike the tactile devices though, this amplitude is realised by the loudness of each frequency.

Taken together, a pixel value is represented by a combination of time (horizontal position), interaural volume difference (horizontal position), frequency (vertical position), and loudness (brightness). The vOICe originally was realised in hardware and featured a resolution of 64 by 64 pixels (Meijer, 1992), but more recent versions are software based and (on a PC) have increased the horizontal resolution to 176 pixels (Ward & Meijer, 2010). Since its creation, the vOICe has become the sensory substitution device of choice amongst the research community – it has been widely used in experiments exploring sensory substitution as an ability (Brown, Macpherson, & Ward, 2011; Haigh, Meijer, & Proulx, 2013), functional brain activity during sensory substitution device use (Amedi et al., 2007; Merabet et al., 2009), and the phenomenology reported by sensory substitution device users (Ward & Meijer, 2010). Many of these studies are explored in greater detail below. The vOICe also serves as the starting point for the second chapter of this thesis.

The vOICe, however, is not the only visual-auditory sensory substitution device. Since its creation, it has been succeeded by a variety of similar systems. Two of the more notable successors both attempt to overcome the one-second refresh rate that is the inevitable result of use of time to represent the horizontal axis by the vOICe. As pointed out by the creators of one of these systems: “the time-multiplexed system of Meijer can be considered as working in real-time, but the refreshing rate (about one Hz) does not allow rapid sensory–motor

interactions, as required in several tasks such as, e.g., mobility or reading" (Capelle, Trullemans, Arno, & Veraart, 1998).

The first of these newer auditory sensory substitution devices was (eventually – not in the original papers) called the Prosthesis Substituting Vision for Audition (PSVA). To represent the horizontal axis without breaking the stream of images into frames, the PSVA instead multiplexes the frequencies. By using increasing octaves to represent the vertical dimension, increasing notes within each octave can be used to represent positions from left to right. This, of course, can only work with a smaller resolution than the vOICe – there aren't 4,096 (64 x 64) musical notes available within a range audible to humans – so the PSVA features a much coarser resolution. To overcome this, it uses a weighted distribution of pixels based on the foveal region of the retina (Capelle et al., 1998). In addition to these innovations, the PSVA also uses stereo panning to assist in horizontal judgements as the vOICe does.

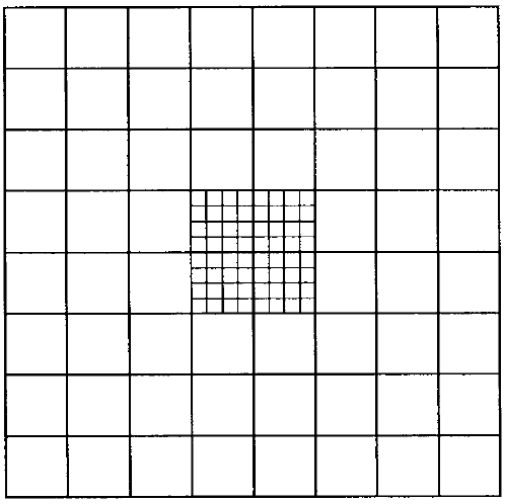


Figure 2: Diagram of the "artificial retina" used by the PSVA (Capelle et al., 1988)

The Vibe is, perhaps, halfway between the vOICe and the PSVA. What sets it apart is that it relies solely on interaural differences to describe horizontal location. Since pitch is once again only used to convey vertical position, the resolution need not be limited in the same way as the PSVA. On the other hand, the acuity of our auditory localisation is still limited. Accordingly, the Vibe divides the scene into approximately 20 evenly distributed "receptive fields" (Durette, Louveton, Alleysson, & Hérault, 2008).

More radically, the SmartSight uses the Sobel edge detection algorithm to break the image into a collection of oriented lines. It is these lines, rather than the pixels of the original image, which are then sonified (Cronly-Dillon, Persaud, & Gregory, 1999). Like the vOICe, SmartSight is time multiplexed and uses pitch (specifically 50 musical notes in this implementation) to represent vertical position.

The number of visual-auditory systems – and their popularity among the research community – is probably due to the relative ease with which one can implement an auditory sensory substitution device. Unlike tactile systems, which necessarily require specialist hardware for

their output, auditory systems only require a set of headphones. This point is demonstrated by the paucity of tactile systems available. Aside from the TDU (which is expensive to buy), tactile systems are rare. Furthermore, when a new tactile system is produced, it rarely features as many "taxels" as the original TVSS or even the TDU. The "Minimal TVSS", for instance, only uses a three by two grid of vibrating motors (Bird, Marshall, & Rogers, 2009) and the Encative Torch (which conveys proximity information rather than a visual signal) uses only one (Froese, McGann, Bigge, Spiers, & Seth, 2012).

Another factor which contributes to the variety of visual-auditory sensory substitution devices is the lack of an approach with the immediate obviousness of that utilised by tactile sensory substitution devices – human auditory localisation has too poor a resolution to rely on spatial cues in the way the TVSS does. This also means that it is difficult to objectively ascertain the optimal transformation for a visual-auditory sensory substitution device to use. Chapters two and three of this thesis each offer a potential strategy to overcome this problem. Chapter two demonstrates the use of interactive genetic algorithms to efficiently search for a set of parameters that approximate an optimal set. In chapter three it is acknowledged that it no one device is likely to be optimal for all users or situations. Chapter three therefore introduces a general purpose, modular framework for the development of sensory substitution devices called Polyglot.

A potentially interesting avenue of exploration is the combination of tactile and auditory signals into a single, hybrid visual sensory substitution device. This is achievable using the Polyglot framework described in chapter three. Indeed, a sensory substitution device called Creole has been produced which uses a touchpad to control the region of interest and simultaneously deliver haptic feedback whilst the user listens to high fidelity auditory output.

1.2.3 Colour devices

All the aforementioned sensory substitution devices convey only greyscale visual information. They take brightness to be amplitude and convey this as either loudness or the intensity of a vibration. Yet lightness is colour's poor cousin: lightness is often enough to infer form, but colour gives vibrancy. Even beyond aesthetics, colour information also assists with several visual functions including scene segmentation (Cheng, Jiang, Sun, & Wang, 2001), object identification (Tanaka, Weiskopf, & Williams, 2001), and face recognition (Yip & Sinha, 2002). Why then, has the research focused predominantly on substituting greyscale vision?

The answer lies in the difficult problem of mapping the additional dimensions. All modern colour models use at least three dimensions (Hunt & Pointer, 2011). On computers colours are

represented by their constituent proportions of red, blue and green. Printers use cyan, magenta and yellow. In psychophysics, we might describe a colour in terms of its “x”, “y” and “Y” properties or in terms of “hue”, “chroma” and “value”. Whereas greyscale vision only requires that we map brightness onto a property of another modality, full colour vision triples this demand.

The time multiplexing approach of the vOICe hints at the dimensional shortage in audition compared to vision, but in fact the problem goes deeper. The core problem is one of the amounts of information that each sense is capable of conveying. In other words, it comes down to bandwidth. Bandwidth may be compared across data types and modalities by leveraging information theory to calculate the number of bits (binary digits) per second (bps). Whereas the eye has been estimated to have capacity for 4.3×10^6 bps (Jacobson, 1951), our ears are only estimated to be able to carry 10^4 bps and our skin (at its most sensitive; on the fingertips) only 100 bps (Kokjer, 1987).

As a consequence, designers of colour sensory substitution devices are forced to either pare down some other aspect of vision or dream up complicated methods of encoding complex information. The former approach is the rationale behind the most common strategy for colour sensory substitution devices: conveying colour information from a single point in the scene. An early manifestation of this was a device which mapped the central point of an image to a named colour, then played a recording of this name (McMorrow, Wang, & Whelan, 1997). Though this could certainly be counted as an assistive technology, its output isn’t so much sensory as symbolic, which makes it a bad fit amongst sensory substitution devices. (The distinction between sensory and symbolic systems is expanded upon in chapter 5.)

From a sensory substitution perspective, the later device produced by Adam Montandon for Neil Harbisson – a sighted artist with achromatopsia (profound colour blindness). Like the example described above, this device samples a single point in the visual scene. The hue information from this point is then mapped to a pitch using a relationship inspired by the electromagnetic frequency of the colour spectrum (Hauskeller, 2012; Wade, 2005). The Kromophone is a similar system, but also uses stereo panning to aid in colour differentiations (Capalbo & Glenney, 2009).

The See ColOR (Seeing colour with an orchestra) system is an example of the latter approach. Instead of selecting a single point to represent, it presents colour information from a row of 25 adjacent regions in the visual scene. Colours are mapped to a restricted set of hue identities,

each of which is associated with the sound of a musical instrument (“blue”, for instance, is represented by the sound of a piano). The saturation of each region is used to determine which note of the instrument sound is to be played. Luminance information is used to determine whether the region should be represented by a colour, disregarded for being too dark, or represented as white. The 25 lateral regions are distinguished using auditory localisation cues (Gomez, Bologna, & Pun, 2010). A later iteration of the system also encoded the depth of each region by altering the durations of the sounds that represent them (Bologna, Deville, & Pun, 2010).

The “Electro-Neural Vision System” (ENVS; pronounced as “envious”) is interesting because it is a tactile colour sensory substitution device. Using a pair of gloves fitted with an electrode for each finger, the ENVS can convey information about proximity and colour for each of 10 vertical slices of the visual scene. The former is represented by the intensity of electro-tactile pulses. The latter involves a user-defined mapping between eight colours and frequencies (Meers & Ward, 2004).

As with aforementioned variability among visual-auditory sensory substitution devices, the number of approaches to representing colour speaks to the underlying lack of an obvious relationship between colour and sound or touch. Some aspects of colour do have well established links to sound. As utilised by most visual-auditory devices, there is a strong cross-modal correspondence between lightness and loudness for example (Marks, Hammeal, Bornstein, & Smith, 1987). The existence of a cross-modal correspondence between hue and pitch is less clear cut and the principle subject of the empirical work described in chapter six.

1.2.4 Sensory augmentation

As mentioned briefly above, it is possible to apply the principles of sensory substitution to entirely novel forms of perception. This is referred to in the literature as sensory augmentation. The best example of sensory augmentation is the FeelSpace project, which takes orientation information from a digital compass and feeds it to an array of vibro-tactile pads mounted on a belt, such that only the pad facing north vibrates at any given moment (Nagel, Carl, Kringe, Märtin, & König, 2005). Participants who train with the belt for six weeks are able to integrate this novel sense with other cues to boost their performance on navigation tasks. Intriguingly, half of these participants also report qualitative changes as they progress through this training:

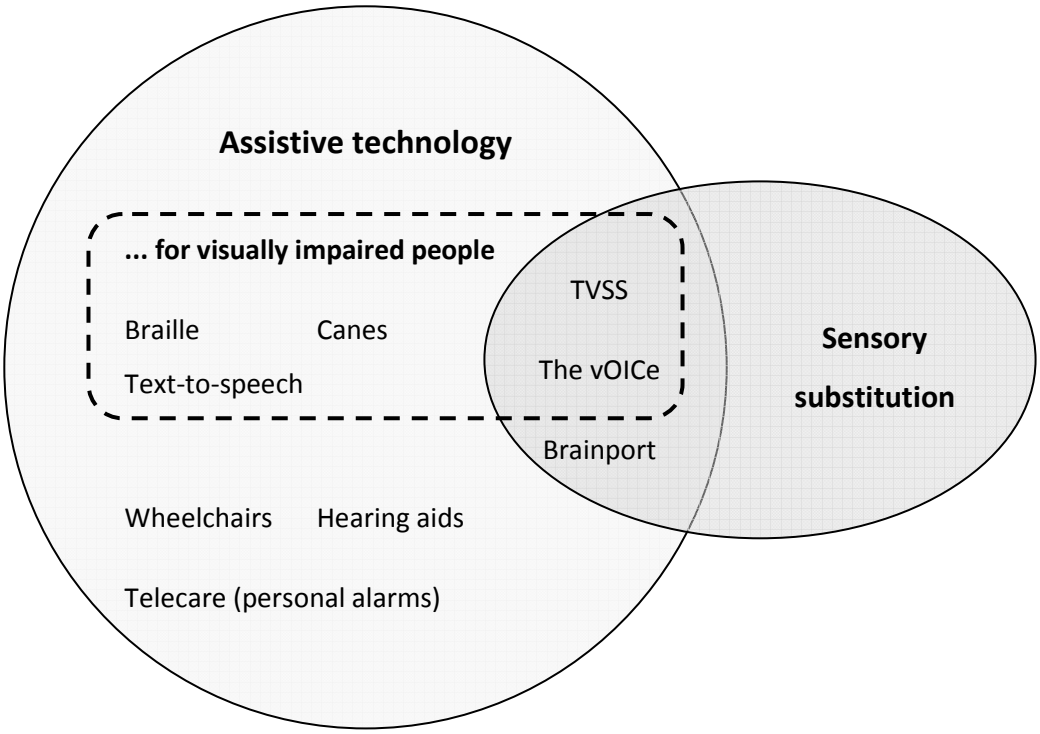


Figure 3: Diagram of the relationship between assistive technology and sensory substitution

"I was intuitively aware of the direction of my home or of my office. For example, I would wait in line in the cafeteria and spontaneously think: I'm living over there."

- A participant on wearing the FeelSpace device (Nagel et al., 2005)

Another example of sensory augmentation also involves magnetism. That similarity aside though, they could not be more different. Whereas FeelSpace uses complicated electronics to communicate the orientation of the wearer relative to the earth, some researchers are implanting magnets into their fingertips in order to detect man-made magnetic fields (Hameed, Harrison, Gasson, & Warwick, 2010). Though the functional applications are hard to identify, these (and other; e.g. Berg, 2012) researchers report some fascinating experiences.

Sensory augmentation is a theme picked up in a couple of chapters of this thesis. Firstly, the discovery of a multisensory facilitatory effect presented in chapter four supports the idea that sensory augmentation devices could become deeply integrated with existing senses. Secondly, sensory augmentation devices are considered among other, related devices in chapter five.

1.2.5 Summary

Though the sensory substitution devices described above are predominantly concerned with substituting vision, there is nothing inherent in the concept of sensory substitution that gives primacy to sight. Indeed, the fact that so many approaches to sensory substitution appear to work makes it seem likely that we have barely scratched the surface so far.

The relationship between assistive technology and sensory substitution devices is also a strange one. Again, all the devices described so far are (by virtue of having been designed for visually impaired people or people with bilateral vestibular loss) assistive technologies. Indeed, a sensory substitution device will normally be assistive technology. Yet this is not a necessary property of a sensory substitution device. Assistive technology is described by its (intended) use. A visual sensory substitution device used by a sighted, blindfolded person is therefore not assistive technology. This relationship is illustrated in Figure 3. Both the relationship between senses and sensory substitution and between assistive technology and sensory substitution devices are considered in greater detail in chapter five.

Despite the fact that most sensory substitution devices are ostensibly designed for visually impaired people, adoption in among those with a visual impairment is very low. An active mailing list community shows that there are active users of the vOICe, but no reliable estimates exist for the overall adoption of sensory substitution by visually impaired people. Anecdotally, except for those sought out because of their use of a sensory substitution device, none of the visually impaired people I spoke to during my research had heard of sensory substitution. The word may spread as the technology continues to mature, but for now it is clear that the primary users of sensory substitution devices are researchers.

1.3 Key findings and discussion

It is not surprising that sensory substitution devices are of great interest to the research community. Ignoring for a moment the utility of a functioning sensory substitution device for visually impaired people, sensory substitution is also a tool with which we can explore the mind, brain and the processes by which our perceptions of the world come to be. Research into sensory substitution may be divided roughly into the behavioural, the neurological and the philosophical.

1.3.1 Behavioural experiments

The behavioural study of sensory substitution is as old as the first sensory substitution device. As mentioned above, Bach-y-Rita's original paper includes a description of the first experiment performed using the TVSS. In this experiment, six participants (with almost no prior visual experience) were trained for between 20 and 40 hours firstly using geometric shapes, but later images of physical objects. Impressively, the authors report that during the experiment the participants begin to discover attributes of vision such as perspective and occlusion (Bach-y-Rita et al., 1969).

A subsequent paper reports experiments in which a mix group of blind and sighted participants (with varying levels of training on the TVSS) is tested on a range of basic measures (White et al., 1970). When asked to identify a geometric shape, participants began at chance, but eventually reached 100% accuracy and latencies of under a second. When asked to judge in which direction a checkerboard pattern was slanted, they found that their blind participants were more accurate than the sighted sensory substitution device users. This paper introduces the concept of visual acuity to sensory substitution, but does not use common measures of visual acuity as some later studies do. It is also in this paper that the importance of the user having agency over the device is first stressed.

These initial experiments presciently set the stage for many of the exploratory avenues to follow. Object identification and localisation are common themes throughout the behavioural experiments using sensory substitution devices. The popularity of these paradigms is perhaps in part due to their combination of simplicity and basis in everyday tasks (sensory substitution devices are, after all, intended to be used by visually impaired people). A large number of studies in this style have been conducted using the vOICe. Some of the key findings include:

- Participants describe their sensory substitution device use as more visual when performing more spatial tasks (e.g. localisation) and more auditory when attempting to identify objects (Auvray, Hanneton, & O'Regan, 2007). It is possible that this differential finding is the result of matching objects based purely on their auditory features.
- Participants may still perform above chance without prior training, if they have had the conversion algorithm explained to them (Kim & Zatorre, 2008).
- Using a head-mounted camera is preferable during localisation tasks, but using a handheld camera is preferable when seeking to identify an object (Brown et al., 2011).

Measuring the visual acuity afforded by sensory substitution devices is another common trope. Using the TDU and a common ophthalmic test known as the "Snellen tumbling E" (wherein participants must judge the orientation of a rotated letter E), naive participants were found to be capable of an acuity of 20/860, but with training were capable of doubling this to 20/430 (Sampaio, Maris, & Bach-y-Rita, 2001). As mentioned above, the WHO definition of blindness is 20/400. Using the same task with the vOICe, seven of the nine early blind participants were able to achieve visual acuities of 20/200 (the WHO blindness threshold) or better (Striem-Amit, Guendelman, & Amedi, 2012). In a similar study, 20/480 was found to be the upper limit at which participants maintained above-chance performance (Haigh et al., 2013). These

experiments are hard to compare because of the extent to which their participants, their training and procedures differ. It is clear though, that although no training is required to get started with a sensory substitution device, it certainly improves performance. It is disappointing, but perhaps not unexpected, that the best sensory substitution device users are still unlikely to display acuities better than the threshold for being classified as blind. It remains possible however, that long term sensory substitution device users may perform better than the participants who took part in these experiments.

A related style of experiment is the identification of two-dimensional figures. The SmartSight system was evaluated partly by participants to identify geometric shapes and partly by their ability to recreate visual scenes (Cronly–Dillon, Persaud, & Blore, 2000; Cronly-Dillon et al., 1999). A shape discrimination task was also used as a component of an assessment of the vOICe (Brown et al., 2011).

Another popular form of assessing the behavioural performance of sensory substitution device users is the obstacle course. This approach has been used to good effect in the evaluation of the See ColOR and ENVS systems (Bologna et al., 2010; Meers & Ward, 2004). In both cases, the choice to conduct these navigation tasks appears to have been driven by the desire to demonstrate the benefits of colour and depth information in a highly ecologically valid setting. The conference paper presenting See ColOR device, for instance, shows a blindfolded participant avoiding parked cars.

Comparable performance among naive users of the Vibe in a U-shaped car park suggests that colour and depth are not essential cues for mobility (Durette et al., 2008). The obstacle course paradigm is also one of the components of the Brown *et al.* study using the vOICe. Both studies show a significant learning effect. The study using the vOICe also compared head-mounted to handheld camera position, but in this task found no difference with regard to either speed or accuracy. Given that a key anticipated application of sensory substitution devices is mobility, asking participants to navigate through a maze or around obstacles, is a highly ecologically valid test to perform. Unfortunately, it is also difficult and time consuming to administer.

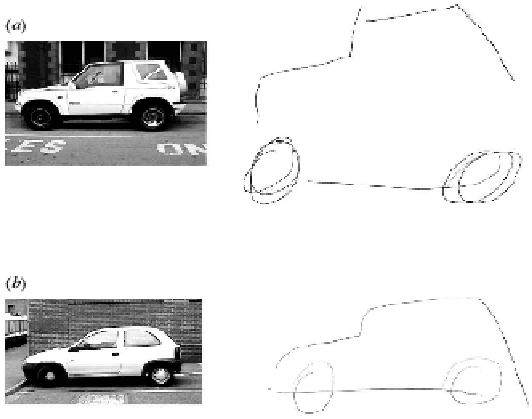


Figure 4: Drawings by blind users of the SmartSight system (adapted from Cronly-Dillon *et al.*, 2000)

A few, somewhat more esoteric, approaches to demonstrating sensory substitution devices provide some flavour to our understanding. Mentioned above, the use of the SmartSight system in the reproduction of visual scenes by visually impaired people (see Figure 4) illustrates nicely the fact that sensory substitution devices are treated fundamentally as sources of visual information. In a somewhat different vein, the demonstration of the efficacy of visual illusions delivered via the PSVA further demonstrates this point (Renier, Laloyaux, et al., 2005).

The key message from all of these behavioural studies is that sensory substitution devices work. Even though the visual acuity they afford is not sufficient to trigger the reclassification of their visual impaired users, sensory substitution devices allow for accurate object identification, object localisation, two-dimensional figure discrimination, and navigation in both realistic and contrived settings. An addition interest finding is that explicitly explaining the way in which sensory substitution devices operate is not a necessary precursor to their successful use, but training does quickly improve performance.

1.3.2 Neurological

The neurological foundations of sensory substitution have long fascinated researchers. How does the brain process this information? Which areas of the brain are responsible? Does training with a sensory substitution device lead to any neuroanatomical changes? The first study that investigated these questions was a Positron Emission Tomography (PET) study using geometric figures and the PSVA. Interestingly, they found that sensory substitution device use resulted in greater activity in the occipital lobes of early-blind participants compared to sighted controls (Arno, De Volder, et al., 2001). Moreover, despite low level of activation from general auditory stimuli, activity in the occipital lobe was greater when performing sensory substitution tasks.

A subsequent study using PET and the PSVA investigated the effects of depth perception. Again, regions of the brain associated with vision were found to be activated (Renier, Collignon, et al., 2005). A third study applied repetitive Transcranial Magnetic Stimulation

(rTMS) over the region activated in the first study. This resulted in diminished performance in the sensory substitution task for the blind, but not the sighted, participants (Collignon, Lassonde, Lepore, Bastien, & Veraart, 2007).

Contemporaneously, PET studies were also being carried out using the TDU. Six early-blind participants and six matched controls undertook a "tumbling T" test. During an intensive seven day training period, both the blind and sighted participants became faster and more accurate. Despite this, only the blind participants showed any activation of the visual cortex (Ptito, Moesgaard, Gjedde, & Kupers, 2005). Interestingly, the same researchers later found that applying TMS over the visual cortex after training with the TDU resulted in tactile sensations on the tongue for four of their blind participants (Kupers et al., 2006).

Similarly, functional magnetic resonance imaging (fMRI) research shows that processing shape information present in auditory information from the vOICe results in activation of the Lateral-Occipital tactile-visual (LOtv) area in both sighted and blind participants (Amedi et al., 2007). This is significant because the LOtv is strongly associated with identifying objects by vision and touch, but not with any auditory processes. A later study demonstrated that this activation is functional (i.e. not epiphenomenal) by showing that successful sensory substitution device use can be disrupted by the application of TMS to this region (Merabet et al., 2009).

Taken together, this data has two strands. Firstly, the differential activations of the visual cortex demonstrated in early-blind participants suggest a degree of plasticity caused by a sustained lack of visual input. Secondly, the fact that some areas of the brain are recruited for sensory substitution tasks in both sighted and blind participants suggests that some areas of the brain previously thought of as "visual" are task-based rather than modality-based.

1.3.3 Philosophical

Embedded in most of the studies described above, there is some attempt to shed light on what sensory substitution *really is*. Is the information treated more like the substituting modality, or the substituted modality? Perhaps it all just feeds into a sense of space. Or does it represent something new entirely? Inevitably, these discussions are entwined and interwoven with debates around the definitions and boundaries of our more conventional sensory modalities. In a review written with my supervisor, we consider some of these issues in the context of synaesthesia (Ward & Wright, 2014). The descriptions below summarise the various criteria for classifying senses that we introduced in that paper.

Bach-y-Rita was of the opinion that senses should be defined by their function: “If a subject without functioning eyes can perceive detailed information in space, correctly localise it subjectively, and respond to it in a manner comparable to the response of a normally sighted person, I feel justified in applying the term ‘vision’” (Bach-y-Rita, 1972). This view is not universally accepted.

At the other end of the spectrum from Bach-y-Rita, there are those who assert that senses are defined by the sensor (Keeley, 2002). In this model, vision can only be vision if it enters through the eyes. This position mirrors Müller’s “law of specific nerve energies”, which states that the effect of stimulating a sensory system is dependent on the system being stimulated, rather than the mode of stimulation (Müller, 1826; cited in Norrsell, Finger, & Lajonchere, 1999).

A related view is that senses are defined by the region of the brain they activate. This approach is rarely stated explicitly, but often lurks behind the reasoning of studies that use neuroimaging techniques. If this is true, the sensory nature of sensory substitution device use depends entirely on the balance of evidence given in the previous section.

There are two, more nuanced, approaches which rely on phenomenology and the sensorimotor account respectively. The phenomenological approach to classifying the senses involves self-report. A good phenomenological account of the vOICe strongly supports its classification as visual (Ward & Meijer, 2010), but accounts by early-blind individuals inevitably present difficulties (Guarniero, 1974). The sensorimotor account states that the senses are defined by the rules that govern how our actions lead to changes in sensory experience (the “sensorimotor contingencies” of a sense). Vision, for example, has the property of occlusion, whereby placing an object behind will render it hidden. This does not happen with sound. Accordingly, if a sensory substitution device demonstrates this rule, we might conclude that it is behaving in a visual way (see O’Regan & Noë, 2001a).

Finally, there are some who suggest that accurate use of a sensory substitution device can be explained by spatial perception, without any reference to vision (Block, 2003). Sensory experiences associated with vision (e.g. brightness, colour) pose problems for this explanation, but these are all but impossible to probe objectively.

The more theoretical aspects of this thesis (particularly chapter five) deliberately avoid this debate, preferring instead to focus on the relationship between sensory substitution devices and other sensory tools.

1.4 The present thesis

The chief motivation behind each of the chapters that follow is the production of more useful sensory substitution devices. That isn't to say that theoretical concerns will henceforth be disregarded: the theories underpinning sensory substitution are inescapably intertwined with the practicalities of producing sensory substitution devices.

The first empirical chapter (chapter two) demonstrates the use of interactive genetic algorithms to determine an optimal set of parameters for a sensory substitution device based on the vOICe. Interactive genetic algorithms were used because they allow a large, highly dimensional problem space to be searched without needing to do so exhaustively (which in this case would involve testing 65,536 combinations for the parameters under test). In order to generate novel sensory substitution devices during the course of an experiment, a configurable device ("Polyglot") was created.

As well as introducing the technique and the tools, this chapter also presents data from three interactive genetic algorithm based experiments that use this new sensory substitution device. The first of these is designed to evolve a sensory substitution device which sounds more aesthetically pleasing. The second aimed to evolve the most intuitive sensory substitution device. The third set out to evolve a sensory substitution device which best facilitated discrimination between similar images. For practical purposes (and in line with other research in the field) these experiments, and those in the other chapters of this thesis, use sighted participants.

Combining the results of each of these experiments leads to a recommended set of parameters for future vOICe-like sensory substitution devices. In homage to the vOICe, the new sensory substitution device created to implement these recommendations was named "Vox" (i.e. Polyglot is "many vOICes" and Vox represents a single selection of these vOICe variants). This chapter was previously published in the Quarterly Journal of Experimental Psychology (Wright & Ward, 2013).

The use of interactive genetic algorithms in chapter two is fundamentally about exploring the landscape of possible sensory substitution devices. Chapter three radically expands on this theme by introducing a general purpose, modular framework for developing visual-to-auditory sensory substitution devices. This framework is the fuller realisation of the Polyglot device introduced in the first chapter and is based on the principle of End-User Development (EUD). By breaking up the sensory substitution device into four types of module, it has been possible

to create a mix-and-match system for combining aspects of functionality. Three of these modules map onto the three functional units that have traditionally been part of the general description of this class of device (this notion is expanded upon in chapter 6 and is referred to throughout as the "tripartite model"): there are acquisition modules for capturing images, output modules for generating sounds and transformation modules for coupling them together. Additionally, some of the spatial behaviour of existing sensory substitution devices (e.g. scanning in the vOICe) is abstracted to a fourth type of module ("pointer" modules), which the transformation modules can use to select regions on the visual scene.

Whereas the version described in chapter two allows for parameters to be dynamically configured, this new version of Polyglot has three levels of customisability. At the most superficial level, each module may have options that alter some aspect of their operation. Next is the composition of a sensory substitution device by selecting which modules to use. Third, Polyglot also contains the support structures necessary for the creation of entirely new modules. The Polyglot framework has been released as open-source software (Wright, 2013).

In chapter four, a novel method of evaluating sensory substitution devices using eye-tracking is introduced. This technique allows for the interaction between vision, audition and visual attention to be investigated. Does the presence of partial visual stimuli assist in a judgement using an auditory sensory substitution device? Can a sensory substitution device drive the attention of natural vision?

In this paradigm, participants are simultaneously presented with a visual scene and the sonification of the same scene with an embedded visual target. (In the control condition, participants hear the scene and with the target, but do not see anything.) Participants then seek to localise the target to the correct quadrant of the scene. Meanwhile, an eye-tracker is recording the position of their gaze on the visual scene. As expected, participants are able to make judgement correctly at rates better than chance and this accuracy improves over time. More interestingly, the data shows both that the co-presentation of visual stimuli assisted localisation and that gaze predicted the target location more reliably than the behavioural responses. Preliminary results from this study were presented as a poster at the 13th annual meeting of the International Multisensory Research Forum (Wright, Ward, Simonon, & Margolis, 2012).

Chapter five explores the relationship between sensory substitution devices and the wider context of "sensory tools". The label "sensory tool" is a novel one, which we define as "a

device whose primary function is to manipulate rich, abstract sensory information and present the product to a user who retains agency over the sensory experience." Examples of sensory tools consequently include telescopes, cochlear implants, and subdermal magnetic implants.

By treating sensory substitution devices as sensory tools, it is possible to compare and contrast them with systematically with other, more conventional sensory tools. To facilitate this, a taxonomy of sensory tools is proposed. In this taxonomy, sensory substitution device are renamed as between-sense referral devices and are placed alongside compensatory prostheses (e.g. reading glasses), within-sense referral devices (e.g. long canes), and novel-sense referral devices (i.e. sensory augmentation devices). Further, each of these categories may be subdivided into examples of direct and mediated sensory tools. The motivation behind this proposal is to provide a set of solid foundations for future sensory substitution research. This work has been submitted for inclusion to the Proceedings of the British Academy.

Finally, in chapter six data is presented from an experiment exploring the existence of a crossmodal correspondence between hue and pitch. The results from chapter two suggest that crossmodal correspondences can be useful scaffolding whilst learning to use a sensory substitution device. The existence of a crossmodal correspondence between hue and pitch would therefore be a useful asset if one were producing a colour sensory substitution device. Although the data were not consequently helpful from the perspective of a sensory substitution device designer, they do hint at an interesting phenomenon: rather than a direct relationship between hue and pitch, the data seem to suggest a relationship mediated by the geometry of the range of visible colours.

1.5 General discussion

The following chapters represent a multi-disciplinary cross-section of sensory substitution research. Taken together, they reflect the nature of the field. Sensory substitution is still a young one phenomenon and there is lots left to be discovered.

In concrete terms, this thesis represents the contribution of an improved version of the vOICe (Vox; chapter two) and a framework for the rapid development and configuration of novel sensory substitution devices (Polyglot; chapter three). Chapters four and six represent novel empirical findings which should be helpful to those involved in sensory substitution and further afield. Chapter five is a contribution to the ongoing conversation about the nature of sensory substitution and offers a conceptual framework for comparative evaluation.

A key feature of the thesis as a whole is that more questions are raised than are answered. (This author believes that this is a fundamentally positive thing.) In terms of taking the work forward, there are two key areas for potential development: building on the technical contributions, and further exploring the empirical findings.

1.5.1 Technical contributions

The most obvious example of the former is the Polyglot framework. This has the potential to massively accelerate sensory substitution device design and enhance their therapeutic significance. By allowing common components (e.g. webcam modules) to be reused, designers can focus on just the new feature that exemplifies their interesting new idea.

Because of the introduction of a standard file format for storing configuration parameters, Polyglot is highly suited towards guided discovery, whereby a mobility officer (or other healthcare professional) helps a visually impaired user to select the most suitable configuration. Because the core logic is separated from the composition logic, the visually impaired person in this scenario would be able to run a stripped down version of the Polyglot engine, which would reduce the chance of accidentally changing the settings. A stripped down Polyglot engine would also be the best option for a mobile version.

Polyglot therefore affords opportunities both in its development and in its application. At the most basic level, the code for Polyglot has already been released as open-source: interested researchers can download the code and improve it, as long as they contribute their improvements back to the community. As mentioned above, it is anticipated that most developers will contribute modules, rather than alter the core framework.

A core feature of Polyglot is the application of the principles of End-User development. It is here that the opportunities for development are richest. A specially adapted Integrated Development Environment (IDE) or Software Shaping Workshops (Costabile, Fogli, Fresta, Mussio, & Piccinno, 2003) might make it easier for users to become developers. This takes on special significance when one considers that some visually impaired person might not be able to use standard IDEs because of problems with accessibility.

As well as Polyglot itself, there are a few other innovations in the code-base which may be of benefit to researchers in the future. The first of these is a general purpose library for creating localised tones in the C# programming language. This is an extension of the popular NAudio library and utilises natural localisation cues such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD). Using this code is as simple as providing a pitch (in Hz), an

amplitude (between zero and one), and an azimuth (from minus one on the left to plus one on the right).

The second key innovation is the use an Android device as a positional controller with haptic feedback. The Android application which makes this possible has also been released as open-source code and the Polyglot project features an example implementation of a desktop application with which the Android application can interface. The Android application is capable of using either the standard Android vibration functions, or the high fidelity vibration available in devices certified by Immersion Technologies.

1.5.2 Empirical findings

Two of the key empirical findings in this thesis pose interesting questions worthy of future exploration. The first, that natural vision interacts bi-directionally with substituted vision, is perhaps the more directly applicable to sensory substitution research. Future work could explore the distinction between reflexive and volitional saccades – can this be distinguished with EEG and are reflexive saccades better at predicting the location of the target in the substituted signal? The most obvious applications of this finding are in sensory augmentation and in sensory substitution devices for those with partial vision (i.e. due to a scotoma). It would therefore be highly interesting to test the findings with a sensory augmentation device. The findings predict that a partially sighted user would be advantaged as their remaining vision meshes with the substituted vision.

The second key empirical finding is that of a relationship between hue and pitch that appears to be mediated by innate familiarity with the geometry of visible colour space. For starters, this is a bold claim which would benefit enormously from independent verification. Once the basic effect is confirmed, it would be very interesting to explore how it interacts with other known cross-modal correspondences. We know that brightness, for example, maps to both loudness and pitch (Marks et al., 1987), but there is not currently a full model of visual-to-auditory cross-modal correspondences. Taken to the logical extreme, it would be fascinating to fully quantify these relationships, such that the model would be able to predict the most appropriate sound for a given visual stimuli.

1.5.3 Challenges for sensory substitution

The work cited in this chapter is a testament to the fact that sensory substitution devices are fascinating. They are hugely valuable tools for research, but are currently underused by their target audience. For sensory substitution devices to achieve their therapeutic potential however, sensory substitution research must undergo a step-change. Sensory substitution

devices are still unknown by the vast majority of visually impaired people. That so many researchers are using sensory substitution devices is encouraging, but will count for little if current practices are not developed upon.

Current rates of sensory substitution device usage represent a missed opportunity. It is possible that most visually impaired people would not want to regularly use an sensory substitution device (a lack of empirical data makes this hard to gauge), but they should at least be made aware of the technology in order to make an informed decision.

The worst instance of this is arguably in developing countries – home to 90% of global visually impaired people (Thylefors, 1998). A simple visual-to-auditory sensory substitution device does not require any specialised equipment and should therefore be very easy to distribute. A computer with a webcam or a modern mobile phone is the only hardware required. The business of sensory substitution is handled by software, which is easily distributed over the internet. Despite the persistence of inequalities (Petrazzini & Kibati, 1999), more and more people from developing countries have access to computers, smart phones, and the internet (Chinn & Fairlie, 2010; Gitau, Marsden, & Donner, 2010). With support from international health organisations, sensory substitution devices could quickly become a normal part of treating visual impairments.

Fundamentally, if sensory substitution devices are to become more than just academic curiosities, we must take their production beyond a cottage industry. Chapters two and three of this thesis are suggest approaches that would help this process of professionalization, but this thesis does not discuss the social or policy aspects of sensory substitution device adoption. Studies that would be useful include:

1. an audit of current sensory substitution device usage
2. in-depth analysis of how these people use them
3. a survey of awareness amongst healthcare professionals
4. an evaluation of the best methods of training visually impaired people to make use of sensory substitution devices

This research could potentially all feed into global promotion backed by informed and well-trained professionals.

1.5.4 Looking ahead

Beyond straightforward application of sensory substitution devices, there are several other ways in which the field may be of practical benefit. Though perhaps unlikely comparison, a

precedent is being set by robotic exoskeleton systems. For fans of science fiction, these systems will immediately bring military application to mind. The reality however, is that the first practical exoskeleton systems are being used to help with personal mobility (for elderly or disabled people) and subsequent models are likely to be designed for those who routinely handle large objects (e.g. construction workers and furniture movers). What makes sensory substitution comparable with these exoskeleton systems is their generality: nothing inherent in either concept defines a particular way of using them. Though the applications may be less immediately obvious, it is likely that the significance of sensory substitution technology will grow both within the visual impairment sector and beyond.

One intriguing use of sensory substitution technology would be to act as a backdrop for other forms of assistive technology for visually impaired people. In particular, there is a need for a mechanism for screen-readers (computer based text-to-speech systems) to convey the overall "shape" of a document. Because screen-readers treat the text of a document as a one-dimension string, many of the visual cues used in reading are lost (Yesilada, Harper, Goble, & Stevens, 2004). It is impossible, for instance, to gauge how much of a paragraph remains using a screen-reader. Additionally, though many systems are able to announce elements such as headings or bullet points, these mechanisms are often reliant on the adherence by the software (or website) creator to coding conventions and guidelines. A screen-reader that also sonified the shape of the document would benefit the user by giving presenting exactly the same information as is available to a sighted person, whilst at the same time not requiring the user to extract graphemes from the sonification. As a bonus, figures and illustrations would be more accessible too.

In a similar vein, sensory substitution technology could improve the usefulness of navigation devices. Sonifying maps would be beneficial both to visually impaired people and sighted people. In the case of the former, there are already specially adapted personal GPS units, which are designed to assist with navigation whilst on foot. By sonifying their maps, a visually impaired person could gain a better understanding of their surroundings than afforded by instructions delivered by a synthesised voice. These sonified maps could also be explored offline to build confidence prior to the start of a journey. For sighted people, the benefits come more from the need to keep eyes on the road whilst driving. This need is acknowledged in the work of Moulster & Stockman, who devised a method for communicating upcoming turns using a code comprised of paired tones (Moulster & Stockman, 2011). This could be taken further if the principals of sensory substitution were applied to the visual output of an

in-car GPS system: the driver would be able to gain a rich sense of the route ahead without needing to look down at a screen or decode a series of tones.

Finally, there are the many cases in which expanding a person's visual field (spatially, or to incorporate additional information) might be of enormous benefit. This form of sensory augmentation is alluded to in various places throughout this thesis (particularly in chapter four). As well as assisting people with partial visual fields (such as might be caused by scotoma or hemianopsia) or with attentional disorders (e.g. hemispatial neglect), it is possible that sensory augmentation devices will find use among rescue workers, remote operators and pilots. Rescue workers may benefit, for example, from a system that could superimpose the signal from a thermal imaging device over their natural vision. Indeed, there already exists a patent for such a device (Havey, Gibson, Seifert, & Kalpin, 2007). A remote operator (or machinery or a vehicle, for instance) might benefit from being able to monitor more than one visual feed simultaneously. This may be made possible by presenting the focal region visually and sonifying a composite of the other visual feeds. Finally, pilots may wish for certain information from sensors to be made intuitively available in a form meshed with the view from the windscreen. The output of the weather radar and the virtual horizon might both be candidates for this form of representation. This information could be presented using an augmented reality heads-up display system (Milgram, Takemura, Utsumi, & Kishino, 1995), but there may be benefits to offloading some to another (less overloaded) sense.

1.5.5 Conclusion

In short, sensory substitution research is still in its infancy. Though the last fifty years have seen huge advances, the field still presents a range of fascinating technological opportunities and still poses a number of interesting theoretical questions for the research community. Moreover, for sensory substitution technology to reach its full potential among both visually impaired it must attract the interest of healthcare professionals and policy makers. The research presented in this thesis gives tantalising glimpses of what may be possible in the future and also offers suggestions on how these ideas may be carried to fruition.

Chapter 2: The evolution of a visual-to-auditory sensory substitution device using interactive genetic algorithms

2.1 Abstract

Sensory substitution is a promising technique for mitigating the loss of a sensory modality. Sensory substitution devices (SSDs) work by converting information from the impaired sense (e.g., vision) into another, intact sense (e.g., audition). However, there are a potentially infinite number of ways of converting images into sounds, and it is important that the conversion takes into account the limits of human perception and other user-related factors (e.g., whether the sounds are pleasant to listen to). The device explored here is termed "polyglot" because it generates a very large set of solutions. Specifically, we adapt a procedure that has been in widespread use in the design of technology but has rarely been used as a tool to explore perception—namely, interactive genetic algorithms. In this procedure, a very large range of potential sensory substitution devices can be explored by creating a set of "genes" with different allelic variants (e.g., different ways of translating luminance into loudness). The most successful devices are then "bred" together, and we statistically explore the characteristics of the selected-for traits after multiple generations. The aim of the present study is to produce design guidelines for a better SSD. In three experiments, we vary the way that the fitness of the device is computed: by asking the user to rate the auditory aesthetics of different devices (Experiment 1), and by measuring the ability of participants to match sounds to images (Experiment 2) and the ability to perceptually discriminate between two sounds derived from similar images (Experiment 3). In each case, the traits selected for by the genetic algorithm represent the ideal SSD for that task. Taken together, these traits can guide the design of a better SSD.

This chapter was previously published in the Quarterly Journal of Experimental Psychology (Wright & Ward, 2013).

Sensory substitution is a process in which information from one sensory modality is represented in another modality, the most common application being visual impairment, with vision represented in either sound or skin-based stimulation (mechanical or electrical). Sensory substitution is enacted by a sensory substitution device (SSD): a system composed of a sensor (e.g., a camera), a coupling process (the software), and a stimulator (e.g., headphones, vibrotactile array). Within a few hours of training, novice participants have some ability to localize and recognize objects (Auvray, Hanneton, & O'Regan, 2007; Brown, Macpherson, & Ward, 2011) and generalize to new objects (Kim & Zatorre, 2008). Expert blind users recruit "visual" cortices to process the substituted sense (Amedi et al., 2007; Merabet et al., 2009; Poirier, De Volder, Tranduy, & Scheiber, 2007; cf. Pollok, Schnitzler, Stoerig, Mierdorf, & Schnitzler, 2005). Users may report visual phenomenology to sounds or touch (Ward & Meijer, 2010) and have been shown to be susceptible to visual illusions delivered via a substituting sense (Renier et al., 2005). Despite these impressive findings, there remains a lack of knowledge concerning how visual images should be converted into sounds to enable efficient perception and learning. Here we present an original approach to this problem that could be an important tool for perception research itself, outside of the more limited domain of sensory substitution.

2.2 Sensory substitution

Tactile-based systems continue the tradition of Bach-y-Rita and his original tactile vision sensory substitution (TVSS) device (Bach-y-Rita et al., 1969), which acted on the skin of the back. More recent tactile systems have used a fingertip (Kaczmarek, Tyler, & Bach-y-Rita, 1997) and the tongue (Bach-y-Rita et al., 1998). In all these tactile systems, pixel position is mapped to stimulator position, and luminosity is mapped to the intensity of the stimulation.

Despite the lack of an immediately obvious set of mappings, auditory SSDs also share a common set of basic relationships: Vertical position tends to be represented by sound frequency, and luminosity tends to be represented by sound amplitude. This basic assumption is grounded by experimental research suggesting that, in sighted people at least, there is a tendency for pitch and vertical position to interact (e.g. Ben-Artzi & Marks, 1995) and similarly for loudness and luminance (e.g. Marks et al., 1987). However, a significant challenge for auditory devices is the representation of space because spatial resolution is generally considered to be poorer in the auditory domain than in vision or touch. One device, the Vibe, is similar to the tactile systems in that it presents the whole field of view at once and relies on the natural localization abilities of the ear by expressing horizontal (left/right) position by

controlling the relative amplitude in each ear (Auvray, Hanneton, Lenay, & O'Regan, 2005). The vOICe (which forms the basis for this study and is described in more detail below) encodes horizontal position temporally—that is, the image is heard piecemeal over time. The PSVA (prosthesis substituting vision for audition) uses pitch to encode position in both the horizontal and the vertical axes. The PSVA also implements a bias inspired by the foveal region of the human eye, which dedicates more "space" to pixels in the centre of the field of view (Arno, Capelle, Wanet-Defalque, Catalan-Ahumada, & Veraart, 1999). An alternative approach is to sonify only those pixels that represent edges rather than surfaces. This occurs in the SmartSight system in which the user is presented with the a sound generated from the pattern of extracted visual features in a scene (Cronly-Dillon et al., 1999).

Given success in the laboratory and the potential therapeutic benefits, one may wonder why there are so few users of these devices in the real world. There are likely to be many reasons for this including lack of information: costs (particularly true for tactile devices), and the time it can take to become an expert user (seemingly more the case for auditory devices). With regard to the latter, one may be able to develop better conversion algorithms that are more intuitive to use because they are optimized with respect to human perceptual abilities. Whilst one can base a judgement on known properties of the auditory system, sounds derived from images will have special properties compared to naturally occurring sounds. This occurs because images have certain regularities (e.g., light tends to come from above; spatial smoothness—the intensity of a pixel tends to correlate with that of its neighbours). These regularities would then become a property of these particular sounds, but would not be a meaningful property of sounds in general. As such, there is a need for research to determine the optimal solution for converting images into sounds. The problem for conventional approaches is that the number of potential conversion algorithms to explore is huge. In the experiments described below, we consider a problem space of 65,536 conversion algorithms ($4 \times 4 \times 4 \times 4 \times 2 \times 2 \times 8 \times 8$, or 216) in our "polyglot" device. Clearly, a conventional approach is not feasible: We could not test each unique condition over multiple participants, and nor would we find it easy to interpret the eight-way interaction generated by the analysis of variance (ANOVA). An alternative way of approaching this involves the use of interactive genetic algorithms.

2.3 Genetic algorithms and interactive genetic algorithms

Genetic algorithms (GAs) are an established method for rapidly approximating an optimal solution from within a large, highly dimensional search space. As implied by their name,

genetic algorithms are inspired by the way in which nature has (over many generations of incremental change) produced organisms that are highly adapted to exist in a particular ecological niche. The fundamental components of a genetic algorithm are “genomes”, which describe an individual member “organism” of a “population”. The basic process is that each genome is assessed. Depending on how well it performs, each genome may be used as a starting point for a new batch (“generation”) of genomes. The genomes of this new generation are also subsequently evaluated. This cycle continues until either an adequate solution is found or after a predetermined number of generations. Though they have not yet become a mainstream technique in psychology, the usefulness of GAs in other fields is firmly established. Examples of their success can be found in areas as diverse as 2D packing (Hopper & Turton, 1999), protein folding simulations (Unger & Moult, 1993), and jazz improvisation (Biles, 1994). For a more detailed account of genetic algorithms, see Haupt and Haupt (2004).

In order to be solved using a genetic algorithm, a problem space must first be formalized as a genome. The most simple form of genome is a string of binary digits (“bits”), where the simplest gene is a single binary digit (i.e., 1 or 0). To take a simple example, a single bit could be used to code whether a light bulb is switched on. Multiple bits can be combined to represent more complex aspects; for example, if we had three coloured light bulbs, we could use three bits to represent any of eight (2^3) colours.

Once the problem space has been mapped to genomic data, an initial generation of randomly generated genomes are tested to obtain a “fitness” score for each genome. The score of each genome is determined by a “fitness function”. In our coloured lights example, the fitness function could be the proximity to a target colour. In the sensory substitution domain, a fitness function could be the participants' ability to hear certain aspects of a sonified image. These fitness scores are used as the basis for “selection”, which is the primary genetic operator used to produce the next generation of genomes. The specific mechanism used to drive selection can vary, but the present study utilized a popular method known as weighted-stochastic selection. Essentially, the higher the fitness score then the greater the probability that it will be selected to “mate” and, hence, the greater probability that those traits will be inherited by the second generation. Other forms of selection, such as tournament selection, operate in a broadly similar fashion. In experiments using a greater number of generations, tournament selection may be more appropriate (Blickle & Thiele, 1996).

After selection, the genomes are copied a digit at a time to the next generation. At this point, two genetic operators come into effect. The first is “crossover” and requires the selection

process (described above) to choose two parent genomes from the previous generation for each new genome. The new genome is generated by copying from the old genomes one bit at a time. The first genome to have been selected is active and will be copied from, but each copying operation carries with it a possibility (the crossover rate) that the active genome will switch. Crossover is equivalent to organic mating. The second operation is “spot-mutation” and is essentially the (much smaller) chance that one of the digits will be changed. In the case of a binary digit genome, the change can only be an inversion. Crossover and spot-mutations are both important to ensure that the solutions converge towards an end result, but not at the expense of getting stuck in “local optima”. Optionally, the fittest genomes can be progressed from one generation to the next without modification—this is called “elitism” in the GA literature. This process of evaluating, selecting, and recombining the genomes is cycled for either a predetermined number of generations or until a predefined “stopping condition” is met. Once this is finished, the fittest genomes should represent good approximations for optimal solutions.

Interactive genetic algorithms (IGAs) are a subset of genetic algorithms whose fitness function incorporates a response from a human participant. IGAs have enjoyed success in a wide variety of disciplines. The above example of jazz improvisation is a good example of this. Music lends itself to formal representation, and computers are able to generate melodies, but they cannot evaluate what makes a jazz solo great (Biles, 1994). Other examples include computer graphics and animations (Sims, 1991) as well as architecture and product design (Soddu, 2002). IGAs have also made inroads into psychological research and have helped, for example, determine an “idealized” female face (Johnston & Franklin, 1993).

2.4 The present study

The sensory substitution algorithms that we explore can be considered as variants of “the vOICe” (Meijer, 1992), or, rather, “the vOICe” (the capitalized letters phonetically spelling “Oh, I see!”) can be construed as existing on the dimensions used within our problem space. This system has been widely studied by researchers. These rules underlying the vOICe were derived by reversing the transformation applied in the generation of a spectrograph. The resulting sound is referred to as a “soundscape”.

To explore which parameters, if any, could be improved upon, we re-implemented the vOICe so that every aspect of its operation could be altered as desired. We named this new software “Polyglot”. Our conversion algorithms are conceptually similar to the vOICe insofar as frequency is always used to represent vertical position, time is always used to represent

horizontal position (the image is heard over 1 second from left to right with panning), and sound amplitude is always used to represent luminance. However, other detailed parameters of the device were free to vary.

In three experiments, we varied the task that participants performed and, thereby, varied the fitness function that controlled the evolution of the device. In the first task, we simply asked participants to indicate their aesthetic preference for one soundscape (generated by the IGA) relative to another (the vOICe). In the second, we use an objective fitness function using a two-alternative forced choice in which participants had to determine which image a soundscape was derived from. In the third task, the fitness function was based on participants' ability to make a same/different discrimination between two soundscapes. These three tasks were chosen on the basis that users need to be able to discriminate changes in the soundscape (Experiment 3), link sound to vision (at least in those with residual vision or prior visual experience; Experiment 2), and not find them aversive (Experiment 1).

2.5 General methods

Given that the same genome design is used in all of the studies, it is outlined here first. We then describe the general method for evolving over time.

2.5.1 Genome design

The genome for the SSD consisted of eight different “traits” coded by a total of 16 bits (i.e., generating 216 unique genomes). The eight traits consist of the following:

i) X-resolution (XRes). This is the horizontal resolution used when the image is down-sampled and corresponds to the number of discrete time points in the soundscape. This had eight levels from 10 to 80 in steps of 10. The vOICe has 176 but such a resolution could not be achieved with the present software due to the computational demands of manipulating a larger number of algorithms.

ii) Y-resolution (YRes). This controls the down-sampling of the image in the vertical dimension and also corresponds to the total number of discrete frequencies that are allocated. Again, this had eight levels from 10 to 80 in steps of 10. The vOICe has a Y-resolution of 64.

iii) Minimum frequency (MinF). The lower bounding (floor) frequency could be one of four levels: 250 Hz, 500 Hz, 750 Hz, or 1,000 Hz.

iv) Maximum frequency (MaxF). The upper bounding (ceiling) frequency could be 2,500 Hz, 5,000 Hz, 7,500 Hz, or 10,000 Hz. Note that the vOICe uses frequencies between 500 Hz and 5,000 Hz.

v) The distribution of frequencies between the floor and ceiling was calculated in four ways: linear, musical (Western), musical (constrained), and inverse logarithmic. The simplest is the linear distribution, where each row in the image is allocated a frequency that is proportional to the number of the row:

$$f = MinF + \frac{i \cdot (MaxF - MinF)}{YRes - 1}$$

where f is the resulting frequency, and i is the (zero-indexed) number of the row whose frequency is currently being allocated.

The frequencies may instead by allocated using a distribution that uses intervals based on Western music, such that each octave is composed of 12 notes (semitones) and that notes one octave apart are exactly double in frequency. The following formula ensures this distribution:

$$f = \frac{MinF + MaxF}{2} \cdot 2^{\frac{2i - YRes + 1}{24}}$$

In this formula, MinF and MaxF are used to centre the distribution but do not provide hard constraints on the actual upper and lower frequency bounds. In order to enforce the bounds, the number of discrete notes that can occur for a given doubling of frequency must not be fixed at 12, but should be free to vary as in the following formula:

$$f = MinF \cdot 2^{\frac{i \cdot \log_2\left(\frac{MaxF}{MinF}\right)}{YRes - 1}}$$

This is effectively a logarithmic distribution. As the frequencies increase, so too do the intervals between them. Both of these musically based distributions approximate psychoacoustic performance (Stevens & Volkmann, 1940). As a fourth option, we can generate an approximation of the symmetric distribution—an inversely logarithmic mode of frequency allocation.

$$f = MinF + \frac{(MinF - MaxF) \cdot \log_{10} i + 1}{\log_{10} YRes}$$

This last option is also bounded by the floor and ceiling frequencies, but has decreasing intervals between frequencies as the frequencies increase. Figure 5 illustrates the transfer function in each case.

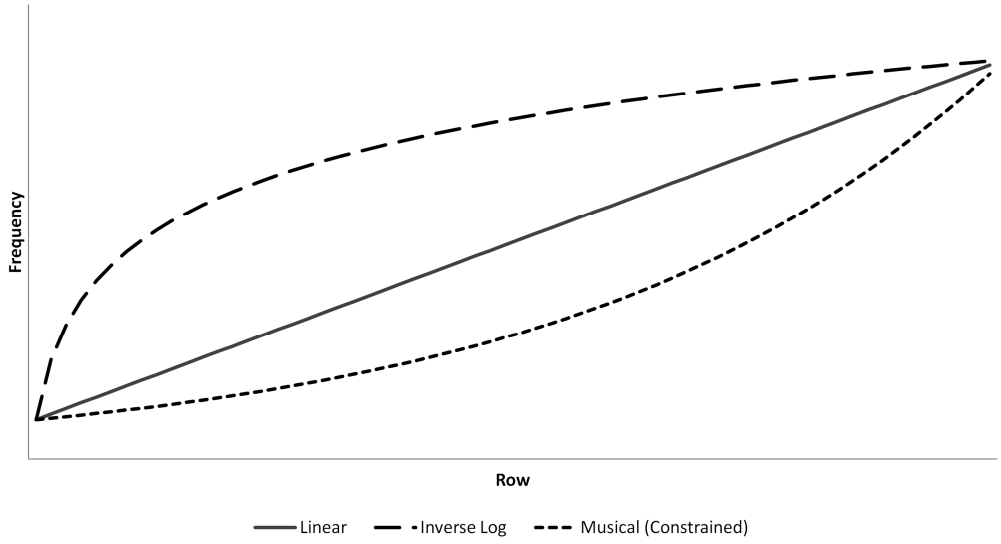


Figure 5: Comparison of frequency allocation modes: linear, inverse log, musical (constrained).

vi) Contrast function. This determines the way in which luminance is mapped to amplitude (relating to perceived loudness) and is set to four levels. The first option is for no contrast adjustment to be made. In this case, there is a linear relationship between luminosity and loudness. That is, a = v where a is the amplitude of the sound (0 to 1), and v is the luminance value (0 to 1). This is the setting used by the vOICe. The other three potential settings all involve the application of a sigmoid function, which causes values to be moved away from the middle area. Increasing the steepness of this function causes light greys to become lighter and dark greys to become darker. These three options are described by the following:

$$a = \frac{1}{1 + c^{-((v \cdot 20) - 10)}}$$

where c is the curve steepness and takes one of three values, 2, 8, and 32, corresponding to small, medium, and large contrast adjustments. The latter effectively renders the image as two-tone, meaning that each frequency is either maximally loud or silent. The luminosity values are scaled to range between –10 and 10 in order to make the asymptotes of the sigmoid approach 0 and 1 within the operational range.

vii) Normal/reversed contrast. In normal contrast, bright is loud, and in reverse contrast, bright is quiet. Reverse contrast is achieved by adjusting the formulae in (6) to

$a = 1 - v$ and $a = 1 - \frac{1}{1+c^{-((v \cdot 20) - 10)}}$

viii) Pitch–space relationship. In the normal setting, high pitch is allocated to the top of the image, and in the reverse setting it is allocated to the bottom of the image. In (5) this achieved by incrementing i either up from 0 to YRes – 1 or down from YRes – 1 to 0.

2.6 Experiment 1: Auditory aesthetics

The first aspect of the vOICe that the present study sought to improve was the aesthetic properties of the sounds it generates. As IGAs have often been used with liking/preference as a fitness function, the first experiment offers a proof of principle that it can be extended to the sensory substitution domain. At a pragmatic level, an unpleasant sound from an SSD may limit their uptake among visually impaired people and should be an important consideration for these kinds of devices in general (Song & Beilharz, 2008). There are also theoretical insights to be gained in terms of understanding how aesthetic judgements depend on the underlying architecture of perception. Are the features that are selected for in a soundscape on aesthetic grounds the same as those that optimize objective performance on discriminating or identifying the soundscape? In some theories, aesthetic judgement is underpinned by the same mechanisms as those that support perception (Zeki, 1999), whereas, in other theories, aesthetics is far more related to reward and experience (e.g., via motor resonance) than the characteristics of perception (Cinzia & Vittorio, 2009).

2.6.1 Method

2.6.1.1 Participants

Twenty students (15 female, aged between 18 and 39 years) were recruited from the University of Sussex and were awarded course credits for their participation. In this and subsequent experiments, ethical approval was granted by the Life Sciences & Psychology Cluster-based Research Ethics Committee at the University of Sussex. Similarly, in this and subsequent experiments, all participants reported normal hearing and normal (or corrected-to-normal) vision.

2.6.1.2 Materials

The stimulus material consisted of 30 natural images of everyday indoor and outdoor scenes (three examples are given in Figure 6). One image was used in each block selected randomly, without replacement, from the pool of 30 images. On each trial, an image was sonified twice: once using the vOICe and once using one of the conversion algorithms selected in that generation. Consequently, participants never saw the images—they only heard them. Their task was to indicate their preference as described in detail later.

Figure 6: Three examples of source stimuli used in Experiment 1

2.6.1.3 Procedure

Participants were instructed that they would hear two different sounds, and their task was simply to rate their degree of preference for one sound over the other. They were given no further information about the origin of the sounds (i.e., that they were based on images).

They were seated at a computer screen (337 × 270 mm) at a comfortable viewing distance and wore headphones (Sony, MDR-XD100). On the screen was a horizontal visual analogue scale and a "play" button on each side of the screen. The participant was required to click the buttons with the mouse to listen to each sound. Participants listened to each sound twice. They were then required to move a pointer on a visual analogue scale on the computer that was initially placed in the centre of the line. The two ends of the line were defined as "Prefer Sound 1" and "Prefer Sound 2". The distance along the line (from the "vOICe" sound to the evolved sound) was computed, and this served to define the fitness function. The "vOICe" soundscape was randomly allocated as either Sound 1 or Sound 2. There were 10 trials in each block and 15 blocks. At the end of each block, participants were given a self-paced break and were asked to press a button to continue. At the end of each block, the computer generated a new set of genomes to be used in the next block. The experiment typically lasted for 40 minutes.

2.6.2 Results

In order to assess the performance of the IGA, we propose that it should be determined whether any traits are more (or less) common than would be expected by chance. Here, our statistics are based entirely on the final generation, although we show graphically how selection emerges across generations. Each organism is treated as an independent observation in a chi-square test, and we apply a Bonferroni correction to take into account the fact that we are exploring eight traits (i.e., an alpha of .05/8). This analysis determines whether selection has occurred (across the sample of genomes) but it does not tell us about the selection

behaviour of the sample of participants—that is, whether the group as a whole made that selection, or whether it was biased by the performance of a few participants. To asses this, we additionally compared the proportion of a given trait in the final generation against the expected rate based on chance using (post hoc) one-sampled t tests.

The chi-square analyses revealed that four traits showed selection. Figure 7 shows the proportion of genomes containing different frequency allocation methods across generations; at the final generation, $\chi^2(3, N = 200) = 21.52$, $p < .001$. In this example, one trait—namely, "musical (Western)"—is selected against (i.e., appears less common in the population than expected). This pattern of selection was found across participants: $t(19) = –2.624$, $p < .05$, for "musical (Western)"; other traits not significant from .25.

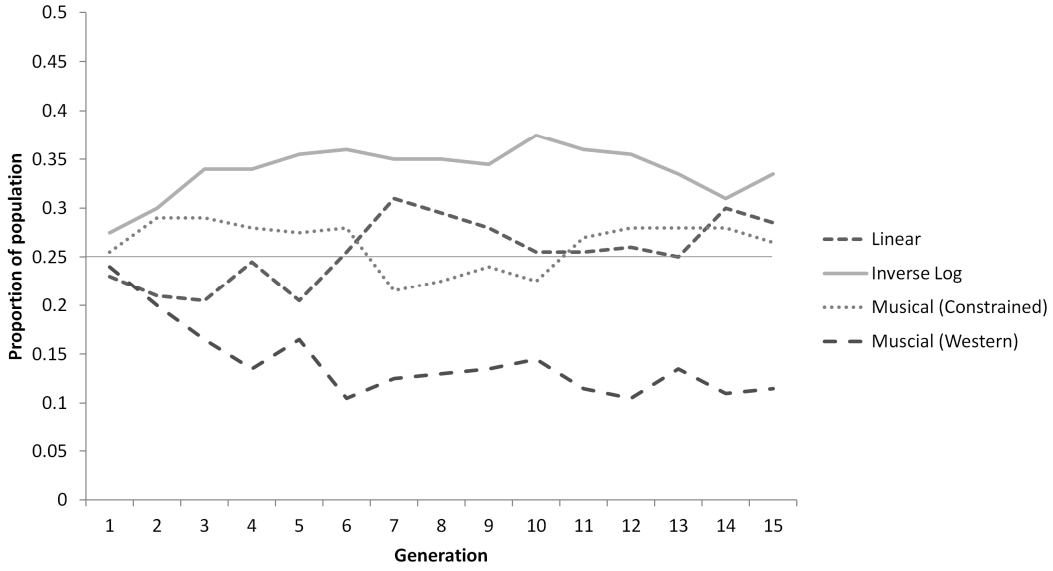


Figure 7: Proportion of each frequency allocation mode over 15 generations in Experiment 1 as selected by 20 participants. The trait of "musical (Western)" is selected against.

The second trait that exhibited selection was the contrast function, which, in the auditory domain, relates to the distribution of different amplitudes, $\chi^2(3, N = 200) = 65.08$, $p < .001$. This is shown in Figure 8. In this instance, one trait is selected *against* (medium contrast adjustment) and another is selected *for* (low contrast adjustment). This is confirmed by post hoc *t* tests (low adjustment: $t(19) = 4.36$, $p < .001$; medium adjustment: $t(19) = –8.72$, $p < .001$). This demonstrates that selection can be very specific even when given a trait that varies monotonically.

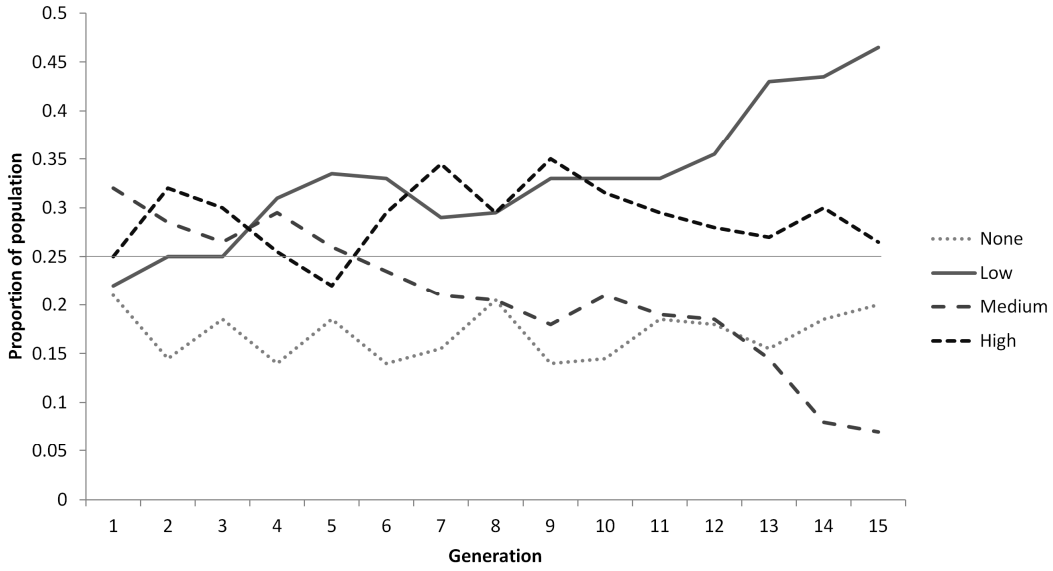


Figure 8: Proportion of each contrast enhancement mode over 15 generations in Experiment 1 as selected by 20 participants. Whereas a small contrast enhancement is selected for, a medium contrast enhancement is selected against.

Another monotonically varying trait that showed selection was the Y-resolution, $\chi^2(7, N = 200) = 35.28$, $p < .001$. In the auditory domain, this refers to the number of discrete frequencies that are heard. This is illustrated in Figure 9, collapsing the 8 traits into 4 bins. In this instance, there is a monotonic relationship between the number of discrete frequencies and their likelihood of selection (such that more frequencies are preferred). Statistically, resolutions of 10–20 are reliably selected against, $t(19) = -2.25$, $p = .036$, and resolutions of 70–80 are reliably selected for, $t(19) = 2.34$, $p = .030$, with intermediate values not reaching significance. A similar pattern is found for the upper bound frequency, MaxF, $\chi^2(3, N = 200) = 17.08$, $p = .001$, with the highest frequency, 10 kHz, reliably selected against by participants, $t(19) = -2.83$, $p = .011$, and the lowest frequency, 2,500Hz, reliably selected for, $t(19) = 2.28$, $p = .035$. This is shown in Figure 10.

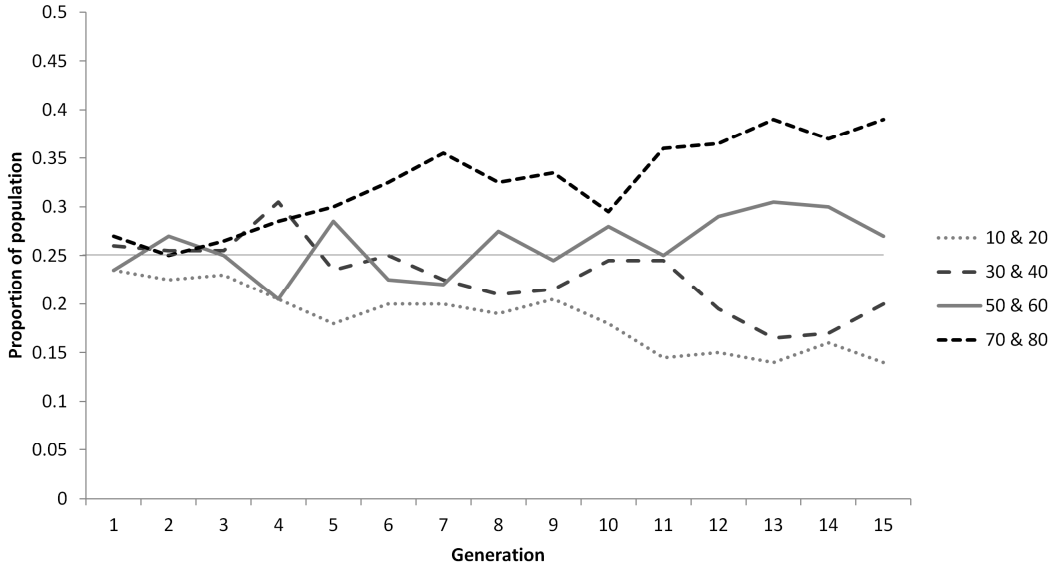


Figure 9: Proportion of genomes containing a given Y-resolution (number of discrete frequencies) over 15 generations in Experiment 1 as selected by 20 participants. There is a monotonic relationship between resolution and prevalence in the final generation.

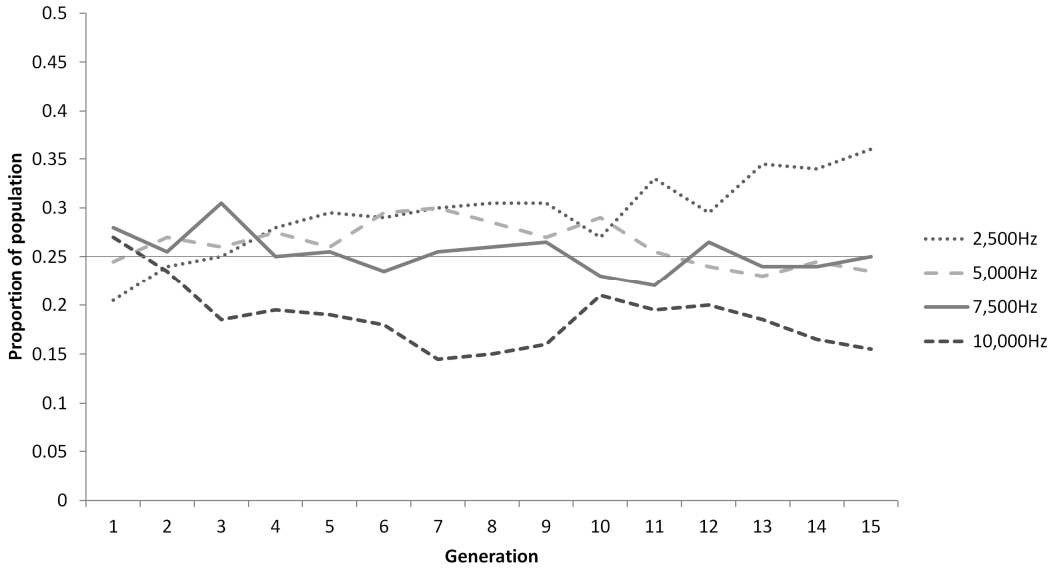


Figure 10: Proportion of frequency range ceilings over 15 generations in Experiment 1 as selected by 20 participants. Note that 2500Hz is selected for and 10,000Hz is selected against.

2.6.3 Discussion

This experiment has demonstrated that IGAs can be used to inform the design of conversion algorithms, such as those used in sensory substitution, by rating the pleasantness of the resulting sounds. In this instance it was done by comparing the aesthetics of an evolving device ("Polyglot") with that of a fixed conversion algorithm in widespread use in the literature (the "vOICe"). Importantly, the aesthetically optimized properties are not necessarily those that would be predicted from the perceptual performance of the auditory system. If aesthetics

were tied closely to perceptual performance, we would predict that a “musical” (logarithmic) distribution of frequencies would be positively selected when in fact, if anything, it is selected against. (Indeed in Experiment 3, we show that such a trait is selected for when the fitness function is perceptual rather than aesthetic.) The sensory substitution device of Cronly-Dillon et al. (1999) is based on the Western musical system (albeit using concert pitch). An adapted system that sonifies in a musical key (i.e., using a subset of the 12 semitones in the octave) may fare better, but would reduce the overall number of tones that can be used to represent the image (which, in our study, was positively selected). In sonified images there will be a natural tendency for adjacent notes to be played together (because the intensity at a given pixel tends to be correlated with that of its neighbours) but this rarely occurs in music and is perceived as highly dissonant.

With other natural sounds (e.g., made by animals or objects), unpleasantness has been linked to high energy in the 2,500–5,000-Hz range (Kumar, Forster, Bailey, & Griffiths, 2008). There is some evidence consistent with this in our study—a ceiling of 2,500 Hz was selected for (and a very high ceiling selected against). This has been linked to the fact that sounds in the 2,500–5,000-Hz range are perceived as subjectively louder (ISO 226:2003; International Organization for Standardization, 2003), but may also depend on an interaction with other acoustic features (e.g., temporal modulation; Kumar et al., 2008).

2.7 Experiment 2: Audiovisual matching

The second experiment consisted of presenting participants with two images and a single soundscape, which was derived from one of these images using, in the first instance, a randomly generated genome (with mating in subsequent generations). The participants' task was to decide which image the soundscape was derived from. As such, the resulting fitness function is based on a performance measure (correctness). From an applied perspective, it needs to be borne in mind that blindness and visual impairment represent a spectrum of functioning with people having differing levels of residual vision and differing levels of visual history. For many blind individuals, the function of an auditory SSD may be to integrate the auditory information with residual vision rather than being a true substitution.

From a theoretical perspective, there are reliable “rules” that people adopt when linking together auditory and visual features—for instance, between pitch and size (Parise & Spence, 2009), pitch and space (Melara & O’Brien, 1987; Pratt, 1930), loudness and luminance (Marks et al., 1987), and pitch and shape (Marks et al., 1987; Parise & Spence, 2009). Many of these are present from a very early age, suggesting that they are not learned (e.g. P. Walker et al.,

2010). However, this literature is based either on preference measures for audiovisual associations (Ward, Moore, Thompson-Lake, Salih, & Beck, 2008) or on interference-based measures showing, for instance, a modulation of response time by a task-irrelevant incongruent modality (Marks et al., 1987) or a disruption of temporal order judgements for bound relative to unbound audiovisual stimuli (Parise & Spence, 2009). By showing that these associations are selected for in an audiovisual matching task, we aim to demonstrate that these associations may also enhance accuracy-based performance when congruently paired.

2.7.1 Method

2.7.1.1 Participants

Twenty sighted participants (12 female, aged between 18 and 35 years) were recruited from the University of Sussex and were compensated with course credits. None had participated in Experiment 1.

2.7.1.2 Materials

In a departure from Experiment 1, natural images were not used in this experiment. In order for selection to occur, there needs to be sufficient variability in performance across trials that is neither at floor or ceiling. In pilot studies with natural images, our controls were close to chance across many trials. For the genetic process to be meaningful, the fitness function should also be meaningful, which in this case implies that participants need to be performing better than would be expected from random choices. Instead, participants were asked to choose between two images taken from the cartoon TV show “The Simpsons”. The surface areas of constant luminosity combined with small details made the images a good balance between simplicity and variability. There were 20 images available, cropped to be square, and each genome was evaluated using a new, randomly selected pair from the pool.

2.7.1.3 Procedure

Participants were given a basic description of the process by which the images are converted into sounds—this did not include any allusions to the parameters under test, but did make clear that the sounds scanned from left to right over the image over the course of one second. They were instructed that they would hear one of these sounds and see two images, one on each side of the screen. After listening to the sound twice, their task was to indicate which image they believed the sound to have been generated from using a horizontal visual analogue scale. Participants were instructed to move the pointer (initially located in the centre) towards ends labelled as “Image 1” and “Image 2” according to their decision and their degree of certainty in it.

As in Experiment 1, each participant was seated at a computer screen (337 mm × 270 mm) at a comfortable viewing distance and wore headphones (Sony, MDR-XD100). The genetic algorithm used the same parameters as those in Experiment 1. The sole exception to this was the number of generations: Participants found this task more taxing, so the number of iterations was reduced from 15 to 10. It took approximately 30 min to complete.

2.7.2 Results

As the responses in this task are either objectively correct or incorrect, we can first examine the overall scores by generation. This allows us to verify whether the process of selection is working as expected. In this case, regressing the mean score of each participant against generation number reveals that scores did improve (R^2 = .048), $F(1, 198) = 9.96$, $p < .01$. The estimated means in Generations 1 and 10 were 53.7% and 62.0%, but note that even in the final generation there is a range of genomes present (many of which will not be optimal).

As in Experiment 1, we report all traits for which a significant result was obtained by employing a chi-square test on the final generation. Three traits showed evidence of selection.

The upper bound frequency (MaxF) showed selection, $\chi^2(3, N = 200) = 29.60$, $p < .001$. The highest frequency of 10 kHz is selected against, $t(19) = -7.393$, $p < .001$, with others not differing from baseline. This is shown in Figure 11. In this instance, the same trait is selected against both when the fitness function is a simple preference (in Experiment 1) and also when the fitness function is based on task performance (matching a sound to a picture).

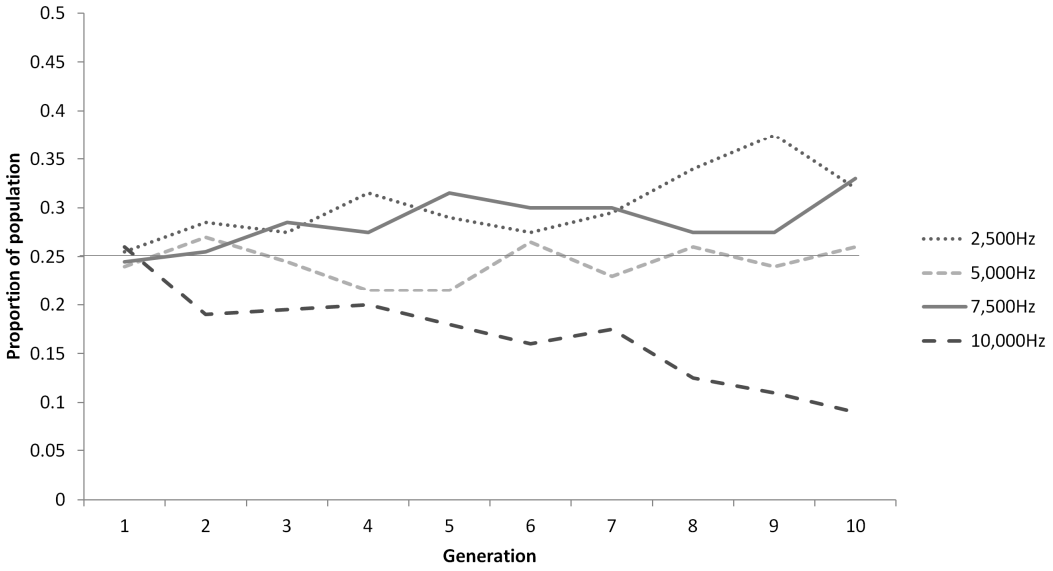


Figure 11: Proportion of frequency range ceilings (in Hz) over 10 generations in Experiment 2 as selected by 20 participants. Note that 10,000Hz is selected against.

The other two traits that were selected for were those identified from previous research on audiovisual interactions—namely, pitch–space, $\chi^2(1, N = 200) = 19.22$, $p < .001$, and loudness–luminance, $\chi^2(1, N = 200) = 12.50$, $p < .001$. Specifically, the tendency for high frequency to be linked to high space (rather than low space) was selected for as was the tendency for brightness to be linked to high-amplitude sounds (rather than silence). This is shown in Figure 12. (Note: with a binary trait, selection *for* one trait necessarily implies selection *against* the other trait.) Again, the results were found when we consider the behaviour of participants (pitch–space: $t(19) = 2.66$, $p = .015$; loudness–luminance: $t(19) = –2.16$, $p = .044$).

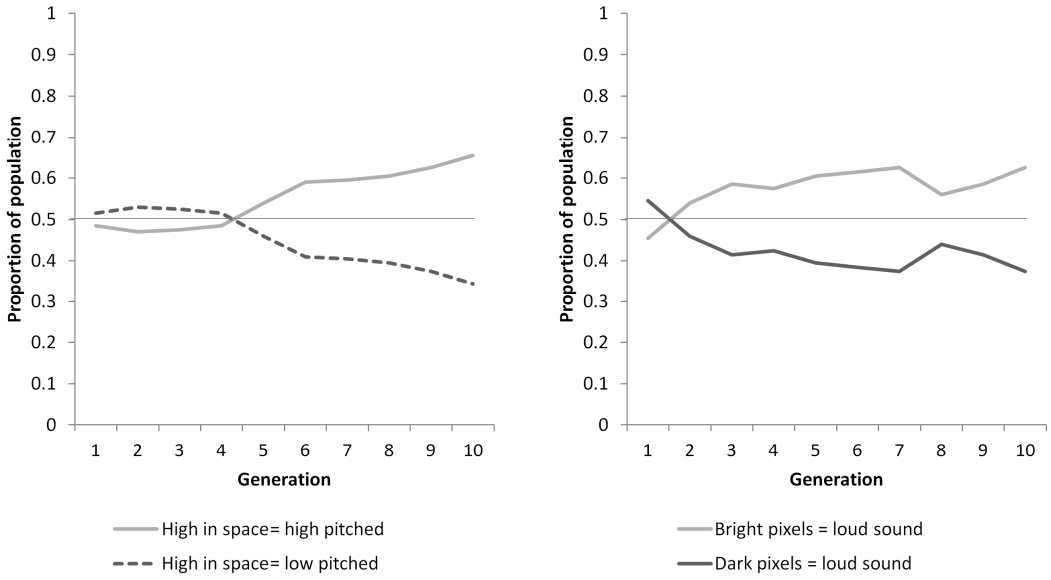


Figure 12: Proportion of pitch-space genomes (left) and luminosity-loudness genomes (right) over 10 generations in Experiment 2 as selected by 20 participants.

2.7.3 Discussion

In addition to selecting against a 10,000-Hz ceiling frequency, the most important findings of this study are that "congruent" luminance–loudness relationships (bright = loud) and "congruent" pitch–space relationships (high pitch = high space) are selected for; that is, these associations serve a functional role in enabling soundscapes to be linked to visual information. This is likely to be important for blind users of such a device who have some degree of residual vision. For these individuals, an optimal sensory substitution device may enable the best integration of auditory-derived vision and residual vision, rather than necessarily being the most efficient psychoacoustically. Interestingly, certain traits that are likely to enhance auditory discrimination itself (e.g., a logarithmic pitch series) were not selected for. Whilst this could reflect a lack of statistical power, our final experiment suggests that this may not be the

case. Specifically, such traits are selected for when the task is solely auditory rather than auditory–visual.

2.8 Experiment 3: Auditory discrimination

A key advantage of visual–auditory SSDs over visual–tactile SSDs is the ability to increase the resolution of the encoded image without modifying the hardware—in theory, the only limit to resolution in auditory systems is the ability of users. In this experiment, only auditory stimuli were used. Participants listened to two soundscapes, generated via the same algorithm, and were asked to determine whether they were the same or different. As such, the fitness function in this experiment was an objective measure of performance (how well the soundscapes could be discriminated), as in Experiment 2.

2.8.1 Method

2.8.1.1 Participants

Twenty sighted participants (11 female, aged between 18 and 28 years) were recruited from the University of Sussex and were compensated with course credits. Some participants (N = 4) had previously taken part in Experiment 1, but this was deemed to be non-problematic as Experiment 1 was a preference task whereas this experiment required skill—it was not possible to cheat or to bias the outcome.

2.8.1.2 Materials

We used the 20 (square-cropped) images taken from the cartoon TV show “The Simpsons”, as in Experiment 2. Each image was then used to generate another, by rotating a randomly designated segment by 180 degrees. These segments were squares with sides that were 50% of the length of the whole image, such that they had an area equal to 25% of the total image area. This operation was chosen because it disrupts the shapes in the image without altering the overall contrast or luminosity. The size of the segment was determined by previous pilot research.

2.8.1.3 Procedure

As in Experiment 1 they heard two soundscapes derived from the images. Participants were asked to press a button (marked “play”) and to listen to each sound twice before indicating whether they believed that they were the same or different. Participants did not see any images and were not informed that the sounds were generated from images. Each genome was used twice—once to sonify a pair of unmodified images (“same” condition) and once with

an unmodified image paired with a modified image (“different” condition). Participants were naïve as to how the sounds were constructed.

Rather than use a visual analogue scale, this experiment used a two-alternative forced-choice paradigm (buttons labelled “same” and “different”). This was because in previous studies we observed that participants tended to resort to a binary placement along the visual analogue scale rather than using the entire range of values.

Each genome started with a fitness score of 0. If the participant responded correctly, the score was increased by 0.45 each time, so that a maximum score of 0.9 could be obtained. If the participant responded incorrectly, the score was increased by 0.05, so that the minimum score each genome could obtain was 0.1. (Values of 0 and 1 were not used since any 0-scored genomes would not be represented in the weighted-stochastic selection process.) Given that the scores in this case were discrete rather than continuous, the elitism employed in the previous experiments did not take place. Due to the additional time spent on each genome (as they were evaluated twice), the number of genomes per generation was reduced to 7, and the number of generations was 10. All other aspects of the genetic algorithm were as described for Experiment 2. The experiment took approximately 45 min to complete.

2.8.2 Results

Once again, regressing the mean score of each participant each generation against generation number shows that scores improved ($R^2 = .051$), $F(1, 198) = 10.58$, $p < .05$. The estimated means in Generations 1 and 10 were 50.4% and 59.5%. Four traits showed evidence of selection when assessed in the final generation.

Figure 13 shows that there is a distinct advantage conferred by the musical types of frequency allocation, $\chi^2(3, N = 140) = 17.20$, $p = .001$. This trend is visible from the fifth generation. Collapsing across the two musical modes reveals that participants showed a reliable selection bias for these pitch series, $t(19) = 2.92$, $p = .034$. This fits our understanding of the distribution of sensory resources in the ear: The resolving ability of the cochlear is greater (following a roughly logarithmic pattern) at higher frequencies (Steinberg, 1937). However, it is interesting to note that these were not previously selected for when the fitness function was auditory aesthetics or audiovisual matching even though the auditory soundscape was task relevant in all three experiments.

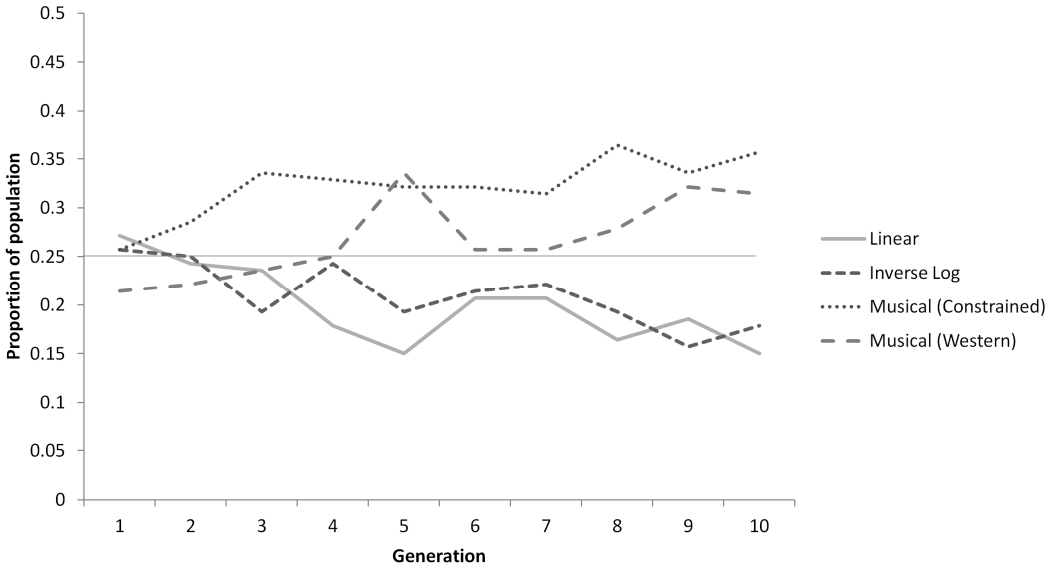


Figure 13: Proportion of genomes containing a given frequency allocation modes over 10 generations in Experiment 3 as selected by 20 participants. Note that musical (i.e., logarithmic) distributions of discrete frequencies are selected for.

Figure 14 shows that the frequency range floor (i.e., the lowest frequency in a soundscape) of 750 Hz is strongly selected for in the final generation, $\chi^2(3, N = 140) = 30.23$, $p < .001$, and is reliable across the group of participants, $t(19) = 2.87$, $p = .01$. This is likely to be the result of competing pressures: towards a lower frequency in order to expand the range and towards a higher frequency in order to avoid the lowest frequencies. More research is needed to clarify the exact mechanics at play here.

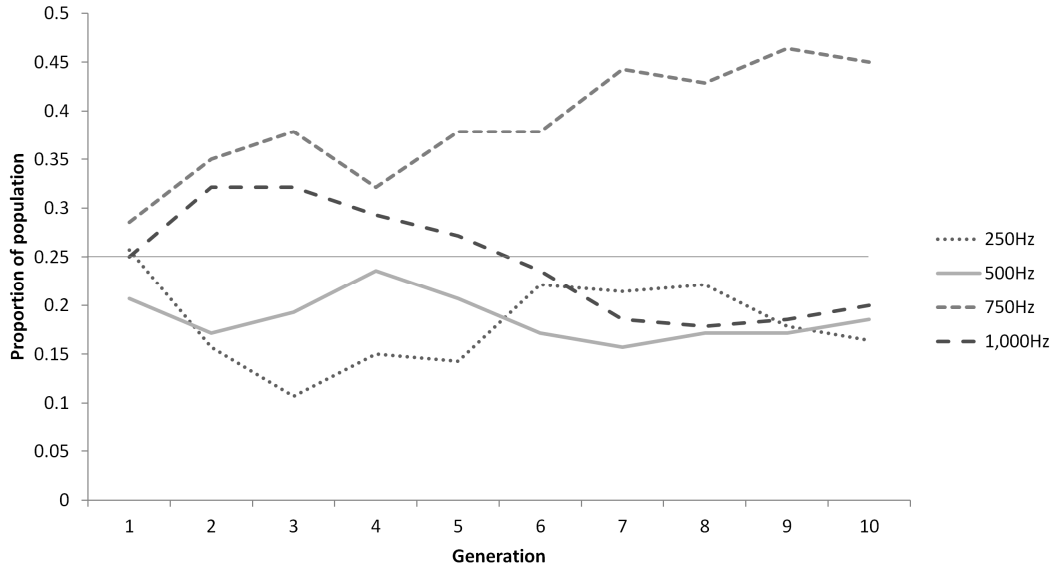


Figure 14: Proportion of genomes containing a given frequency range floor (in Hz) over 10 generations in Experiment 3 as selected by 20 participants. Note that 750 Hz is selected for.

The X-resolution (i.e., number of discrete time points) showed evidence of selection when assessed in the final generation, $\chi^2(7, N = 140) = 22.50$, $p = .001$. Inspection of the data revealed that selection was based against the two lowest resolutions, and Figure 15 illustrates this, collapsing the 8 resolutions into 4 bins. When looking at these binned data across all participants, it is clear that this selection is the only significant deviance from baseline, $t(19) = -2.77$, $p = .012$. Interestingly, the other X-resolutions do not show evidence of being selected for, and nor is there a monotonic trend for greater resolution to offer the greatest benefits. Beyond a value of 30, there is no further observable benefit (at least in naïve participants), which suggests a perceptual resolution of users that is far less than the technology can deliver (recall that the vOICe has an X-resolution of 176).

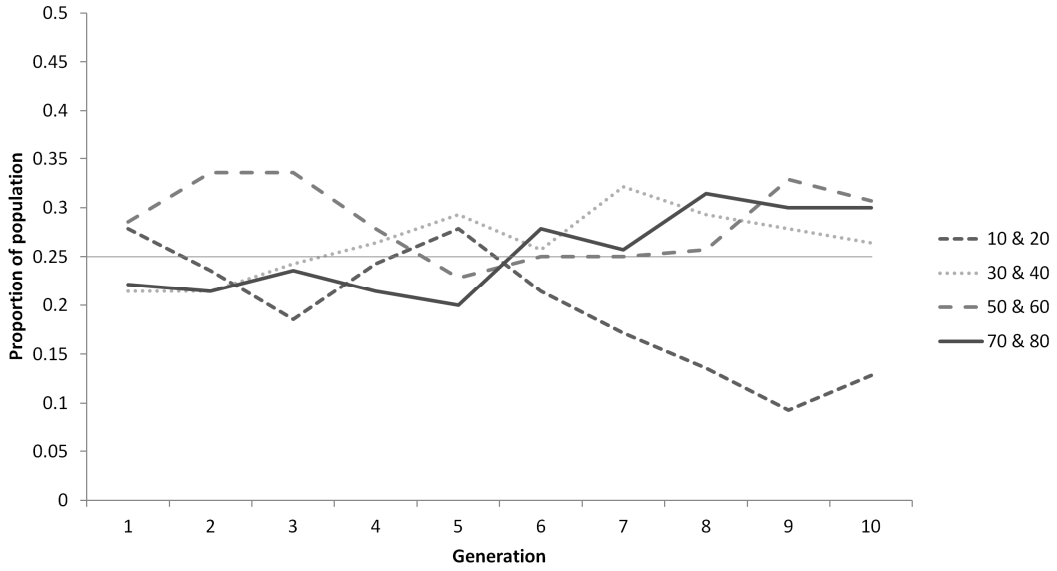


Figure 15: Proportion of genomes containing a given X-resolution (number of separate time points) over 10 generations in Experiment 3 as selected by 20 participants.

The last significant result in this study is shown in Figure 16. Surprisingly, perhaps, in a task of auditory discrimination there is a benefit from having the pitch–space association inverted; that is, high spatial positions coded by lower frequencies are selected for: across genomes, $\chi^2(1, N = 140) = 19.31$, $p < .001$; across participants, $t(19) = 2.65$, $p = .016$. Natural images (and cartoon images of the real world) tend to be visually busier in the bottom half than in the top half. The latter is due to the greater presence of plain surfaces such as walls and the sky at the top. There is also a tendency for the top part of images to be brighter (they normally contain a light source and fewer shadows). Both of these factors may potentially contribute to this effect although note that if the images simply had too many loud components to resolve then we would have expected loudness–luminance inversion (i.e., bright = quiet) to have been selected for, rather than pitch–space.

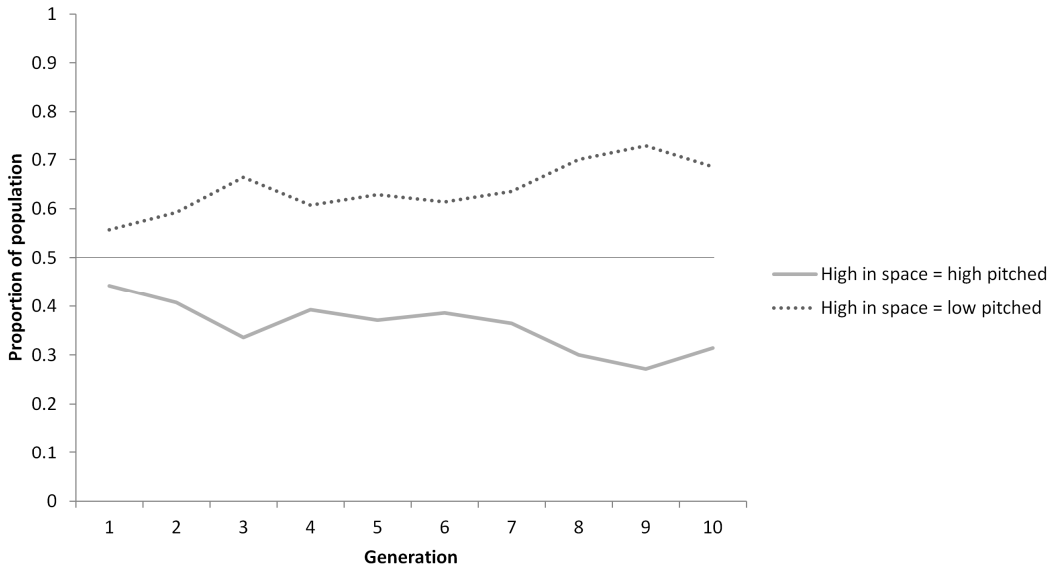


Figure 16: Proportion of genomes containing pitch-space inversions over 10 generations in Experiment 3 as selected by 20 participants. Note that high space = low frequency is selected for.

2.8.3 Discussion

As expected, when "Polyglot" evolves on the basis of auditory discrimination, there is a tendency for musically based (i.e., logarithmic) distributions of frequencies to be selected for. Moreover, there needs to be sufficient temporal variability (X-resolution) in the soundscape (more than 20 Hz). However, other findings are unexpected. First, we may have expected that greater spectral variability (Y-resolution, number of discrete frequencies) and greater amplitude variability (luminance–loudness contrast adjustment) would have been selected for, as both give rise to an acoustically richer soundscape. During the evolution process, there tends to be a moderate degree of "epistasis"—the parameters interact with each other to control the transformation (Haupt & Haupt, 2004, p. 32). For instance, the selection of one trait (e.g., distribution of frequencies) may interfere with selection of other traits (e.g., number of frequencies). There could also be trivial reasons for a null result (e.g., too few generations, the fitness function not sufficiently discriminating). A second unexpected finding is the selection of an inverted pitch–space association. We speculate that this is due to the statistical regularities in the top and bottom halves of images that are then translated into the soundscapes (i.e., bottom halves are darker and more crowded, on average). Given that the images were selected to be representative of scenes that might be encountered by a user of a sensory substitution device, these statistical regularities are artefacts of the ecological validity of the experiment. However, further testing in which the image properties are varied in a more systematic way would be needed to confirm and understand this finding.

2.9 General discussion

In the present study, interactive genetic algorithms were applied to a configurable sensory substitution device that we termed “Polyglot”. The key advantage of this method is that it allows researchers to explore a much larger problem space than is conventionally possible and to converge on solutions relatively quickly (e.g., as little as 15 hours of collective testing per experiment). In the General Discussion, we consider the theoretical and methodological implications of our study, considering, first, the implications for sensory substitution research and, secondly, the wider applicability of this method in psychological research.

2.9.1 Implications for sensory substitution research

Previous research on sensory substitution devices has tended to test only a single device at a time, giving little insight into the merits and pitfalls of each approach. More recently, Brown et al. (Brown et al., 2011) explored different settings within the “vOICe” device—for instance, comparing contrast settings (bright = loud vs. bright = quiet) and the length of the soundscape (1 s vs. 2 s) in a 2 × 2 design. This is one of the first attempts to determine the optimal parameters for perceiving sonified images in a sensory substitution device, but the number of parameters that can be varied in a given experiment is very low. The use of interactive genetic algorithms marks a step-change in our ability to explore this. It enabled a large number of parameters to be evaluated and a way of comparing optimal parameters across tasks. The results of the parameters selected for (and against) in the three tasks employed here are summarized in Table 1.

Parameter	*Experiment 1*	*Experiment 2*	*Experiment 3*
Frequency allocation	**-** Musical (Western)		**+** Musical (Western) +Musical
Contrast function	**+** small, - medium		
Frequency range floor			**+** 750Hz
Frequency range	+2,500Hz, **-** 10,000Hz	**-** 10,000Hz	
X resolution (time)			**-** Small
Y resolution	**+** Large		
Pitch-height		High Pitch = Top	High Pitch = Bottom
Luminosity-loudness		Bright = Loud	

Table 1: Summary of the results from all three experiments

Across the three experiments, all of the eight parameters that we varied were subject to selection at one point or another. However, the results reveal that the particular parameters that affect performance in one task are not the same across tasks. This is of interest given that the potential pool of soundscapes (as specified by the genome) was common to all tasks. Differences in the images between Experiment 1 and Experiments 2 and 3 are unlikely to be the main source of difference in results, given the diversity within each image pool and the fact that all stimuli simulated everyday scenes. As such, the optimal properties of an auditory sensory substitution device are driven as much by the task as by the limits of the ear and auditory system (and the stimuli used). An interesting comparison here is between Experiments 2 and 3, in which not only was the auditory genome the same but the images from which the soundscapes were derived were also the same. When one has to discriminate two soundscapes from each other, a logarithmic distribution of frequencies is beneficial (as expected from the performance of the ear). However, this does not apply when one has to match a soundscape to the visual image from which it is derived. Similarly, allocating high frequencies to represent the top of an image is beneficial when the task is to match images to soundscapes but not when discriminating between soundscapes themselves.

At an applied level, we can offer empirically derived suggestions for what an optimal configuration of a sensory substitution device would be that satisfies all three task constraints. Specifically, one may wish to develop a device that operates in the lower frequency range (up

to 2,500 Hz) using a musical (non-Western) distribution of frequencies, a small contrast adjustment that maps high luminance to high amplitude, at least 30 time points, and up to 80 discrete intervals. The pitch–height relationship may be task dependent. This could be explored in future work.

It would be important to test such a device against others, such as the vOICe, and to extend the research to the blind and visually impaired. Although visually impaired people tend to perform better (Arno, Vanlierde, et al., 2001) and undergo functional changes to their brains (Kupers, Chebat, Madsen, Paulson, & Ptito, 2010; Ortiz et al., 2011), blindfolded sighted participants can complete sensory substitution tasks and are not necessarily qualitatively different despite being quantitatively worse. Finally, such devices may be useful in the sighted population itself by offering a dual-coding of vision—that is, by supplementing natural vision with an auditory presentation of vision.

As a notable limitation, the present research omits one of the most important components of learning to use a sensory substitution device—namely, the motor component. In order for the participant to link an auditory component (e.g., a high-pitch sound in the second time point) to an external location/object that can be acted upon, they must also "embody" the device itself (Brown et al., 2011; O'Regan, 1992). For instance, if a camera is worn on the head, then the position in space that the sound denotes is determined by the current orientation of the head in addition to the properties of the sound itself. The extent to which the parameters explored above would affect this process of embodiment is unknown, but at least one of them is expected to be important from current evidence. Specifically, the link between vertical space and pitch may be akin to a sensory–motor affordance in which there is an intuitive link between pitch and space (and this is presumably independent from vision, although there are no known data on that). In terms of perceptual discrimination, high-pitched sounds are perceived to emanate from higher locations (Pratt, 1930), and infants associate these dimensions together in preferential looking (P. Walker et al., 2010). Thus, maintaining a link between high frequency and high space may remain the optimal configuration for such a device even if it transpires that, from a purely psychoacoustic point of view, sonified images are easier to discriminate when the reverse mapping is applied.

An interesting consequence of the currently presented data is to largely confirm that cross-modal correspondences apply to sensory substitution. Previous experimental work has shown relationships between pitch and vertical position (e.g. Ben-Artzi & Marks, 1995) as well as loudness and luminance (e.g. Marks et al., 1987). Experiment 2 replicates these findings in the

sensory substitution domain, validating the design assumptions of the vOICe and other devices. These associations appear to be useful when linking audition and vision. It is possible that they help the user to "bootstrap" the learning process.

2.9.2 On the use of interactive genetic algorithms in psychology

Recent research in psychology has seen an increase in so-called data-driven approaches using methods such as multidimensional scaling (e.g. Jaworska & Chupetlovska-Anastasova, 2009). IGAs are conceptually similar in that they aim to reduce a large problem space either to an ideal solution in that space or by creating a smaller problem space (e.g., by eliminating parameters that are not selected for). In other respects they differ. In multidimensional scaling the structure is determined by the data themselves, whereas in an IGA the range of possible structures is constrained by the design of the genome. That is, the experimenter must have some knowledge of the likely problem space.

The IGA method lies someway between being an entirely data-driven approach and the more conventional hypothesis-testing methodology. It is possible to test hypotheses using this method. For instance, we hypothesized that participants would select for a bright = loud mapping and a high space = high frequency mapping, and this hypothesis was confirmed. The advantage of the present method is that it enabled us to evaluate a whole host of additional variables alongside hypotheses that were predicted from existing theory.

In perception research there are many domains in which IGAs could be applied. Music would be an ideal system in which this could be applied because musical structure can be easily specified. Consider a recent study by Mesz, Trevisan, and Sigman (2011) in which a group of composers were asked to create musical pieces to denote tastes (e.g., salty music, sour music). The experimenters then analysed the compositions for certain features (e.g., salty music tends to be staccato). This would be easily achievable using IGAs in which initially random excerpts are rated for "saltiness", and then the saltiest excerpts are bred over generations. One obvious advantage in this example is that the participants need not have any formal musical knowledge, and it could be easily done over the internet to generate cross-cultural perspectives. The perception of voices is again another area that is well suited to this approach (for instance, the study by Baumann & Belin, 2010, concerning the role of acoustic features in speaker identification could be done using an IGA). Faces are another candidate for study using this method, although the potential structural components of a face are harder to specify a priori (in contrast to, say, music). As already noted, there is an IGA study exploring what makes a female face beautiful (Johnston & Franklin, 1993). There is also a growing literature on how

perceived social traits (e.g., dominance) are related to facial characteristics such as the facial width-to-height ratio, and many of these studies would lend themselves to an IGA approach (Nestor & Tarr, 2008; Rojas, Masip, Todorov, & Vitria, 2011).

It would also be very interesting to use physiological measures (e.g., galvanic skin response, heart rate, or electroencephalography, EEG) to drive a genetic algorithm. These signals have long been used to determine psychological aspects of a participant, such as the emotional state or degree of arousal (e.g. Lisetti & Nasoz, 2004). Such a system would require a human participant, but would not be interactive in the strictest sense, since the participant is expected to have no conscious control over their response. This type of physiological genetic algorithm could be used, for example, to drive the evolution of an SSD based on the physiological response to the soundscape it produces.

In summary, the question as to how to translate an image into a sound represents an interesting theoretical question and one that has potentially important applied consequences. We have shown that interactive genetic algorithms, based on the perceptual performance/judgements of participants, offer a significant advance in this field.

Chapter 3: Introducing Polyglot

A framework for visual-to-auditory sensory substitution devices

3.1 Abstract

This paper introduces Polyglot – a general purpose, modular framework for visual-to-auditory sensory substitution devices (SSDs). We outline why the exploration of possible SSDs is a huge task and why SSD users may be best placed to lead this exploration. Polyglot aims to allow End User Development (EUD) of novel SSDs by providing a tiered framework of opportunities for engagement: firstly through the customisation of modules, then through runtime composition and finally by the development of new modules. The programming structures developed to assist EUD in Polyglot are discussed, as well as potential avenues for further development of the framework. Lastly, we also consider the potential utility of Polyglot to the research community, both directly and indirectly.

The Polyglot framework has been released as open-source software (Wright, 2013)

3.2 Introduction

Sensory substitution is “the artificial conveyance of rich, abstract sensory information of one sense via a different modality” (Ward & Wright, 2014). Since the creation of the TVSS, the first Sensory Substitution Device (SSD), in the late 60s (Bach-y-Rita et al., 1969), the field has largely been concerned with the substitution of vision in blind people. The TVSS represents the brightness information of 400 pixels as the intensity of the vibration of an array of 400 solenoid stimulators.

More recent systems have predominantly used sound as the substituting sensory modality. This is partly due to the increased available bandwidth (Jacobson, 1950; Kokjer, 1987), but is also likely to be due to the comparative ease and flexibility of generating sounds compared to tactile sensations. This flexibility is a somewhat of a double edged blade: the lack of an immediately obvious set of audio-visual mappings means that researchers are unconstrained in how these two senses are paired, but it also means that finding optimal solutions is non-trivial.

In a previous paper, we described a method for fine-tuning the parameters of a particular implementation using interactive genetic algorithms (Wright & Ward, 2013). To facilitate this, we re-implemented a popular visual-to-audio SSD called “the vOICe” (Meijer, 1992) in such a way as to allow for the customisation of 8 key parameters. Since our re-implementation could produce many variants of the vOICe, we named it “Polyglot”. By rapidly iterating and evaluating permutations of potential configurations under three test conditions, we were able to find good approximations of an optimal configuration. We implemented this configuration in a new SSD, which (since it is very similar to the vOICe) we called Vox.

Somewhere between the choice of substituting modality and this fine-tuning, there are important decisions to be made about the way in which one information from sense is translated into the primitives of another. These decisions are best illustrated by the current diversity of visual-to-audio SSDs.

3.2.1 Current SSDs

The aforementioned vOICe SSD scans each image from left to right over the course of a second, sonifying the image as series of columns of pixels (Meijer, 1992). Each pixel takes a frequency based on its height in the column and amplitude based on its luminance. In this way, a single white line at a 45° angle against a black background will sound like a single rising tone.

Conversely, the Prosthesis for Substitution of Vision by Audition (PSVA) uses pitch to represent space in both dimensions, with vertically adjacent pixels being different by an octave. This means that it is able to operate in real-time, whereas the vOICe only presents a new soundscape once a second. The PSVA is also noteworthy in that it mimics the retina by having a higher density of pixels in a central "fovea" (Capelle et al., 1998).

The Vibe also operates in real-time, but rather than using pitch to encode both spatial dimensions, it relies on audio panning to encode horizontal position. Like the vOICe and the PSVA, vertical position is encoded by pitch and luminance is encoded by amplitude (Auvray et al., 2005).

These three devices differ primarily in the way they handle visual horizontal position in the auditory domain. They all map the same primitives of vision: vertical pixel position, horizontal pixel position and pixel luminance. The auditory primitives these are mapped onto vary little: loudness, frequency, time (the vOICe) and stereo balance (the Vibe). Pixels are the basic primitive of computerised images, which may explain the predominance of systems which relay on pixel-based information. Yet although these three devices are the most studied, audio SSDs need not be limited to these primitives: the SmartSight SSD, for instance, segments visual scenes into oriented lines and uses musical primitives as a means of output (Cronly–Dillon et al., 2000; Cronly-Dillon et al., 1999).

The mere fact that all these systems appear to work to some degree says something fundamental about the plasticity of the brain, but also raises questions about which approach is "best". What makes this judgement impossible is the huge variance in the needs of the users. In this paper, we argue that the best way to maximise the utility of SSDs is to put the control of these decisions into the hands of the end users. We outline why SSD users are likely to be suitable candidates for "End User Development" (EUD). We then introduce Polyglot (Mark II) – a general-purpose framework for visual-to-auditory SSDs.

3.2.2 End-user development

Within the field of computer science, "end-users" are those people for whom a piece of software was written. Typically this designation is intended to contrast against the developers of the software. "End-user development" (EUD) may therefore seem initially to be somewhat of an oxymoron, yet is something that the reader likely engages in on a regular basis.

End-user development activities range from customization to component configuration and programming (Fischer, Giaccardi, Ye, Sutcliffe, & Mehandjiev, 2004). Microsoft's Excel

spreadsheet program is a classic example. Within Excel one can use in-cell functions to achieve programmatic outcomes. On top of this, Excel has a domain-specific version of Visual Basic, which can be used to write macros that automate tasks. Matlab (*Matlab*, 2013) and E-prime (W Schneider, A Eschman, & A Zuccolotto, 2002), used to create psychology experiments, are also examples of EUD environments with domain-specific languages.

In essence, products that enable end-user development may be thought of as comparable to construction toys such as Lego. The manufacturer produces the components and may even publish guides for the construction of particular configurations. In the same as a child is free to arrange Lego bricks in any combination to arrive at the desired model/toy, so may the user of a spreadsheet use the provided formulas to analyse experimental results, keep track of expenditure, or prepare a tax return.

The value of EUD is rapidly becoming realised in a variety of different settings. In the corporate environment: “The empowerment of end users to tailor their applications will render appropriation processes more effective and thus lead to more economical IT investments” (Wulf & Jarke, 2004). Among users with domain-specific expertise (e.g. surgeons), EUD allows for equal participation alongside software engineers (Costabile, Fogli, Mussio, & Piccinno, 2007).

3.2.3 The users

There are several factors that make SSD users ideal candidates for EUD.

Firstly they are all, by definition, “early adopters” – people who are keen to seek out and experiment with new technologies. They have all, in the absence of any institution promotion or support, sought out an SSD and taught themselves to use it. Moreover, to use an SSD is not comparable to using, say, word processing software – as a skill, SSD use has a significant non-declarative component, which consequently makes the experience fundamentally more personal. SSD users are almost uniformly enthusiastic about the technology and enjoy tinkering with existing systems as well as theorising about possible systems. On the mailing list of the vOICe user group (“The world’s largest sensory substitution network”; Meijer, 2013) for instance, it is not uncommon to see requests for new options, suggestions about alternate methods of implementation, or reports of experiments with new types of camera.

Secondly, the needs of visually impaired people can vary wildly. At one level, visual impairment has a high co-morbidity rate, which means that some users will have another sensory impairment or some restriction on their ability to control a device. Beyond physical

restrictions, some users may prefer one mode of operation over another (e.g. passive scanning instead of touch-based spot sampling) or simply want to experience a different aspect of their environment (e.g. colour instead of luminance). These needs and desires are not something that visually impaired people are passively subject to. Users of SSDs, and visually impaired people more generally, are the unequalled experts in their requirements. Within the healthcare literature, it is widely acknowledged that involving patients (or, in the parlance of assistive technology research, users) in the development of medical technology leads to better outcomes (Bridgelal Ram, Grocott, & Weir, 2008; Shah & Robinson, 2007).

Finally, contrary to potential preconceptions, visually impaired people are not at a disadvantage with regards to computer programming. Naively, people often assume that programming is a very visual task, but text on a computer screen is actually one of the easiest visual forms to make accessible. On StackOverflow, a popular question and answer site about programming, three blind programmers responded to the question “How can you program if you're blind?”

"I am a totally blind college student who's had several programming internships," says Jared. "I usually rely on synthetic speech but do have a Braille display. I find I usually work faster with speech but use the Braille display in situations where punctuation matters and gets complicated. Examples of this are if statements with lots of nested parenthesis's and JCL where punctuation is incredibly important." – (Jared, 2008)

Saqib writes "I'm blind, and have been programming for about 13 years on Windows, Mac, Linux and DOS, in languages from C/C++, Python, Java, C# and various smaller languages along the way. [...] I personally use Visual Studio 2008 these days, and run it with very few modifications. I turn off certain features like displaying errors as I type since I find this distracting. Prior to joining Microsoft all my development was done in a standard text editor like Notepad, so once again no customisations." – (Saqib, 2009)

"I am blind and have been a programmer for the last 12 years or so," writes Mannish. "Have recently been playing around with python, which as other people have noted above is particularly unfriendly for a blind user because it is written using indentation as the nesting mechanism. Having said that, NVDA, the most popular open source screen reader is written completely using python and some of the committers on that project are themselves blind." – (Manish, 2011)

The three quoted programmers casually make reference (in their unabridged answers) to specially adapted programming tools and screen-reader plug-ins for integrated development environments (IDEs). Note also that Manish mentions that NVDA is open source (i.e. created and maintained by volunteers) and that some of the contributors are blind. Programming is not something that visually impaired people are excluded from – far from it. Nor is it merely accessible. Rather, programming takes place on a level playing field and empowers the visually impaired by providing the means to create tools to further its own accessibility.

Fischer and colleagues argue that, in a corporate environment, "the spread of EUD depends on a fine balance between user motivation, effective tools, and management support" (Fischer et al., 2004). In the context of SSDs for visually impaired people, we might interpret this to mean that the success of an EUD project depends on user motivation, effective tools and a supportive online community. Hopefully, the motivation of the users is now beyond doubt. The rest of this paper will primarily consider the tools, but will also touch on the potential means to sustain an online community.

3.3 The Polyglot Framework

The Polyglot framework is written using Microsoft's C♯ programming language and their ".NET" framework. Several factors influenced the decision to adopt these technologies. Firstly, we decided to target the Windows operating system, as it enjoys that largest market share among sighted and (seemingly) visually impaired users. Secondly, the .NET framework is well served by a large range of well-supported multimedia libraries such as "NAudio" and "AForge". Finally, the Managed Extensibility Framework offers unparalleled support for the modular interoperability at the heart of Polyglot.

In order to facilitate EUD, the Polyglot Framework offers three levels of customisation by the end user. Most superficially, each module may offer configurable options. More interestingly, Polyglot allows the user to "mix and match" modules to construct their own system. Finally, Polyglot is open source and has been built with extensibility in mind. If the end user is so inclined, they are able to leverage the provided support structures to develop entirely new modules. These modules in turn could be made available for other users to include in their compositions. This tiered approach to EUD mirrors the "gentle slope" of tailoring: by customization, then integration, and finally extension (Mørch et al., 2004).

3.3.1 Polyglot modules

The modular system underpinning Polyglot is based on the widely accepted general description of an SSD: comprising a sensor, a coupling system and a stimulator. In the parlance of the Polyglot Framework, these are referred to as "acquisition", "transformation" and "output" modules. Acquisition modules are responsible for obtaining an image and passing this data to the Transformation module. The Transformation module then must pass on data to the Output module.

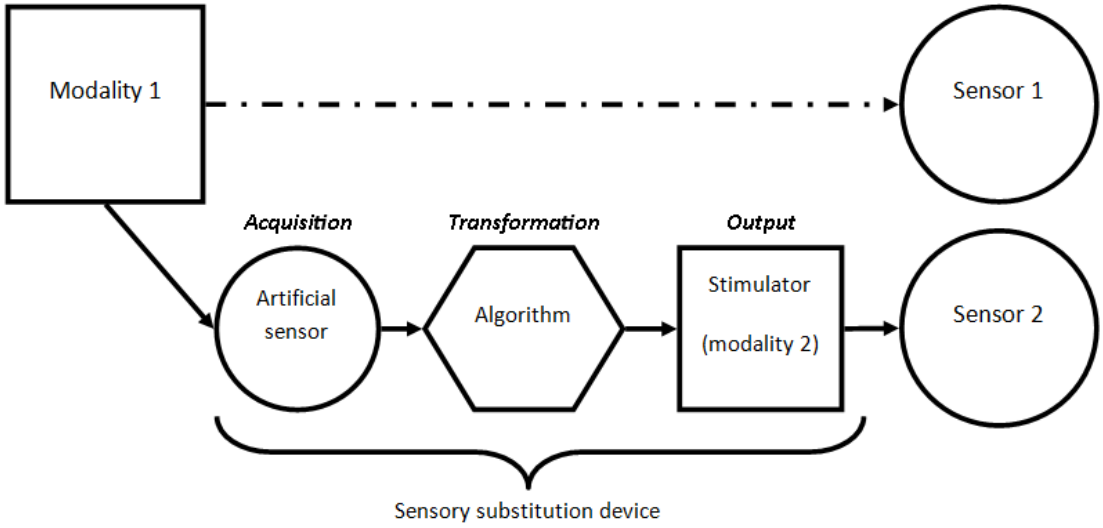


Figure 17: General schema for SSDs, incorporating module types

In order to abstract out some of the behaviour typically accomplished by the coupling system, the Polyglot framework additionally features a class of module responsible for location in the image: the Pointer module. The transformation engine is free to ignore the information supplied by this module, to use either the vertical or horizontal components or to use the full positional information. For example, a configuration based on the Vibe would use the full image at all times and would therefore not require any positional information.

These categories are listed below, along with some example modules for each:

3.3.1.1 Acquisition

Acquisition modules, referring to the sensor of a classically represented SSD, are responsible for acquiring images for conversion. The core Polyglot Framework package comes with three bundled acquisition modules: a test swatch (made up of coloured and monochrome patches), a form for selecting a still image, and a module which reads live video from a webcam connected via USB. Additionally, we have developed a module which captures the current contents of the screen.

3.3.1.2 Pointer

Pointer modules, which have no equivalent in the classic SSD model, determine which location in the image (if not all of it) are currently the focal point. Two such modules are bundled with Polyglot: “CentrePoint” and “ScanningPoint”. The first of these is very basic, and never moved from the centre of the image. This functionality matches an approach commonly seen in colour SSDs (e.g. Capalbo & Glenney, 2009). ScanningPoint moves the focal point horizontally along the mid-line of the image. This behaviour could be the basis for using Polyglot SSDs to emulate

the scanning behaviour of the vOICe; the transformation module would disregard the vertical position and read a column of pixels from the horizontal position.

We have also developed a “TouchPad” module, which interfaces with an Android mobile device running a special application (also available for download) and connected via USB. In addition to the modules we have already developed, it is possible to imagine pointer modules driven by mouse movement or even an eye-tracker. Using a mouse-based pointer module whilst and the screen capture acquisition module would result in functionality similar to the magnification tools often available in assistive software suites. Using an eye-tracker to feed an acquisition module might be of particular benefit to those with partial vision (e.g. scotoma), as a substituted signal could be made to “fill in” the damaged region.

In a deviation from convention, we have also added a single output channel to the specification for the pointer module. Individual pointer modules remain free to dump this data and transformation modules are not obligated to supply it. It can be used however, to create crude audio/tactile hybrid devices. The TouchPad module, for instance, takes this data and correspondingly actuates the vibrating motor in the mobile device. Some computer mice are also support haptic feedback. Alternatively, this output may be used as a convenient monitor channel during experiments.

3.3.1.3 Transformation

Transformation modules are the primary contributor to the unique behaviour of each SSD. These modules correspond to the “coupling system” in the classic model of SSDs. We have implemented two transformation modules.

The first, “VoiceColumn”, takes the horizontal component of the focal point and extracts information from that column of pixels. This column is then sonified such that pitch maps to vertical position increases and amplitude maps to the lightness of each pixel. (For an in-depth description of this mapping, please refer to descriptions of the vOICe/Vox in chapter two.) When coupled with the ScanningPoint module, the behaviour of the resulting SSD is comparable to that of the vOICe/Vox.

The second transformation module bundled with Polyglot is called “MusicalSpot”. This module gets the Gaussian weighted average (with a radius of 5) of the pixels at the focal point. It then extracts the levels of the Red, Green and Blue (RGB) channels and maps these on to the amplitude of three notes (C5, E5 & G5) which form a chord (C major). When combined with the ScanningPoint module, this just gives information about a narrow strip across the middle

of the scene. When paired with the TouchPad module though, the user is able to explore the colours in an image in real-time with their fingertip. This module would be easy to customise (by changing the notes) or expand (perhaps by introducing the ability to specify the notes used at run time).

3.3.1.4 Output

Output modules, corresponding to the stimulator in the classic model, simply generate sounds. They are required to be able to generate a collection of sources, each with a frequency, amplitude and azimuth. Whilst this does not leave much scope for innovation, the Polyglot Framework is supplied with two example modules that demonstrate different approaches to representing azimuth information. The first of these acts exactly like the vOICe: complete stereo panning. When the azimuth is at -90° the amplitude of the left channel is set to 100% and that of the right to 0%. When the azimuth is ±0° both channels are set to 50%. When the azimuth is at +90° the right channel is set to 100% and the left to 0%.

The second module attempts to simulate naturalistic sound localisation. To do this, it incorporates two interaural cues: Interaural time difference (ITD) and interaural level difference (ILD). These are the two primary cues used by humans to localise sounds on the horizontal plane (Grothe, Pecka, & McAlpine, 2010). This produces a much more subtle effect, but one that should be more intuitively spatial, since it leverages our natural ability to localise sounds. ITD is introduced by differentially delaying samples by up to two tenths of a millisecond (the time it takes for sound to travel from one side of the head to the other). ILD is generated by adjusting the amplitude of each frequency according to the level of head shadowing that would normally occur (this varies non-linearly between frequencies).

3.3.2 Runtime composition

A modular system as described in the section above would already meet the objective of allowing code to be reused between systems. In its simplest form however, this would require a developer to recompile the code in order to switch modules. This is a barrier to most users.

In order to encourage EUD in all users, Polyglot supports “runtime composition”. This means that when the Polyglot engine launches, is scans its own code and any 3rd party code in its “modules” directory to discover which modules are available for use. The user is then able to select a module for each of the 4 module categories.

These selections (and the configuration parameters of each module) can be stored in a “Polyglot Configuration File”. These files are XML based and are fully portable (i.e. they do not

contain references to a particular computer, user or composition program). This means that it is possible for the Polyglot engine to be implemented by multiple “front-ends”. The core Polyglot project contains two graphical front-ends to demonstrate this. The first is the Polyglot Composer, which allows the user to view and select each module and save the configuration. The second is the Polyglot Player, which is stripped down and simply reads a configuration file and immediately launches. (Both graphical front-ends were built using the standard “Windows Forms” libraries and utilise all of the accessibility features that these libraries provide. This should ensure full accessibility to visually impaired users in conjunction with all mainstream accessibility tools.)

The interoperable nature of this modular architecture means that modules can be written by any developer and used by any Polyglot user in conjunction with any other modules. Given that users of the vOICe already enjoy a community, it does not seem farfetched to envisage a community driven “App Store” of user generated modules. At some stage in the future, it would be possible to build this functionality into the Polyglot Composer.

3.3.3 Technology

This modular functionality is supported by the Managed Extensibility Framework (MEF) composition layer of the Microsoft .NET framework. The MEF requires that each component implements an “interface”. Interfaces are a .NET means of non-exclusive inheritance; in practical terms a set of obligations to provide certain methods and properties.

Polyglot uses a base “IModule” interface, which is inherited by “IActivatableModule” and “IOutput”. IActivatableModule is in turn inherited by “IAcquisition”, “IPointer” and “ITransformation”. (Only IAcquisition, IOutput, IPointer, and ITransformation are intended to be used directly: IModule and IActivatableModule are simply common ancestors to reduce code duplication.) IOutput is not treated as an “activatable” module like the others, but this is only due to the idiosyncrasies of signal generation: instead of “Activate” and “Deactivate” methods, “Play” and “Stop” are provided.

As well as being “activatable”, Acquisition and Pointer modules may be set to run in either “active” or “passive” mode. When in active mode, these modules will raise an event each time they have new data. In passive mode, on the other hand, these modules will wait for the transformation module to request new data.

The base IModule interface specifies a requirement for a module name and a unique ID string. The former ensures that the user is easily able to select the desired module from a list during

composition. The latter is used internally to identify the module. This separation is designed to prevent collisions in the event of two modules sharing a name. The use of a Universally Unique IDentifier ("UUID") means that collisions between IDs are highly unlikely. Looking forward, one can imagine using these IDs to facilitate automatic updates from a central "App Store" for modules.

3.3.4 Support structures

For those users able and willing to undertake some programming, the Polyglot Framework has been furnished with many support structured in order to simplify development. These include utility classes, third party libraries, custom events and exceptions, and toolboxes for some common groups of functions.

3.3.4.1 Utility classes

Three utility classes are included in the Polyglot framework: "ModuleForm", "TransformationBase" and "ProportionPoint". The first two are abstract base classes (i.e. they needs to be inherited and expanded upon to do anything useful). The first provides basic graphical user interface (GUI) functionality for modules. Module authors therefore do not need to start from scratch if they wish to display information to the screen. The second handles some of the more arcane aspects of the transformation process, so that a developer of a transformation module may focus on the unique behaviour that they wish to produce.

ProportionPoint is a little different, as it underpins the fundamental approach to positional data in Polyglot. Rather than use absolute horizontal and vertical coordinates, Polyglot employs relative values. This means that the absolute width and height of an image need not be known in advance, which means that acquisition modules can be more permissive in what they accept. The ProportionPoint class is simply a pair of floating point numbers corresponding to horizontal (X) and vertical (Y) position, constrained to have values between 0 and 1 inclusive.

3.3.4.2 Third-party libraries

Although module authors remain free to use whichever code libraries they wish, several are bundled with the Polyglot Framework. These have been used during the creation of the supplied modules and have been found to work well.

For capturing images from webcams, the "AForge" library (Kirillov, 2012) works well. The core and "Video" components are bundled with Polyglot, but other components are available that offer image filters ("Imaging") and computer vision routines ("Vision").

To generate sounds, the “NAudio” library (Heath, 2012) is bundled. NAudio is unusual among audio libraries in that it is natively and deeply compatible with .NET (and consequently C#) and also in that it is enormously flexible. This functionality has been exploited by a bundled toolbox, which uses psychometric data to realistically simulate spatial localisation. See below for more.

For advanced mathematical functionality, the “Math.Net Numerics” library (Ruegg & Cuda, 2012) is bundled. Finally, in order to make microsecond precision time measurements, the “MicroTimer” library (Loveday, 2013) is also bundled.

3.3.4.3 Events

In C♯ (as in many other object oriented languages), instances of classes (i.e. modules) are hierarchical. Class A may own a child Class B. In this case, Class A may instigate communication with B by calling one of B’s methods. In order for Class B to instigate communication with A, it must raise an event to which A subscribes.

In the Polyglot framework, the transformation module is the parent of the acquisition, pointer and output modules. In order to support the active modes of the acquisition and pointer modules, we therefore need a set of events. Firstly, the “NewImageEvent” allows the acquisition module to notify the transformation module of new image data. Its event arguments – “NewImageEventArgs” – carry the new image as a payload. The “NewPositionEvent” and “PointerStateChangedEvent” allow the pointer module to notify the transformation module of new coordinates and changes to the pointer state, such as a finger being lifted from a touchscreen. The events both have their own event arguments: “NewPositionEventArgs” and “PointerStateChangedEventArgs”, which convey the new data inside the notification.

Lastly, “ModuleFormClosedEvent” allows modules with graphical interfaces to react to their forms being closed. In this situation, the developer may decide to re-spawn a form, or instead to close stop processing and close the application.

3.3.4.4 Exceptions

Exceptions are a way of handling unexpected or unsupported behaviour without the program crashing. EUD and run-time composition present a particular challenge to robust programming; interoperability is hard to guarantee and users may attempt to use modules in ways the developer did not imagine. Accordingly, Polyglot offers several exception types, which complement those already built into the C♯ language.

Firstly, “ModeNotSupportedException” offers a way to gracefully handle occasions when an acquisition or pointer module is asked to enter a mode (passive / active) that it does not support. Generally, these modules should support both modes, but this is not always possible. In the case of a static image loaded from a file, for instance, it is not clear how an active mode should behave.

Secondly, “NoDataYetException” is to be raised when the transformation module requests data from the acquisition or pointer modules before they have been able to collect their initial data.

Thirdly, “ModuleImplementationException” is a general exception to be used when there are problems in the way a module is behaving. Currently this is only called when the Polyglot Engine is unable to compose an SSD using the specified modules. It is possible to imagine this exception also being used in modules to report unexpected situations.

3.3.5 Toolboxes

In manipulating visual information, there are some tasks which come up time and time again. In order to reduce the need for module authors to reinvent the wheel, a collection of routines are included with Polyglot in “toolboxes”.

3.3.5.1 PixelTools

The first of these toolboxes is “PixelTools”. This contains methods for reading images and extracting basic information. The method called “GetBytesFromImage” reads a bitmap image into an array of bytes. This is a useful first step for any image processing, as the built-in C# tools for reading pixels directly from a bitmap are very slow.

The “GetGaussianPixel” method of PixelTools applies a Gaussian average to a group of pixels centred on a particular ProportionPoint (see above). For a given radius, this method first segments the image, and then calculates the Euclidean distance between each pixel and the origin (as determined by the ProportionPoint). The normal distribution class of the 3rd party Math.Net Numerics library is then used to give each pixel a weighting based on its distance. This produces weightings like those illustrated in Figure 18.

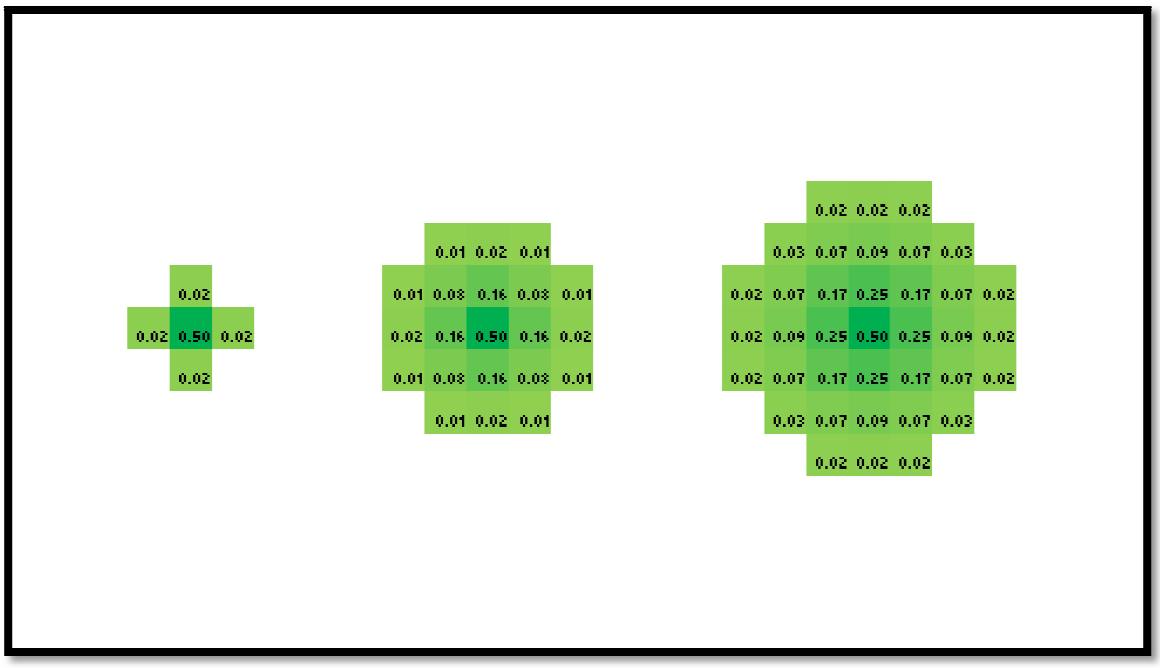


Figure 18: Gaussian weightings for radii of 1, 2 and 3 pixels

In our own experimentation (and that of members of our laboratory), we have found that applying this style of weighted average serves to reduce the amount of noise present in an image, which can lead to distracting “blips” in an audio signal. The utility of this depends strongly on the task at hand: when trying to detect boundaries, a weighted average can make a hard line appear more like a gradual transition. Consequently, there will be times when one simply needs the value of single, un-weighted pixel. For this, GetGaussianPixel has a special case – a radius of 0. For simplicity, this special case is wrapped by its own method: “GetPixel”.

Each pixel, even those constructed from a weighted average, is returned as a C# Colour “struct” (structure; a type of object). These structs consist of 4 bytes, one for the alpha channel (unused by Polyglot) and one for each of the Red, Green and Blue (RGB) channels.

3.3.5.2 ColourTools

The primary contribution of the ColourTools toolbox is to support HSL colour values, as well as the accompanying conversion routines. An HSL colour value consists of a value for Hue, a value for Saturation and a value for Lightness. These values are arguably easier to understand than Red, Green and Blue channels (Berk, Kaufman, & Brownston, 1982; cf. Schwarz, Cowan, & Beatty, 1987). Though its components map only crudely onto perceptual aspects of colour, HSL space is used because it retains a regular shape – unlike the Lab or CIELUV colour spaces. In practice, this means that all within-bound coordinates will return a visible colour.

In order to facilitate use of the HSL colour space, the ColourTools contains an HSL struct, which contains floating point values for hue, saturation and lightness. It also contains a method called RGB2HSL, which converts RGB colour points into their HSL equivalents. The toolbox also contains convenience methods for extracting the hue, saturation and lightness from a supplied RGB colour point.

3.3.5.3 Lenses Toolbox

The Lenses Toolbox provides components for more refined weighting schemes than PixelTools. In doing so, this toolbox allows for easy implementation of the sort of weightings used by systems such as the PSVA (Capelle et al., 1998) and the Vibe (Durette et al., 2008). Significantly, this toolbox allows for these weightings to be abstracted out of transformation modules, thereby providing a further level of interoperability.

"Lenses" consist of a collection of "Facets", with rules for mapping pixels onto them. Facets are simply collections of these mapped pixels, with added positional information and an associated audio frequency. As all Lenses must implement the "iLens" interface, they are interchangeable. The iLens interface simply requires that each Lens have a method called "ReadImage", which takes an image and a proportion point and returns a list of Facets.

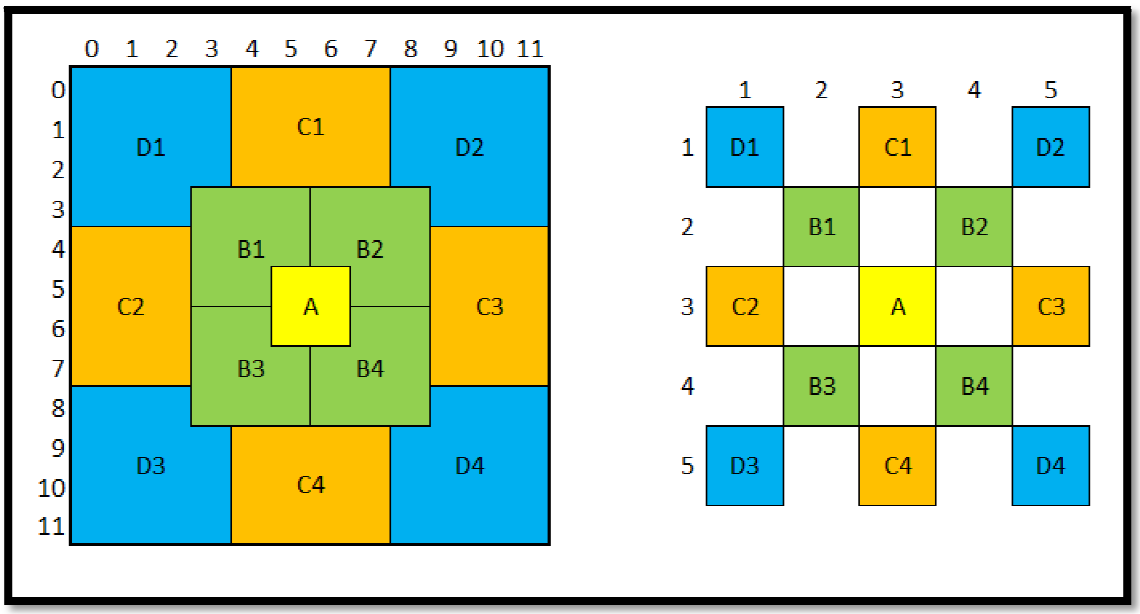


Figure 19: an example of a weighted lens, with 13 Facets (left) and the virtual position to which they are re-mapped (right)

This approach to re-mapping pixels is very flexible. Figure 19 shows an example mapping. The figure on the left shows the central region of an image which has been divided up into regions of varying size (the Facets). The amplitude of each Facet is the average of the amplitudes of

the pixels allocated to it. The figure on the right illustrates how these Facets are modelled in the output. In this example, they are arranged such that the resulting audio signal has 5 frequencies (D1, C1, D2; B1, B2; C2, A, C3; B3, B4; D3, C4, D4) and 5 horizontal positions (D1, C2, D3; B1, B3; C1, A, C4; B2, B4; D2, C3, D4). In other words, 144 pixels are re-mapped to 13 Facets that exist at points within a 5 x 5 grid. Because of the flexibility of this approach, it would be trivial to (for instance) remove the C group of Facets and enlarge the D group Facets to cover their pixels.

Since Lenses created with this toolbox allow for retina-like weightings, they would be highly suited for active exploration of a scene. Achieving an appropriate balance of fine detail and context is a matter for further empirical research, but the Lenses toolbox provides the means for this experimentation to be built into a transformation module.

3.4 Example systems

In the process of developing the Polyglot framework, several SSDs have been built which utilise its design. In fact, the ease with which new SSDs can be composed means that a vast number have been explored. Some of these combinations are of greater interest than others and are revisited with sufficient regularity to be afforded names of their own. Here, we present two of those systems. The first (“Vox”) shows that the core mechanic of an existing device can be replicated using the framework. The second (“Creole”) demonstrates the ability to employ Polyglot to easily design a radically new device.

3.4.1 Vox

Vox largely re-implements the vOICe and is described in detail in chapter two (and in Wright & Ward, 2013). Like the vOICe, it uses the Webcam module in passive mode and the most basic audio Output module. The scanning behaviour of the vOICe is replicated using the ScanningPoint Pointer module in active mode. The Transformation module then reads a column of pixels using the horizontal component of the data from the Pointer module. This column is represented with tones increasing in frequency with vertical position and in amplitude with brightness. Once the ScanningPoint module has completed a full sweep, a new image is requested from the Webcam module.

Like the vOICe, this results in one second “soundscapes” which form a sort of flick-book representation of the video captured by the webcam. A soundscape of a white diagonal line against a black background will sound like a single rising (or falling, depending on the orientation) note. Natural visual scenes are inevitably much more complex.

Vox is used to produce the stimuli presented in the eye-tracking study reported in chapter four. Additionally, Vox has been used in experiments investigating the localisation of small light sources.

3.4.2 Creole

Creole is a demonstration of a much more enactive form of SSD use. Unlike the vOICe / Vox, Creole presents a spatially smaller amount of information, but in real time instead of as one second soundscapes. The portion to be sonified is determined by the location of the users finger on an Android touchpad.

To achieve this effect, Creole uses the Webcam in active mode. Each time a new image is made available, the Transformation module requests a new set of coordinates from the Touchpad module. Using this information, it transforms a region centred on the user's finger using a map of Facets and passes the luminance of these weighted sources to the fully localised audio output. Additionally, the overall luminance is passed to the vibration function of the Touchpad. The result is a sound which is highly responsive to changes in the visual scene and movements made by the user's finger. This responsiveness makes it very intuitive to naive sighted users and it is likely that visually impaired users would find it equally intuitive. Whether early-blind people (who will have grown up relying on touch) find Creole differentially intuitive is a particularly intriguing question.

Creole, and variants that use the outputs to represent attributes of colour, have been used extensively in experiments that seek to explore the relative merits of each dimension of colour in different circumstances.

3.5 Discussion

3.5.1 Facilitating EUD – future work

Whilst it seems very likely that SSD users will explore any provided settings, adoption rates for the two other forms of EUD are hard to predict. In terms of exploring new compositions of modules, adoption might be expected to depend on the availability and quality of modules to experiment with. The future development of a central integrated "app store" of modules could boost this process: descriptions and ratings of modules would boost user confidence and being integrated into a Polyglot UI would eliminate any logistical friction.

With regards to encouraging users to engage with the third form of EUD available in Polyglot, there is much work to be done. The framework described above would allow a competent

programmer to produce a new module, but providing a code framework is really the bare minimum. Projects that aim to facilitate EUD typically provide some form of development environment. As it stands, end users would need to install 3rd party software in order to author a new module. Microsoft do provide a free version of their C# integrated development environment (IDE) called "Visual Studio Express", but this doesn't change the fact that downloading and installing 3rd party software will always add frustration to the EUD process.

In place of a full-fledged (and inevitably programmer-oriented) IDE, some EUD researchers advocate the concept of a "Software Shaping Workshop" or "SSW". (This name refers to the inspiration provided by physical artisanal workshops, rather than a seminar-like event.) Rather than providing an environment for programming, an SSW provides the necessary tools for EUD (Costabile et al., 2003). In the case of Polyglot, an SSW might feature graphical interfaces for the available toolboxes. Specifically, a graphical designer for the lenses toolbox could make development of lenses dramatically simpler.

An additional barrier to development is currently the lack of support materials or community. To foster module development, the code should be augmented by an online tutorial and a community generated "knowledge base" of tips and tricks. Forums to discuss things like programming problems or best practice would provide a virtuous circle of engagement. That all these non-code support structures are currently lacking means that, unfortunately, the development curve is likely to be slow. Further work should therefore seek to provide these resources.

These barriers unfortunately exist in a context which remains hostile towards all but the most dedicated explorers of assistive technologies. Without institutional advocacy and support, use of sensory substitution devices (let alone EUD) will remain rare among visually impaired people. For visual impaired people to benefit from both SSDs and EUD, healthcare professionals (and volunteers such as the RNIB's "Technology Support Squad") will need to be exposed to the technology and persuaded of its merits.

3.5.2 Porting to other platforms

Another possible avenue for future development is the porting of Polyglot to operate on other platforms, such as Mac OS X or Linux. Windows was chosen as the development platform because of its popularity and the variety and maturity of commercial accessibility software available to use with it. However, not all visually impaired people use Windows. Despite early misgivings (Leventhal, 2005), the screen-reader that Apple bundle with Mac OS X ("VoiceOver") is now deemed to be a "viable screen reading option" (Denham, 2008) and

Apple seem committed to furthering the accessibility of their platform. With regards to Linux, the most significant piece of accessibility software for the visually impaired is probably Emacspeak, which was developed by visually impaired developer T. V. Raman. Emacspeak “is a powerful tool for those who are willing to learn it, but a frustrating experience for beginners” (Sajka, 2003).

Luckily, an open source project called Mono provides a cross-platform implementation of the .NET Framework. This implementation is “binary-compatible”, meaning that code compiled using Microsoft’s Visual Studio should run under the Mono interpreter. Unfortunately, the multimedia libraries used by Polyglot (NAudio and AForge; see above) rely on low-level Windows Application Programming Interfaces (APIs) and consequently do not port well to other platforms. This means that any Mac OS X or Linux implementation would require new multimedia libraries to be found. Due to the modular nature of Polyglot, these new multimedia libraries would only require new acquisition and output modules – not a re-write of the entire framework.

3.5.3 Potential implications on the research community

Aside from the obvious implications for end users, Polyglot has the potential to advance SSD research. As a research tool, Polyglot represents a versatile means of systematically evaluating approaches to sensory substitution. By virtue of the fact that it is possible to change very specific aspects of its operation, Polyglot allows for fine-grained exploration, such as that which we undertook using Interactive Genetic Algorithms (Wright & Ward, 2013). At the other end of the scale, Polyglot also supports the rapid switching of large blocks of functionality, thereby enabling comparison of more fundamental aspects of a device’s operation

As well as these primary benefits, widespread adoption of Polyglot could cause a group of visually impaired user-developers join the research community. Given the fact that many of these user-developers will be using an entirely idiosyncratic configuration of Polyglot, incorporating them into the research process may prove problematic. Training these volunteers to produce useful phenomenological accounts could be an approach to deal with otherwise incomparable glut of data. Further, encouraging these volunteers to share and compare their configurations may prove to be a fruitful method for “crowd-sourcing” comparative accounts.

3.5.4 Conclusion

We have shown that end users may be the best placed group to spearhead the development of novel visual-to-auditory SSDs and outlined a mechanism by which this may be

accomplished. The Polyglot Framework provides three forms of opportunity for End User Development (EUD): tailoring module options, mixing and matching modules to form novel devices, and creating entirely new modules. Future work should focus on developing the non-code infrastructure, so as to best foster a community of user developers. With sufficient support, this community could simultaneously develop practical tools for themselves and dramatically accelerate SSD research.

Chapter 4: The presentation of relevant visual backgrounds facilitates localisation of targets presented by means of an auditory sensory substitution device

4.1 Abstract

In a novel experimental paradigm, participants are simultaneously presented with a visual background image and a composite of the background image and a visual target via a sensory substitution device. Whilst participants seek to identify the location of the visual target (which is present only in the substituted stream), their eye-movements are tracked. Comparisons with control groups (who weren't shown the background images) reveal that a partial visual scene facilitates performance in the sensory substitution task. Moreover, the eye-tracking data reveals that gaze is a better predictor of target location than the behavioural responses. Finally, because of the differential approaches to encoding, comparing horizontal errors with vertical errors suggests that more natural spatial cues may offer both a performance and a learning boost. These results have implications for future sensory substitution device design. Sensory augmentation device designers will be encouraged by the apparent meshing of natural vision with substituted vision.

Preliminary results from this study were presented as a poster at the 13th annual meeting of the International Multisensory Research Forum (Wright et al., 2012).

4.2 Introduction

Sensory substitution is a term for processes in which one sensory modality is represented in another modality. Typically, the substituted modality will be vision; the majority of sensory substitution devices (SSDs) are created for people with visual impairments and normally use either touch or hearing as the substituting modality. Sensory substitution is a growing field, but has been the focus for active study for more than four decades. A more detailed account of sensory substitution may be found in earlier chapters (e.g. chapters 1 & 2) or in our recent review (Ward & Wright, 2014).

One of the key theoretical questions surrounding sensory substitution is the degree to which the signal may be classified as belonging to the substituted sense rather than the substituting sense. How visual, for instance, is the signal emitted by a visual-to-auditory device? Should the output of an SSD be treated as auditory because it stimulates the ears, or as visual because changes in the signal obey the sensorimotor rules of vision?

The question at the heart of the present study is closely related – how, if at all, does substituted vision interact with natural vision? That this question has been neglected until now is hardly surprising. Sensory substitution research tends either to be concerned with the interesting theoretical issues raised by replacing vision, or with the important practical development of devices for the blind.

Yet studying the interaction between natural vision and substituted vision (if vision it is) also offers the chance to explore both theoretical and practical questions. Firstly, blindness is not a binary state: many legally blind people have some residual vision. Moreover, potential beneficiaries of sensory substitution technology may exist at many points on the spectrum of visually impaired people.

Secondly, in sighted people there are a number of potentially interesting ways to extend the visual field. This could be done spatially (e.g. eyes in the back of the head) or by adding formerly non-visual information (e.g. by superimposing thermal imaging on vision, as described in a patent granted to Havey et al., 2007). The addition of novel sensory information via an existing sensory pathway is known as “sensory augmentation” and differs from sensory substitution in that it does not seek to replace a lost sense. In terms of both sensory substitution and sensory augmentation, exploiting the novel information is likely to be more easily accomplished if the user can use natural vision as scaffolding.

Lastly, exploring how natural vision interacts with substituted vision is likely to inform us a great deal about the nature of both. If natural and substituted vision do interact, how does this take place? Exploring this interaction may shed light on visual processing, multisensory integration and the way in which spatial stimuli are learnt and encoded.

We used a visual-to-auditory SSD called Vox. Vox is based on a more established SSD called "the vOICe" (the central capitalised letters spelling out "Oh, I see!"), which was developed for visually impaired people (Meijer, 1992). Both Vox and the vOICe use a scanning, one-dimensional array of tones to represent a greyscale version of an image. They sweep over an image, from left to right, over the course of 1 second. Vertical position is represented by pitch. Horizontal position is represented by time (due to the sweep) and with stereo panning. Lightness is represented by loudness. Despite these underlying similarities, the two devices differ in terms of the specific parameters used to achieve the conversion (Wright & Ward, 2013).

Visual-to-auditory SSDs (like Vox) are capable of facilitating visual tasks in blindfolded sighted participants without extensive prior training (e.g. Renier, Laloyaux, et al., 2005; Brown et al., 2011). Moreover, sensory substitution use by this class of participant has been shown to recruit brain regions associated with vision (Poirier, De Volder, Tranduy, & Scheiber, 2007; Renier, Collignon, et al., 2005). These effects occur far too quickly to be explained by any significant changes to cortical organisation and therefore must be explained using mechanisms available to our organic sensory systems. Since the stimulus is physically an audio signal and received by the ears, but results in the functional recruitment of normally-visual areas of the brain, our explanation likely lies in audio-visual multisensory processes.

4.2.1 Mechanisms

The ways in which sound and vision interact to shape our perception of the world are many and varied. Many people will be familiar with the illusion of ventriloquism, in which a sound source appears to originate from a synchronous visual target despite a spatial separation (Choe, Welch, Gilford, & Juola, 1975; Jack & Thurlow, 1973). Here the visual stimuli change the perception of the auditory perception, but the reverse is also possible. In the double-flash illusion, one visual flash can be made to appear as two (or more) when accompanied by multiple auditory beeps (Shams, Kamitani, & Shimojo, 2002). In the stream/bounce illusion a sound causes a dot to bounce off another, rather than pass behind it (Bertenthal, Banton, & Bradbury, 1993). It is also possible for conflicts and ambiguities to be resolved by means other than deferring to one sense. Audio-visual mismatches in speech, for example, can result in

synthetic perceptions, a phenomenon known as the McGurk illusion (McGurk & MacDonald, 1976).

The aforementioned illusions are all perceptual effects, but audio stimulation also has attentional effects on vision, as measured using gaze. Auditory stimuli (e.g. a beep) are not only capable of driving oculomotor saccades (Zambarbieri, Schmid, Magenes, & Prablanc, 1982), but can interact with visual stimuli to decrease Saccadic Response Times ("SRTs"). This bimodal increase in response speed is known as the Intersensory Facilitatory Effect ("IFE"; Colonius & Arndt, 2001). The IFE is primarily explored using single, high contrast targets. When complex background fields have been used in IFE experiments, the decreased signal/noise ratio has been found to reduce, but not eliminate, the facilitatory effect. For instance, when participants have been asked to orient to audio-visual targets, increased noise ratios in the audio stream have been shown to decrease accuracy and increase reaction times (Corneil, Wanrooij, Munoz, & Opstal, 2002).

Saccades to auditory targets are typically driven by activation of neurons in the deep layers of the superior colliculus (Jay & Sparks, 1987; Meredith & Stein, 1986). These neurons are arranged topographically with relation to the "motor error" between the current and desired position of the eye (Yao & Peck, 1997). Since visual stimuli are ordinarily encoded retinotopically, but auditory stimuli are typically encoded relative to the head, this implies that these signals are transformed upstream of the superior colliculus (Jay & Sparks, 1987).

Although the superior colliculus is proximally the source of saccadic control, the ultimate sources of these instructions lay upstream. Moreover, the brain regions responsible for these instructions depend on whether the saccades are reflexive or volitional (Arnott & Alain, 2011). When saccades to auditory targets are reflexive, it is thought that they are ultimately generated by a part of the posterior parietal cortex referred to as the Parietal Eye Field (PEF). Volitional reflexes, on the other hand, seem to originate from the dorsolateral prefrontal cortex and a region of the superior frontal sulcus known as the Frontal Eye Field (FEF).

All of these studies of audio-driven saccades rely on natural forms of auditory source localisation. In humans, sounds are localised using a variety of cues (see Grothe et al., 2010). Horizontal (azimuth) localisation is afforded by Interaural Time Difference ("ITD"; sound reaches the near ear first) and Interaural Level Difference ("ILD"; sound is louder in the near ear) cues. Vertical (elevation) localisation is possible due to the manner in which different

frequencies are reflected by the outer ear (known as the Head Related Transfer Function or "HRTF").

In sensory substitution, spatial position tends not to utilise these localisation cues. Whilst it is possible to simulate the main cues of auditory position using headphones, the closest approximations used by extant sensory substitution devices are poor facsimiles of organic cues. The spatial scheme of Vox (the SSD used in the present study) uses a variety of spatial cues which differ markedly in the extent to which they resemble their natural equivalents. Horizontal position is conveyed using time (as the sound sweeps from left to right) and stereo panning. Stereo panning may be thought of as an exaggerated version of an interaural level difference, but the left –right sweep is based in a spatiotemporal metaphor (Santiago, Lupáñez, Pérez, & Funes, 2007) that is oriented in the direction of the written form of a subjects native language (Tversky, Kugelmass, & Winter, 1991). Vertical position is conveyed by pitch. Although the HRTF depends on the relationship between frequencies and vertical position, this is not comparable or compatible with the representation of vertical position employed by Vox.

We are hence presented with a few testable questions for study. First, do more natural cues afford superior localisation? Since Vox utilises more natural cues for the horizontal axis we would expect localisation to be more accurate in this dimension. Secondly, can recently learnt, arbitrary cross-modal mappings drive saccadic eye movements? That saccades can be manipulated is well established – visual cues can result in congruency effects (Kuhn & Benson, 2007), for example, and auditory stimuli can drive saccades (Frens & Opstal, 1995). If naive participants begin to make saccades based on the artificial cues provided by Vox within a single experimental session, this would suggest a greater degree of plasticity than has previously been demonstrated. Thirdly, can a partial visual scene aid the comprehension of a more complete scene experienced via a sensory substitution device?

With regards to the first of these questions, it is possible that the stereo panning used by Vox is already similar enough to the natural ILD cue to automatically trigger saccades. Even if this is true however, a user of Vox would still have to re-scale their mapping such that the exaggerated level difference corresponds to the width of the encoded image. This manner of adjustment is conceptually similar to the perceptual adaptation first demonstrated by the inverted optics of George Stratton, but expanded upon by many since. Stratton demonstrated that the brain is capable of adjusting to an artificially inverted image falling upon the retina (Stratton, 1896). More recent work has failed to match the degree of adaption reported by

Stratton, but has nevertheless replicated the principle (Richter et al., 2002) and shown that the effects acquired by training in one task generalise to other tasks (Alexander, Flodin, & Marigold, 2011; Morton & Bastian, 2004). If auditory adaptation were to occur in users of an SSD employing panning to map horizontal position, it is conceivable that they would gain an advantage specific to horizontal localisation.

To address these questions the present study involves tracking the eye movements of participants as they attempt to use Vox to determine the location of a target.

4.3 Methods

4.3.1 Participants

Forty-eight students (27 female, aged between 18 and 37) were recruited from the University of Sussex and were awarded course credits for their participation. Ethical approval was granted by the Life Sciences & Psychology Cluster-based Research Ethics Committee at the University of Sussex. All participants reported normal hearing and normal (or corrected to normal) vision. No participants reported any prior exposure to any form of SSD.

4.3.2 Materials

The visual materials consist of scenes with three levels of degradation. The auditory materials consist of "soundscapes" generated from the visual materials but with an additional element (a target) added that is heard but not seen.

4.3.2.1 Visual scenes

Six images were selected as exemplars of ecologically valid scenes. Two were indoor scenes, two were natural outdoor scenes and two were urban outdoor scenes. In each of these three categories, one of the pair was visually dense (many objects) and the other visually sparse (few objects). Each image was cropped and scaled so that they were 760 x 760 pixels in size. The images were also converted to greyscale by desaturation.

These 6 images were then manipulated in two ways to generate a final set of 18 scenes. One manipulation consisted of pixellating the images so as to reduce them to 20 x 20 grids (in which each square in the grid was 38 pixels in width and height). The second derivation involved replacing the image with a solid block of grey with lightness equal to the average lightness of the original image. This was achieved with the help of the "Sample Average Colour" plugin for the Gimp image editor. The purpose of these derivations was to provide

different levels of background information. An example of an original and manipulated image is shown in Figure 20.

Figure 20: From left to right: an original (dense) scene, the Gaussian pixellated derivation, the average lightness derivation.

4.3.2.2 Auditory targets

Into these scenes, we inserted one of 4 high-contrast geometric targets. These were initially designed to contain features that sound distinct when sonified using the Vox sensory substitution device. Examples of such features include areas of solid colour and highly contrasting stripes. A small preliminary study (n = 9 additional participants) was then used to whittle down 7 targets to the 4 that were most easily detected. These 4 targets were then inserted into each of the four quadrants of the scenes, randomly jittered such that they occupy one of four sub-quadrants. By combining the visual images and applying 4 targets in 4 quadrants, a total of 288 unique images were sonified.

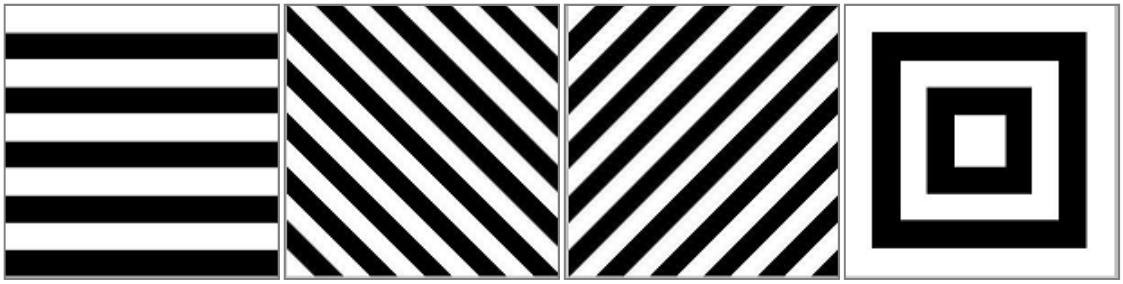


Figure 21: The four targets used in the final study. A) Horizontal lines B) Downward diagonal lines C) Upward diagonal lines D) Concentric boxes

4.3.2.3 Sonification

We used a sensory substitution device called Vox to sonify the composite images. Vox uses frequencies between 750Hz and 2,500Hz and has a resolution of 80 (horizontal in image, time-points in the soundscape) by 80 (vertical in the image, frequencies in the soundscape). For a more detailed description of the sonification process employed by Vox, see our recent paper (Wright & Ward, 2013).

4.3.3 Procedure

We used an Eyelink II head-mounted eye-tracking system to record eye-movements. Fixations were recorded using the pupil of the dominant eye at a rate of 500Hz. Participants were randomly assigned to one of three conditions. For two groups ("directed eye movements" and "free eye movements") the presentation of soundscapes was accompanied by the visual presentation of the scene, but not the target. In the "directed eye movements" condition (n = 12) participants were instructed to use their eyes to scan the scene and to look at the area in the scene corresponding to the location of the target in the composite soundscape. In the "free eye movements" condition (n = 12) participants were instructed to ignore the eye-tracker. Except when comparing the effect of the instructions, these groups are considered together as "Audio + Visual" (n = 24).

The third group ("audio only"; n = 24) were not presented with any visual stimuli during the decision phase. Instead of viewing the background scene, participants were presented with the outline of a square equal in size to the scenes shown to groups A and B and to the composite image shown to all participants in the feedback phase. Participants in group C were told to try picture the soundscape inside the square box and given the same eye-movement instructions as in the "directed eye movements" condition. A between-subjects design was used in order to prevent any training effects being carried over from one condition to another.

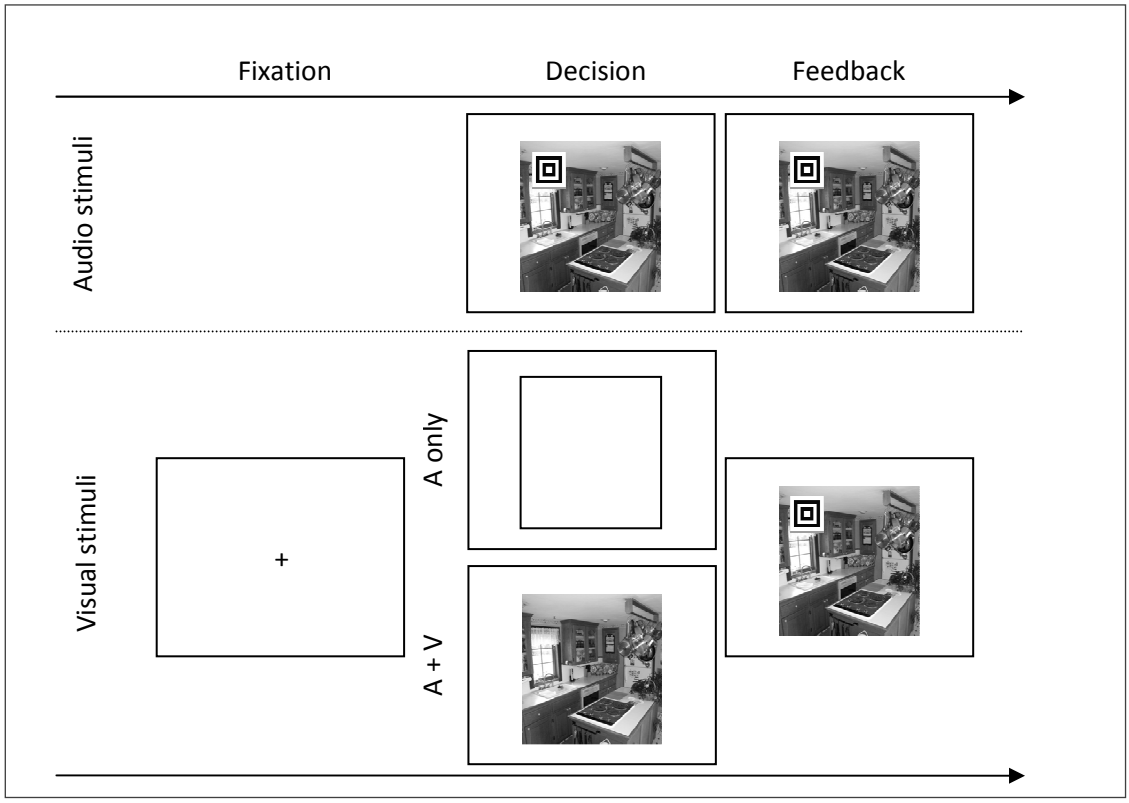


Figure 22: Trial structure showing visual and auditory stimuli for both the audio + visual and audio only conditions

Before starting the experiment participants were given an overview of the Vox image-to-sound conversion process. This overview explained the way in which Vox scans over the image as well as the relationships between vertical position and pitch and between lightness and loudness. They were then instructed that they would hear a soundscape generated by a composite image. Participants were asked to indicate which quadrant of the image contained the target using the central four buttons of a Cedrus RB834 response pad. Participants were not asked to respond to the identity of the target in any way – the variation in targets was included solely for the purpose of reducing declarative learning (i.e. remembering what tareget A in the top corner of background 1 sounds like) in favour or procedural learning (i.e. learning how to interpret soundscapes).

Each trial participants made a decision based on one soundscape, which looped up to 6 times. Participants could respond any time after the first full loop of the soundscape, but were encouraged to do so within the subsequent five loops. This number of repetitions was selected during the development of the paradigm as a sensible balance between making the task realistically achievable and completing the requisite number of trials within a reasonable period.

Button presses interrupted the looping of the soundscape and advanced the participant to a feedback screen. This feedback screen showed the composite image and played the soundscape once more. Each participant responded to each of the 288 soundscapes, presented in a randomised order. To measure improvement over time, these 288 trials were analysed as three blocks of 96 trials each.

Visual stimuli were presented on a Sony Trinitron Multiscan E530 monitor with a diagonal distance of 21 inches and a resolution of 1280 x 1024 pixels. Auditory stimuli were played through an ASIO EMU 0202 audio interface. In-ear monitors (IEMs) were used to avoid obstructing the headset of the eye-tracker.

4.4 Results

4.4.1 Performance data

In determining performance for this task, the dependent variables are the proportion of targets localised to the correct quadrant (where chance would result in 0.25) and the time taken to make a response. The results from the two audio-visual conditions are collapsed because overall performance between they did not differ in terms of the proportion of correct responses (mean = .37; $t(22) = 0.206$, $p = .839$) or response time (mean = 3607ms; $t(22) = 0.857$, $p = .401$). Moreover, the different eye-movement instructions are primarily of relevance to the eye-movement data.

Similarly, the targets were chosen from pilot studies to be maximally identifiable, and a repeated measure ANOVA was performed that confirmed that the target used had no significant effect on the proportion of correct responses ($F(3,138) = 2.09$, $p = .105$; see also Figure 23). As such, remaining analyses collapse across target type.

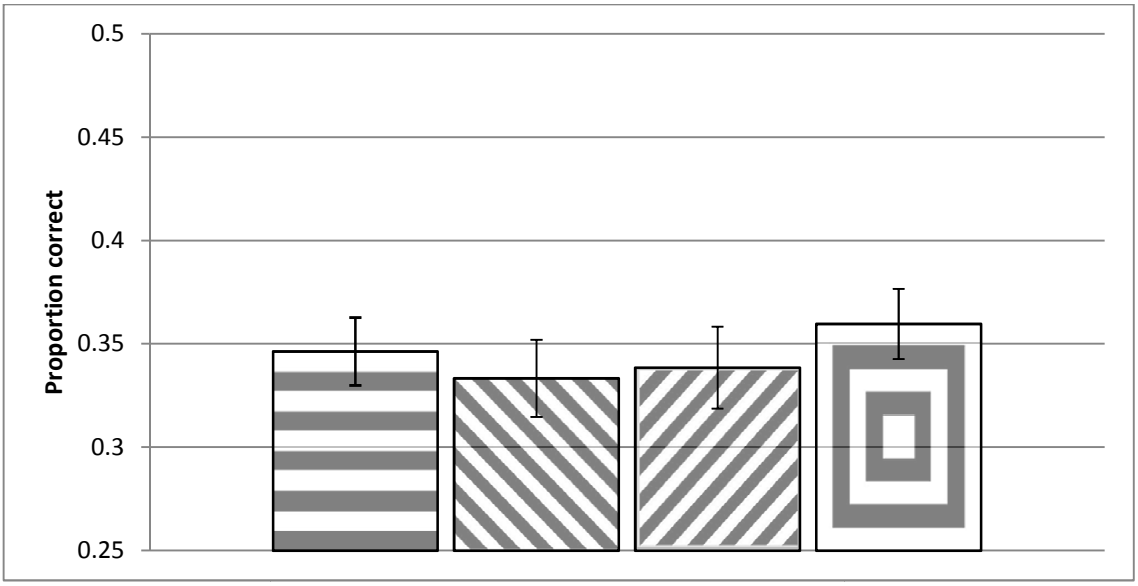


Figure 23: Proportion correct by target type. Bars show the standard error.

Figure 24 shows the proportion correct, across time, by condition and background type. A repeated measures ANOVA was conducted which compared the proportion of correct responses by modality (audio + visual, audio only), time (block 1, block 2, block 3) and by background type (original, pixellated, average lightness). Mauchly's test indicated that the assumption of sphericity had been violated for the background type variable ($\chi^2(2) = 12.35$, $p = .002$), so the degrees of freedom were corrected using Huynh-Feldt estimates ($\varepsilon = .85$). Significant main effects of time ($F(2,92) = 18.22$, $p < .001$), background type ($F(1.70,78.16) = 24.13$, $p < .001$) and modality ($F(1,46) = 4.69$, $p = .036$) were found. No interactions were found to be significant.

Interestingly, the audio-visual condition was superior to the audio only condition even though the target stimulus was only presented in the audio channel. As expected, the complexity of the sonified background affected performance such that the simplest background was associated with the best performance. The significant effect of time demonstrates that participants improved but, the absence of interactions with time, suggests that the improvements were general across all types stimuli types rather than specific to certain stimuli.

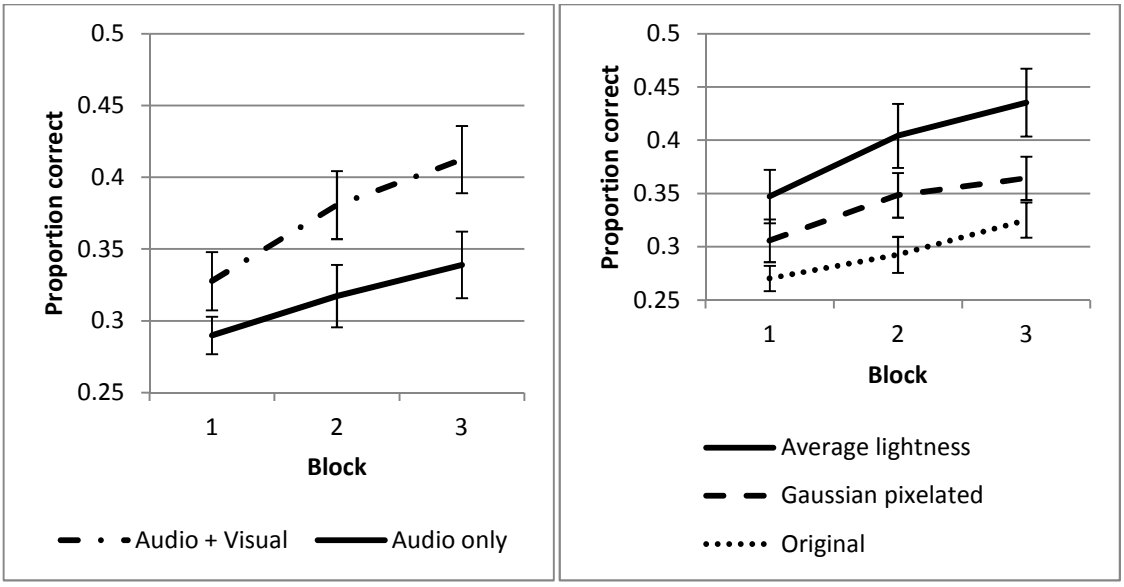


Figure 24: Effect of condition (left) and background type (right) on the proportion of correct responses across blocks. Bars show the standard error.

Performing the same ANOVA on reaction times we find significant main effects of time ($F(1.31,60.42) = 37.72$, $p < .001$; corrected using Greenhouse-Geisser estimates, see below) and background ($F(2,92) = 11.98$, $p < .001$) type, but not modality ($F(1,46) = 0.10$, $p = .754$). Once again, there were no significant interactions. In this instance, Mauchly's test indicated that the assumption of sphericity had been violated in the case of the time variable ($\chi^2(2) = 33.27$, $p < .001$), so the degrees of freedom were corrected using Greenhouse-Geisser estimates ($\varepsilon = .66$). Figure 25 shows the effect of time and background type on the average reaction time. Interestingly, participants became quicker to respond as well as becoming more accurate (in other words, there was no trade off between speed and accuracy). A post-hoc pair-wise comparison (with Bonferroni correction) reveals that both the original ($p < .001$) and Gaussian pixellated ($p = .001$) backgrounds lead to longer reaction times compared to the average lightness backgrounds, but are not significantly different to each other ($p = .429$).

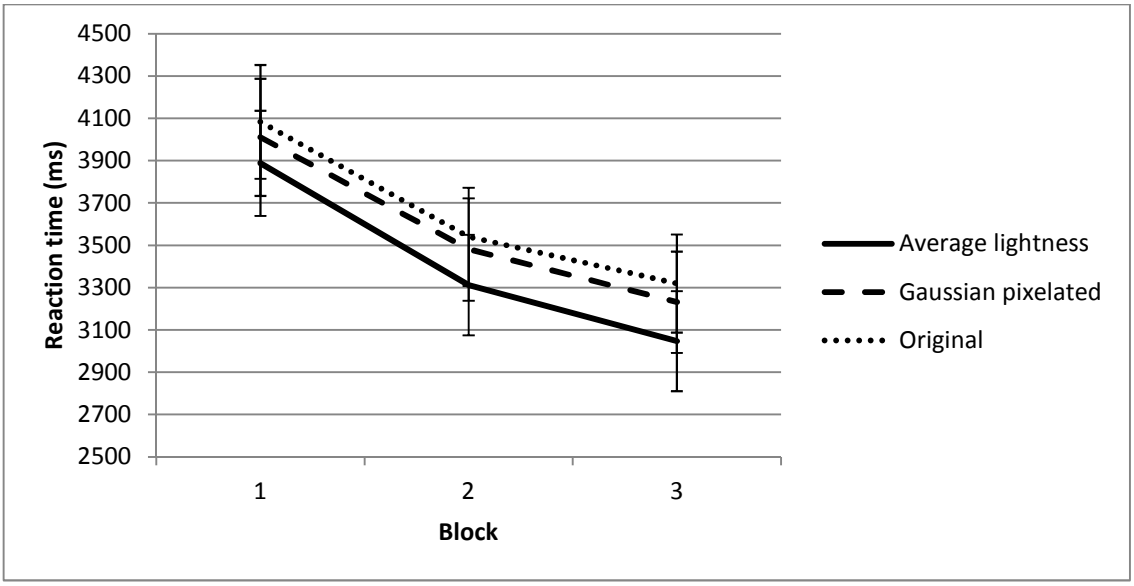


Figure 25: Effect of background type on the average reaction time across blocks. Bars show the standard error.

4.4.2 Eye-tracking data

The Eyelink system records the location of the participant's gaze relative to the visual stimuli. By defining an "interest area", it is possible to quantify the proportion of time spent dwelling in each of the quadrants ("dwell time"). The main quadrant of interest is the quadrant which contained the target. On trials where the participant responded incorrectly, the quadrant corresponding to their response are also of interest. Accordingly, we can re-label these quadrants as the "target quadrant" and the "response quadrant".

The data for the dwell times in the target quadrant across all trials broadly reflect the behavioural data. As before, there is no significant difference between the two audio + visual conditions (i.e. whether participants were instructed to fixate the target or not; mean = .344; $t(22) = 1.73$ $p = .098$), so these are again considered together. Repeating the earlier 3x3x2 repeated measures ANOVA on the eye-tracking data shows that dwell time in the target quadrant across all trials ("target dwell time") increases with time ($F(2,92) = 9.27$, $p < .001$), is greater in trials featuring simpler backgrounds ($F(1.879,86.43) = 19.39$, $p < .001$) and greater when participants are presented with visual as well as auditory stimuli ($F(1,46) = 4.17$, $p = .047$). Interestingly, there is a significant three-way interaction: in the audio + visual condition (but not in the audio only condition) the "Average lightness" and "Gaussian pixellated" background types afford a greater improvement over time than do the Originals ($F(1,46) = 3.46$, $p = .009$). Additionally, the two-way interaction between background type and time approached significance ($F(1,46) = 2.35$, $p = .065$), but this is likely to be driven by the three-way interaction described above. On this occasion, Mauchly's test indicated that the

assumption of sphericity had been violated for the background type variable ($\chi^2(2) = 6.17$, $p = .046$), so the degrees of freedom were corrected using Huynh-Feldt estimates ($\varepsilon = .94$).

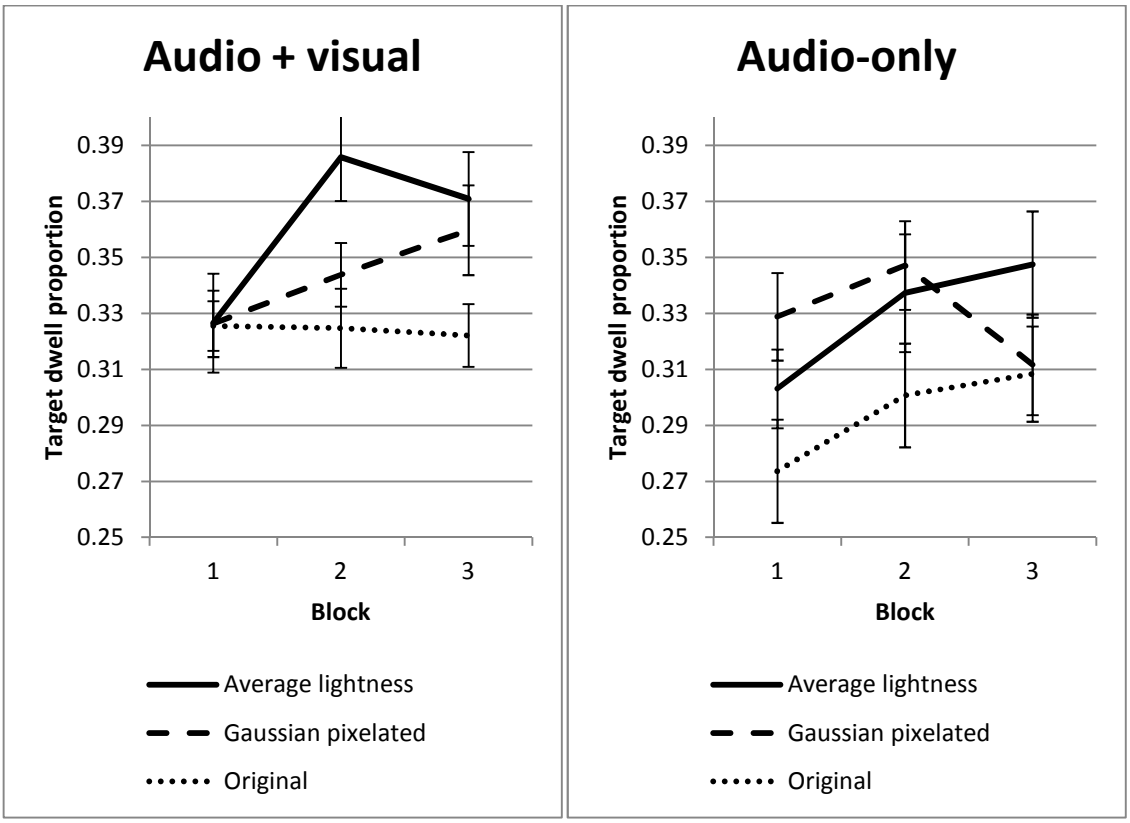


Figure 26: Target dwell proportion by block, background type (lines) and audio-visual condition (top and bottom graphs). Bars show the standard error.

Although the target dwell time on incorrect trials (mean = .271) is significantly lower than the average for all trials (mean = .331), a one-sample T-test shows that it remains significantly above chance ($t(47) = 4.76$, $p < .001$). In other words, the eyes still spend longer than expected dwelling on the quadrant containing the target even if the participant consequently responds incorrectly.

Like the target dwell time, the "response dwell time" is dwell time in the response quadrant. Unlike previous analyses though, this metric reveals a significant difference between the "free eye movements" (mean = .434) and "directed eye movements" (mean = .340) conditions ($t(22) = 3.66$, $p = .001$). Since participants in the directed eye movements condition were instructed to look towards where they believed the target to be located, it is not entirely surprising that their eye movements matched their responses more closely than those of participants who received no such instructions.

Accordingly, a repeated measures ANOVA was conducted to compare the response dwell time by block (block 1, block 2, block 3) and by condition ("directed eye movements", "free eye movements", "audio only"). The response dwell time was found to be significantly different between groups ($F(2,45) = 7.77$, $p = .001$) and to increase with time ($F(2,90) = 15.38$, $p < .001$). There were no significant interactions. Applying a post-hoc pair-wise comparison of the conditions (with Bonferroni correction) reveals significant differences between "directed eye movements" and both the "free eye movements" ($p = .005$) and "audio only" ($p = .002$) groups. In other words, participants who were instructed to look at the area that they believe contains the target spend a greater proportion of the trial time dwelling in the quadrant corresponding to their subsequent response.

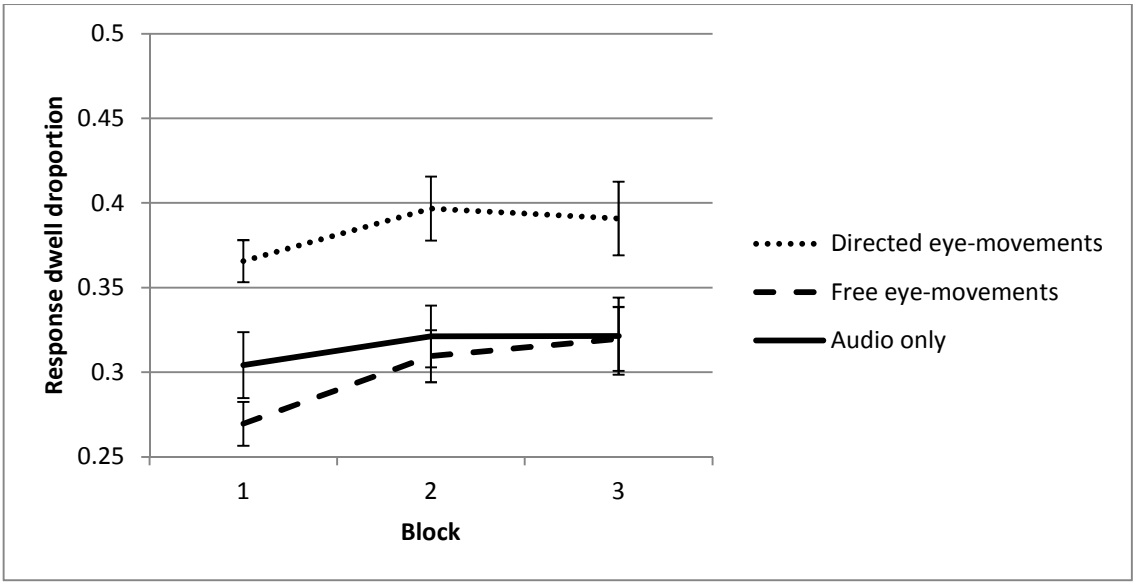


Figure 27: Effect of participant group on the proportion of time spent dwelling in the response quadrant by block. Bars show the standard error.

4.4.3 Horizontal and vertical discrimination

As explained in the introduction, the cues Vox uses to encode horizontal position (stereo panning; time) are more naturalistic than the cue used to encode vertical position (pitch). Therefore, we undertook to additionally analyse the horizontal and vertical components of accuracy individually. A response of "top right" to a target in the top left quadrant would, for instance, be vertically correct, but horizontally incorrect. (Note that this causes the proportion of correct responses expected by chance alone from 25% to 50%.) Initially, a paired-sample T-test shows that accuracy in horizontal discrimination (mean = .60) is significantly greater than in vertical discrimination (mean = .56; $t(47) = 4.19$, $p < .001$). Furthermore, this difference is also significant across all blocks (block 1: $t(47) = 2.49$, $p = .016$; block 2: $t(47) = 4.01$, $p < .001$; block 3: $t(47) = 3.35$, $p = .002$).

Having conducted a repeated measures ANOVA on the vertical accuracy data, comparing time (block 1, block 2, block 3) and condition (audio + visual, audio only), significant main effects of time ($F(2,92) = 8.06$, $p = .001$) and condition ($F(1,46) = 9.59$, $p = .003$) were found. No other effects or interactions were found to be significant. Conducting the same ANOVA on the horizontal accuracy data also reveals main effects of time ($F(2,92) = 15.52$, $p < .001$) and condition ($F(1,46) = 9.59$, $p = .003$). For this horizontal data a significant contrast was also found for the interaction between time and condition ($F(1,46) = 4.25$, $p = .045$). In both cases, participants improve over time and did better in the audio + visual condition. The interaction found in the horizontal data indicates that participants improved more rapidly in the audio + visual condition. Plots of these data can be found in Figure 28.

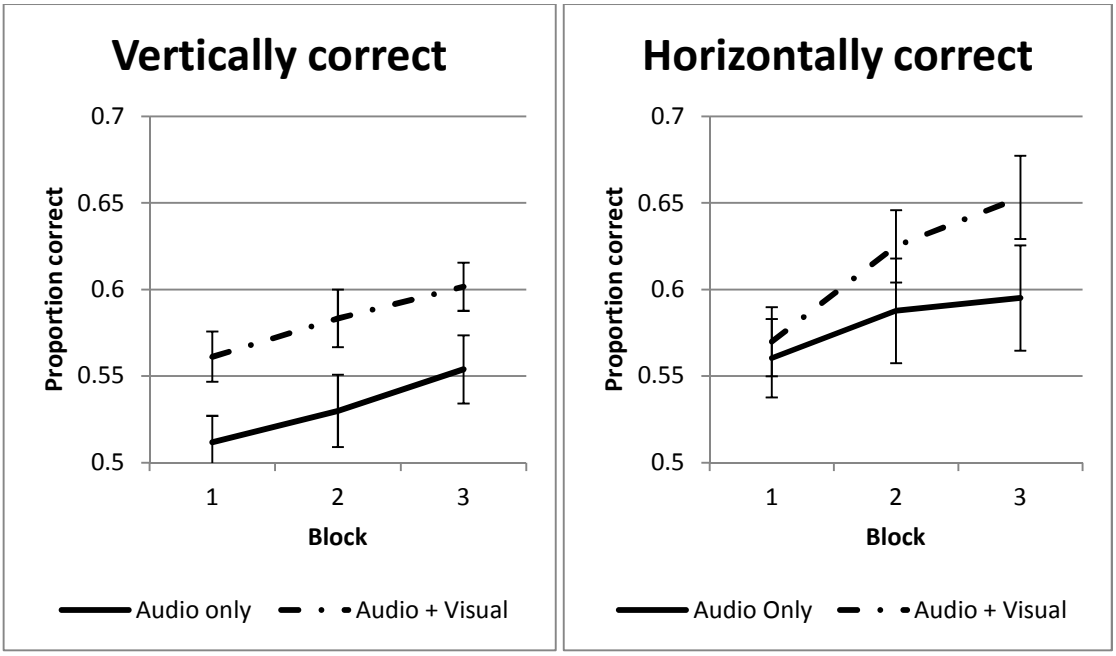


Figure 28: Comparison of proportion of trials where responses were horizontally or vertically correct by block and condition. Bars show the standard error.

4.5 Discussion

Overall, these results show that natural vision and substituted vision are capable of interacting and will do so readily. Indeed, the data presented above strongly suggests that the interaction is bidirectional: On the one hand, participants in the audio + visual condition performed better than those in the audio only condition. On the other, the eye-tracking data shows that participants made saccades to the region of the image corresponding to the location of the target in the substituted stream (even when they subsequently made an incorrect judgement).

This contrasts to previous research in which an audiovisual task – as opposed to a substituted vision task – was found to be hindered by the extra noise of the background (Corneil et al.,

2002). Like the present study, this tracked saccades towards auditory, visual and audiovisual targets with varying levels of "noise". The key difference is that in the present study this noise is informative, as it matches the presented visual stimuli.

It is important to remember however, that our data also shows that increasingly complex backgrounds have a cost to performance. The fact that the simplest backgrounds result in the fastest, most accurate responses may be due to participants being overwhelmed by the amount of information present in the sounds produced from more detailed backgrounds. Alternatively, it may be the result of a greater degree of "complexity contrast": the complex sound of a target more noticeably punctuates the constant sound of a grey background than the complex sound of a detailed background. It would be very interesting to see if this effect persists in more experienced users of sensory substitution devices.

The three-way interaction found for target dwell time suggest that in the audio + visual conditions the average luminance backgrounds allow a greater degree of target localisation by the second block. The pixellated backgrounds only reach an equivalent level of target dwell in the third block. This would seem to support an explanation based in the overall level of information, rather than merely the contrast between the backgrounds and the targets.

This positive effect of visual information is likely to be of key importance in the development of SSDs for visually impaired people who have some residual vision, as well as the development of sensory augmentation devices for people with normal vision. The effect is reminiscent of the Intersensory Facilitation Effect (Colonius & Arndt, 2001). It is likely that a person with low levels of vision would experience a facilitatory effect to their combined visual perception when using a sensory substitution device. For example a person with scotomas may find it useful for a sensory substitution device to "fill in" the missing parts of their visual field. Simultaneously, they may find it easier to learn to use the sensory substitution device than a person with no vision at all.

It would be very interesting to investigate whether facilitation in the reverse direction is possible – can substituted vision aid natural vision? This could be determined using low-contrast visual stimuli and a high-contrast signal from a sensory substitution device. The results of such an experiment would be of direct relevance to designers of sensory augmentation systems. It is easy to imagine, for instance, a thermal imaging sensory augmentation device facilitating the detection and identification of animals in low light situations.

It would also be interesting to investigate whether facilitation could be enhanced by closer integration with natural vision. If the substituted vision behaves as a sort of perceptual overlay, would moving the sensor in synchrony with the eyes lead to more rapid adaptation? Then there is the fact that the sensors used in most sensory substitution devices are webcams with narrow fields of view. What effect would a panoramic sensory substitution device have on the level of integration?

The eye-tracking data shows that visual targets present only in the substituted signal are capable of driving saccades. This provides evidence for a substituted vision to natural vision interaction. This shows that spatial cues from substituted vision are not just readily identified by novice users of sensory substitution devices, but that these cues are at treated as (at least somewhat) equivalent to those present in natural vision. Moreover, this suggests that the integration of these cues occurs at a level low enough to tap into the superior colliculus (Jay & Sparks, 1987). That these saccades take place regardless of whether or not the participants were instructed to move their eyes suggests that some proportion are volitional, whereas some are reflexive.

It would be highly interesting to conduct an EEG/LORETA (Pascual-Marqui, Michel, & Lehmann, 1994) study whilst participants were performing the task used in this study: the high temporal resolution (and adequate spatial resolution) should allow for activity in the PEF and FEF to inform whether each individual saccade is reflexive or volitional (Arnott & Alain, 2011).

It is possible that this question could also be addressed by further analysis using eye-tracking data. Specifically, saccade latency has been used in other work to distinguish between volitional and reflexive movement (Anderson & Carpenter, 2010; R. Walker, Walker, Husain, & Kennard, 2000). Given that Vox uses time to represent horizontal position however, calculating latency for saccades made in response to Vox-generated soundscapes may be challenging.

The interaction between condition and time for horizontal localisation shows that having access to visual information increases the rate at which this aspect of performance improves. This suggests that the natural visual signal facilitates the re-mapping of the cues in the substituted signal. In this way natural vision may be acting like prism glasses for the auditory localisation cues present in the substituted signal, allowing perceptual adaptation to take place, mapping full stereo panning to an Interaural Level Difference.

The generally higher level of accuracy in localisation targets horizontally suggests that the more naturalistic localisation cue (panning) used by Vox to encode horizontal position is more

readily understood than the vertical localisation cue (pitch). Unfortunately, the fact that Vox also uses a time-based cue to encode horizontal location makes this conclusion far from clear cut – it is possible that the enhanced performance in this domain is due entirely to this double representation.

A possible explanation for this effect might lie in the Bayesian framework, which suggests that multisensory integration is the result of statistical comparisons between the individual input modalities (Deneve & Pouget, 2004). By this approach, the task featured in the present study might be understood in terms of subtracting the visual signal from the auditory signal rather than solely determining the location of the target by its own auditory representation. This explanation partially undermines claims that participants are performing true sensory substitution, but since participants still perform above chance in the audio-only condition, this cannot be the full explanation. Moreover, the significance of Bayesian inferences is moot when considering the potential utility of sensory augmentation devices.

Further research will be needed however, to resolve the extent to which naturalistic cues scaffold the early stages of sensory substitution. Since other sensory substitution devices use different cues, they offer a means with which to undertake this further research. The Vibe relies on fully naturalistic cues to represent horizontal location, for instance, and could consequently be used as a point of comparison. Alternatively, the Polyglot framework (see chapter 3) could be used to create pairs of sensory substitution devices differing only in the degree to which they rely on naturalistic cues.

In summary, we have shown that substituted vision and natural vision are capable of bidirectional interaction and that this has measurable effects on the ability of SSD users to perform visual tasks. We suggest that it may be of great utility to the users and designers of sensory substitution/augmentation devices. Further work is required to explore the benefits conveyed by naturalistic localisation cues.

Chapter 5: Cross-modal correspondences between hue and pitch are mediated by geometry of visible colour space

5.1 Abstract

The cross-modal correspondences between visual luminance and aural pitch are well established. Associations between visual hue and pitch have also been reported but it is unclear whether the effects are truly attributable to hue (as opposed to luminance, saturation). These correspondences could be incorporated in sensory substitution technology. Participants performed a speeded 2-alternative forced choice task in which they listened to tones (differing in pitch) and selected the best visual match from two luminance-matched colours. Experiment 1 shows that increasing pitch is linked to an increasing preponderance of yellowy responses and fewer purplish responses. Experiment 2 shows that the mapping is relative rather than absolute (e.g. a tone of 300Hz can be mapped as either yellowy or purple depending on whether the tone is the highest or lowest pitch in the set presented). This cannot be attributed to a direct effect of luminance but we show that the effect can be explained by typical hue-luminance associations, which are driven by the irregularity of the geometry of perceptual colour space (e.g. yellow hues tend to be higher in luminance).

5.2 Introduction

Crossmodal correspondences are links between stimuli of different modalities. Like synaesthesia, these links often seem to be arbitrary. Unlike synaesthesia however, crossmodal correspondences appear to be a largely universal phenomenon, are typically bidirectional, and do not result in the *perception* of a concurrent (Spence, 2011). Crossmodal correspondences are known to exist between most pairings of the senses. Moreover, crossmodal correspondences exist between low-level sensory features (e.g. brightness or loudness) and more abstract features, such as size or shape (e.g. Spence & Gallace, 2011).

Early demonstrations of crossmodal correspondences include the now famous “mil” / “mal” experiment: normal (non-clinical, non-synaesthetic) participants preferentially paired the label “mal” with a larger object and “mil” with a smaller object (Sapir, 1929). Indeed, that sounds may carry inherent meaning may be observed in natural languages (Nuckolls, 1999). Collectively these phenomena are examples of “phonetic symbolism”.

Due to the nature of crossmodal correspondences, many examples will be very familiar to the reader. Consider colour and temperature: most people, when asked, would associate “red” with “hot” and “blue” with “cold” (Morgan, Goodson, & Jones, 1975). Sound and size also exhibit a familiar example: most people associate low pitches with large objects and vice versa (P. Walker & Smith, 1985). Similarly, pitch is associated with vertical position (Martino & Marks, 1999). Perhaps less familiarly, correspondences have also been shown to exist between odours and abstract visual forms (Seo et al., 2010).

Crossmodal correspondences have two very interesting properties. The first is a tendency towards polar dimensions, such as auditory loudness, visual brightness or gustatory sweetness (Smith & Sera, 1992). The second is a propensity to be largely relative, meaning that the absolute loudness of a stimuli (for instance) matters less than how it compares to contemporaneous stimuli (Marks, Szczesiul, & Ohlott, 1986). Both these features contrast sharply with synaesthesia, which tends to exhibit associations which are arbitrary and absolute: for a synaesthete middle C may always be a particular shade of blue (for instance), whereas the neighbouring D may be an equally specific shade of red (Grossenbacher & Lovelace, 2001).

5.2.1 Crossmodal correspondences in sensory substitution

Sensory substitution devices are systems which convey information belonging to one sensory modality via a different sensory modality. The typical application is to restore the functionality

of some lost sense; often vision. The first sensory substitution device – the TVSS – was a visual-tactile device that used a TV camera and 400 solenoids to allow users to experience an image on the skin of their back (Bach-y-Rita et al., 1969). More recent devices have demonstrated the feasibility of visual-audio ("the vOICe"; Meijer, 1992) and vestibular-tactile ("the TDU"; Danilov et al., 2006) sensory substitution.

In our work with sensory substitution devices, we have seen how integral crossmodal correspondences can be. We have shown that sensory substitution devices that obey the rules of crossmodal correspondences are more optimal than those that do not and have suggested that this is because it allows the user to "bootstrap" their acquisition of the sensory substitution skill (Wright & Ward, 2013).

The visual sensory substitution devices mentioned above both convey information about lightness, but, like most visual sensory substitution devices, do not convey any information about colour. Colour is an important aspect of the experience of vision, and helps in object identification and scene segmentation. Since colour is so intrinsic to normal vision, it is not surprising that there have been several attempts to produce a colour sensory substitution device.

The first such attempt took a single point of colour and mapped it to a colour name, a recording of which was then played to the user (McMorrow et al., 1997). This colour-to-speech device may be of practical use to a visually impaired person and is certainly an example of an assistive technology, but isn't really what most people would call a sensory substitution device.

More recently there have been a couple of more interesting attempts. Artist Neil Harbisson is sighted, but has achromatopsia – total colour blindness. Adam Montandon built him a device that samples a point of colour and maps the hue to an audible tone (Hauskeller, 2012; Wade, 2005; Montandon, n.d.). Similarly, the Kromophone device maps a point of colour to a tone, but also uses stereo panning to further aid discrimination (Capalbo & Glenney, 2009).

Unlike the devices mentioned thus far, the See ColOR device sonifies a central row of 25 pixels. Like the colour-to-speech device, the See ColOR maps hue to a colour identity, but then proceeds to assign these identities to musical instruments. Shades of red, for instance, are represented by an oboe. Saturation is likewise reduced to one of four discrete levels and then mapped to a musical note and stereo panning is used to distinguish between the 25 lateral pixels (Gomez et al., 2010). The See ColOR has also been adapted to convey the depth at each of these 25 pixels by adjusting the duration of each sound (Bologna et al., 2010). Two modes of

depth representation are offered by the See ColOR: the first relies purely on extending the duration of sounds representing more distant regions of a scene; the second employs this approach up to four meters, after which it begins to increase the volume of the sounds.

Not all colour sensory substitution devices use sound though. The Electro-Neural Vision System (ENVS) uses electro-tactile stimulation to encode depth and colour. Users wear a pair of gloves with an electrode on each finger corresponding to one of 10 horizontal regions of the image. Each electrode pulses with an intensity corresponding to the depth and a frequency corresponding to the colour (Meers & Ward, 2004). Again, the colour of each region is mapped to one of 8 discrete colours. Interestingly, the selection of these 8 colours and the 8 corresponding frequencies is left to the user.

Though each of these systems uses a different mapping, they are all united in their lack of empirical underpinning. That is not to say that they do not work – on the contrary, it seems they have each enjoyed a degree of success. Rather, the mapping of colour in each case is based entirely on an arbitrary decision by the designer. As mentioned above, we have previously found that optimal sensory substitution devices will draw on crossmodal correspondences for their mappings. The aim of the current study is therefore to explore whether there exists a crossmodal correspondence between colour and pitch.

5.2.2 Correspondences between colour and pitch

Many studies have previously examined possible correspondences between colour and sound. The best studied aspects of colour are lightness (or luminance) and saturation (or chroma). We know, for instance, that lightness corresponds with loudness and pitch. When asked to match metaphorical statements (e.g. “bright squeak”) with physical stimuli (such as tones and lights), both children and adults strongly favour this relationship (Marks et al., 1987). In other words, relatively lighter (e.g. pale grey compared to black) stimuli are associated with louder, higher pitched sounds. Further, ascending melodic intervals are associated with lighter visual stimuli and descending intervals with darker stimuli (Hubbard, 1996).

Research into sound-colour synaesthesia has exposed interesting patterns that appear to be shared by synaesthetes and normal controls. As well as confirming the link between pitch and lightness, relationships were also found between chroma and both pitch and timbre (Ward, Huckstep, & Tsakanikos, 2006). No significant link with hue was found.

Additionally, correspondences have been found between visual lightness and saturation and musical tempo and mode. When these are incongruent, it is the effect on tempo which

dominates. The team behind these findings assert that these links are mediated by emotion (Schloss, Lawler, & Palmer, 2008). Indeed, they support this assertion by demonstrating that the same set of colour-mood relationships correspond to facially expressed emotional states (Palmer, Langlois, Tsang, Schloss, & Levitin, 2011).

Colours	Mood	Music
Light & Saturated	Happy	Faster & Major
Dark & Desaturated	Sad	Slower & Minor
Saturated & Dark	Strong	Faster
Desaturated & Light	Weak	Slower

Table 2: Summary of relationships discovered by Schloss et al. (2008)

The aspect of colour whose relationship with sound is least well understood is hue. An early study seemed to indicate that higher pitches are matched by yellow hues, whereas lower pitches are matched with blue hues (Simpson, Quinn, & Ausubel, 1956). The prevailing view however, is that the failure of this study to control the perceived lightness or physical luminance of the coloured stimuli means that this result is likely to be affected by the known relationship between lightness and pitch. As Spence observes, yellow stimuli tend to be lighter than blue stimuli (Spence, 2011).

A subsequent study failed to find any effect between hue and pitch (Bernstein, Eason, & Schurman, 1971). This study suffers however, from a small number of participants (four, of whom two were experimenters) and only investigated the difference between red and blue stimuli. The results from this study cannot therefore be treated as conclusive.

In recent years, the relationship between hue and pitch has been neglected as an experimental topic. The very existence of such a relationship has not (to the present authors' satisfaction) been conclusively proved nor disproved. Consequently, sensory substitution research community have no choice but to make arbitrary decisions when attempting to represent colour with sound. The purpose of the present study is therefore to revisit this question using modern techniques to rigorously control variables such as lightness.

If a relationship between hue and pitch were found, what might this be like? As mentioned above, the strongest crossmodal correspondences appear to exist between dimensions with polar extremes, such as luminosity or loudness. As a dimension, hue does not fit this description: though it is often described with a scalar value, hue has the unusual property of circularity. It is therefore unclear what shape a hue-pitch relationship may take. Indeed, if a

relationship is found, its shape may have profound implications for theories of colour perception.

Pitch, on the other hand, does appear to be a polar dimension in that any given two pitches can be ordered linearly. That said, the phenomenon of musical harmony adds a level of intricacy to the dimension of pitch. In Western music, a note one octave above another (and consequently having double the frequency) shares a certain invariance that is perceived as a harmonic relationship. From this perspective, it is easy to see similarities with colour – despite having very different lightness, a pair of colours may share the same hue. The combination of the monotonic pitch trend with the cyclical musical pattern suggests that (perceptually at least) tones may exist in a spiral shaped dimension (Patterson, 1986), or a more complicated variant (Shepard, 1982).

Could a spiral shaped dimension provide the stepping stone between hue and pitch? The lack of a starting point in the hue dimension remains a problem, but it is possible to imagine hues aligning with notes as chroma and/or lightness increase along the spiral. Indeed, the basic premise of this model dates back to Isaac Newton, who mapped a colour wheel onto musical notes (as told in Greated, 2011).

This distinction brings to mind the “classes of continua” described by Steven (1957). Whilst raw pitch may be thought of a as a “class I” (or “prothetic”) continua, hue and musical note identity might more properly be considered to be “class II” (or “metathetic”) continua.

The possible shapes of a relationship between hue and pitch allow us to set out several competing hypotheses:

1. There is no relationship between hue and pitch
2. There is a relationship between hue and frequency
3. There is a relationship between hue and musical note

Finally, it is possible for these relationships to be either relative or absolute. That is, it could be the case that a particular shade of red always maps onto middle C, or it could be the case that reds are always higher pitched than blues. Given what we know about crossmodal correspondences, the latter seems much more likely.

5.3 Experiment 1

The purpose of this first experiment is to determine whether a relationship exists between pitch and hue and, in the event that such a relationship is found, to identify its properties.

5.3.1 Methods

5.3.1.1 Participants

Twenty students (18 female, aged between 18 and 24 years, mean age 19) were recruited from the University of Sussex and were awarded course credits for their participation. Ethical approval was granted by the Life Sciences & Psychology Cluster-based Research Ethics Committee at the University of Sussex. All participants reported normal hearing and normal (or corrected-to-normal) vision.

5.3.1.2 Materials

Participants were seated in front of a Dell D1626HT Trinitron monitor in a darkened room and wore a pair of Sennheiser HD497 headphones. The volume of the computer was adjusted to a comfortable level. The gamut of the monitor was recorded using a Cambridge Research ColorCAL (MK1) colorimeter. In subsequent plots, the gamut is illustrated by bounding the area within the maximal activation of the red (R_{max}), green (G_{max}) and blue (B_{max}) phosphors.

Five colours were selected. By iteratively selecting colours in the CIELUV colour-space, which was designed for perceptual uniformity (Hunt & Pointer, 2011, pp. 74–75; Tkalcic & Tasic, 2003), it was possible to match lightness and chroma (saturation) across these colours, whilst carefully controlling variations in hue. The hue angles of these colours were evenly spaced at every 72° ($\frac{2\pi}{5}$ or 1.27 radians). The first iteration was chosen so that the first colour had a hue angle of 0°. This set is plotted in Figure 29.

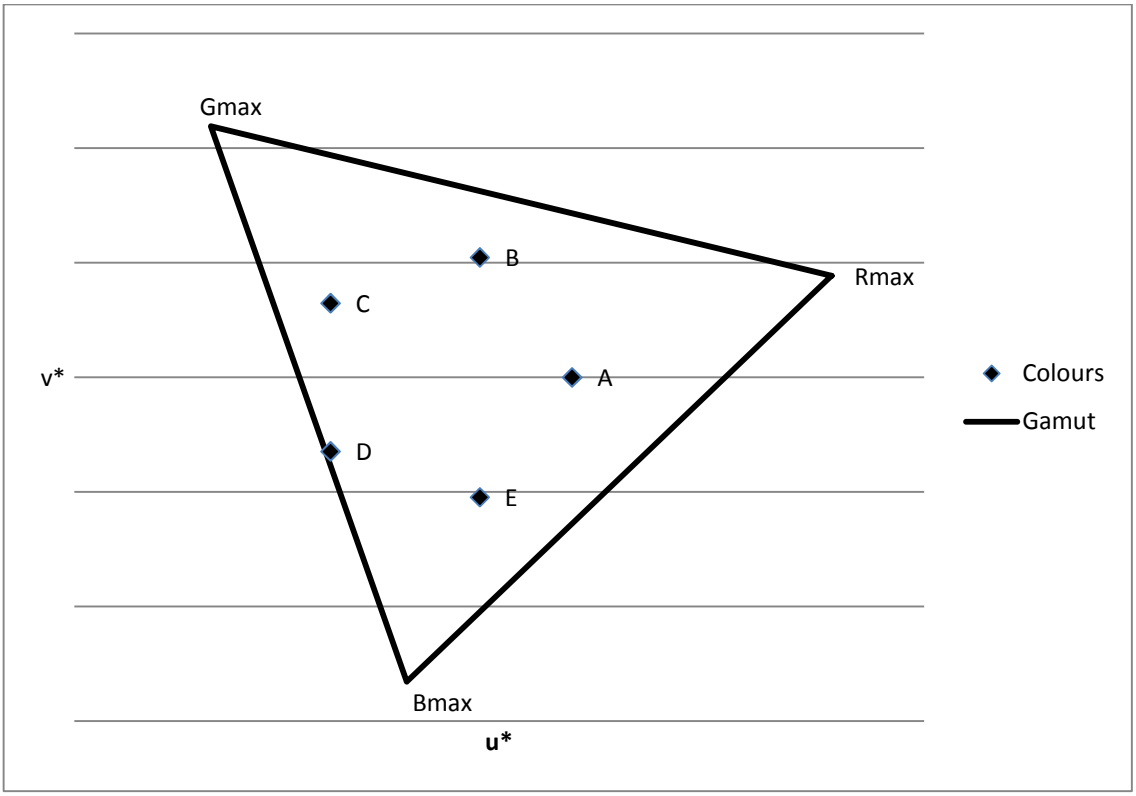


Figure 29: Monitor gamut and initial colours

This naive selection of hue angles restricts the maximum chroma of the five colours within the gamut of the monitor. Specifically, a chroma of 55 places point D very close to the green-blue edge of the gamut. To increase this maximum, these hues were adjusted such that the edge joining the bluest and the greenest colours (i.e. C & D) became parallel to the edge between the maximum possible blue and green values. This results in an offset of 18.39° (0.32 radians). This allowed the chroma to be increased to 60. These colours are mapped in Figure 30.

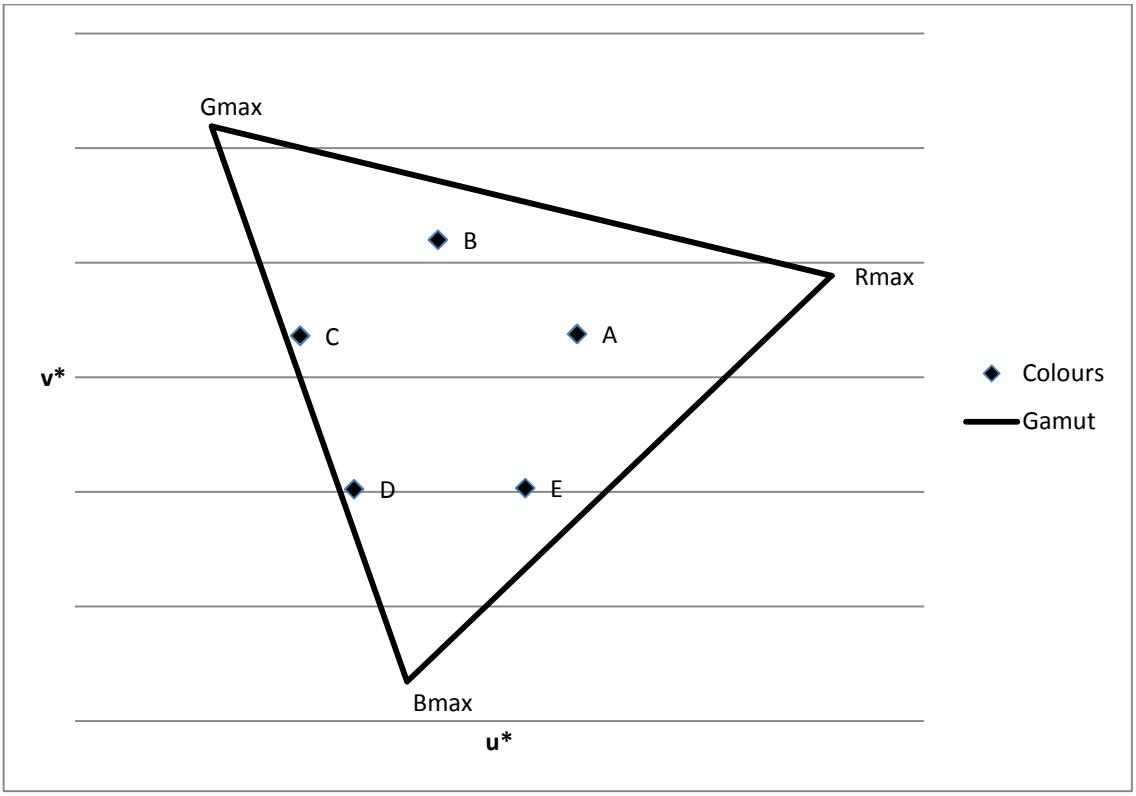


Figure 30: Adjusted colours within the monitor's gamut

Adjusted like this, the colour at D is very close to the "focal" shade of blue – representing that which is "most linguistically 'codable' and the most easily remembered" (Heider, 1972). This presented an opportunity to investigate whether proximity to a focal shade influences pitch preference. To fully exploit this opportunity, a further slight adjustment to the chosen hues was performed. As shown in Figure 31, by applying another offset of 6.61° (0.115 radians), it was possible to align the hue angle of point D with that of focal blue.

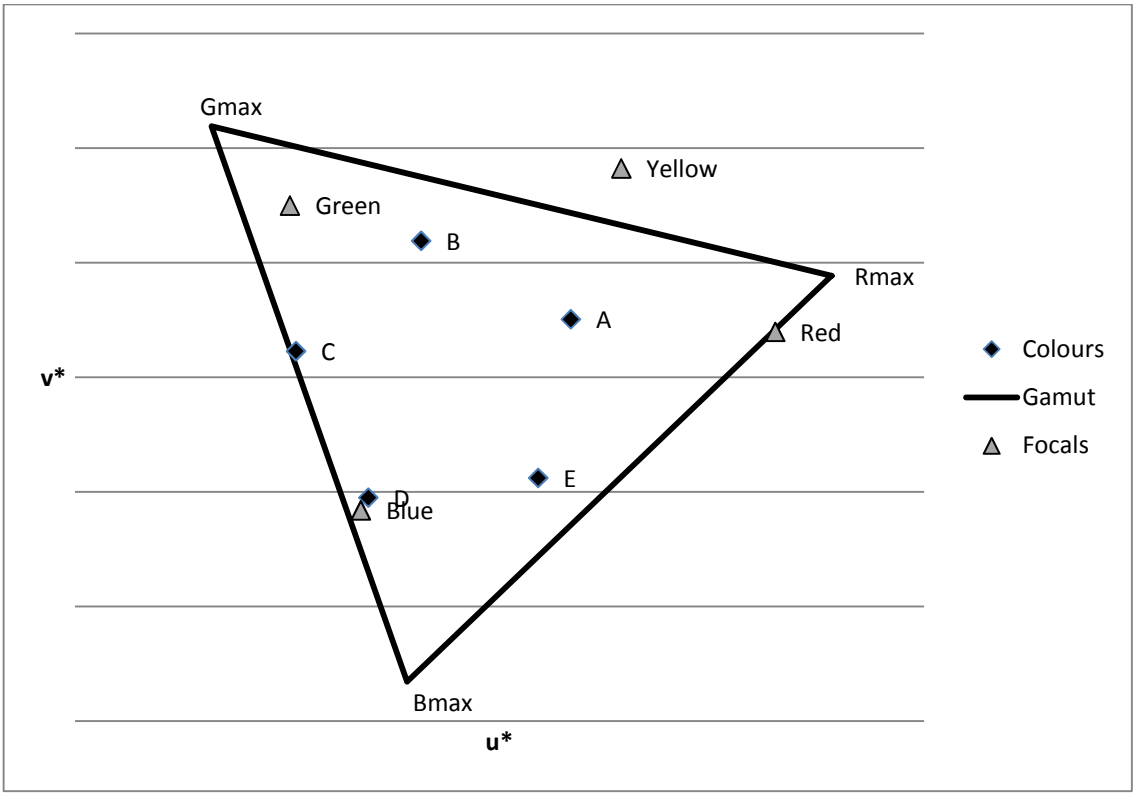


Figure 31: Final five colours shown with focal colours and monitor gamut

The values of these colours are shown in Table 3 along with an approximate representation. These colours were located in RGB space using a Cambridge Research ColorCAL (MK1) colorimeter. To facilitate comprehension of these colours, we provide the following descriptions: the colour labelled A is a brick red colour, B is chartreuse, C is turquoise, D is a sky blue, and E is mauve. Each colour presentation took place against a luminance-matched background of the D65 standard illuminant. All colour conversions consequently used D65 as the white point.

Label	Hue angle	u*	v*	x	Y	Y
A	0.436	54.378	25.357	0.411	0.353	20
B	1.693	-7.312	59.553	0.370	0.460	20
C	2.950	-58.898	11.449	0.237	0.375	20
D	4.206	-29.089	-52.477	0.234	0.257	20
E	5.463	40.920	-43.881	0.324	0.251	20

Table 3: Five colours used in the experiment

The selection of the auditory stimuli was subject to a similar degree of constraint. Since volume is a known contributor to crossmodal correspondences, a frequency range was selected to minimise variation in perceptual loudness (Fletcher & Munson, 1933; D. W.

Robinson & Dadson, 1956). Another consideration was the ability to analyse the results in terms of pitch (the perceptual correlate of frequency), note identity, and octave membership. The relationship between notes and octaves is significant: since each note identity is repeated in each ascending octave, both aspects are derived from pitch. Crucially however, the relationship between pitch and note identity is periodic, whereas that between pitch and octave membership is linear. Accordingly, twelve musical notes were selected for the present experiment. These range from C5 (the C in octave 5; 523.25 Hz) to A7 (the A in octave 7; 3520 Hz) and were evenly spaced such that the same four notes (C, D♯, F♯ and A) were represented in three octaves (5, 6 and 7).

5.3.1.3 Procedure

Participants were seated in front of the monitor in a dark room. Participants were asked to wear the headphones and adjust the volume to a comfortable level. Each trial the participants saw two coloured patches and heard one tone. Each pairing of colours was presented in both orders (i.e. BA as well as AB), resulting in 20 pairs of colours. All 240 combinations of colour-pairs and tones were presented in a random order in 10 blocks of 24.

The trials consisted of a one second fixation phase, followed by a one second presentation phase, finally followed by a one second response phase. These timings were kept deliberately brief in order to minimise opportunities for developing strategies: it is important that responses rely on intuition as far as possible. During the presentation phase the participants heard the tone and saw two coloured patches on the screen. The transition between the presentation and response phases was announced by the appearance of a message asking the participant to make a response. During the response phase the tone continued to play and the patches remained present on the screen. Participants responded by pressing the ‘a’ key to indicate the colour on the left or the ‘l’ key to indicate the colour on the right.

To minimise any bias resulting from differences in cumulative exposure, the timing of these phases was strictly controlled. Responses did not hide the coloured patches, stop the tone or advance the participant to the next trial. Similarly, a failure to respond during the response phase did not extend the presentation of the stimuli or delay the progression to the next trial. Between trials, participants heard a short burst of white noise to reduce carryover effects. This white noise was normally played at the same level as the experimental tones. In the event that a participant failed to register a response to a particular trial, the subsequent white noise was louder and acted as a warning.

5.3.2 Results

The frequency of "missed" trials was only 142 in 4800 trials, or 2.96%. The worst performing participant only missed 17 trials, equating to 7%.

Before seeking to establish the outcome of our main hypotheses, several checks were made to ensure that no systematic bias had been introduced. Since reaction times are known to be affected by crossmodal congruence, it is possible that the distribution of "missed" responses is symptomatic of the relationship we are looking for. To explore this possibility, we first filtered out all "missed" trials and performed a χ^2 (chi-squared) analysis on the identities of the stimuli present in those trials where a response was recorded. This proved to be non-significant ($\chi^2(4) = .101$, $p = .999$) suggesting that all colours were presented an equivalent number of times. Similarly another χ^2 test shows that, though the right-hand-side did have numerically more responses, there was no significant left-right bias in the responses ($\chi^2(1) = 1.113$, $p = .291$).

It is also interesting to see whether any one of the five colours is preferentially selected for or against. Unlike with the previous χ^2 tests, the distribution of colours selected by participants differs significantly from the distribution predicted by chance ($\chi^2(4) = 17.76$, $p = .001$). This is driven primarily by a preference for colour B (chartreuse) – typically considered to be aesthetically displeasing (Palmer & Schloss, 2010). When this colour is excluded from the analysis, the remaining colours conform reasonably well to a bias-free distribution ($\chi^2(3) = 1.449$, $p = .694$).

5.3.2.1 Main effect

To alleviate any bias effects (such as might be caused by "missed" trials), the total number of opportunities for each pitch and hue combination was tallied. The total number of times a hue was selected for each pitch can then be divided by this opportunity count to give a preference score. If the hue choices were all made randomly, this score would be consistently 50%. Scores over 50% indicate that, for the pitch in question, a hue has been selected for, whereas a score below 50% indicates that the hue has been selected against. These scores are plotted in Figure 32.

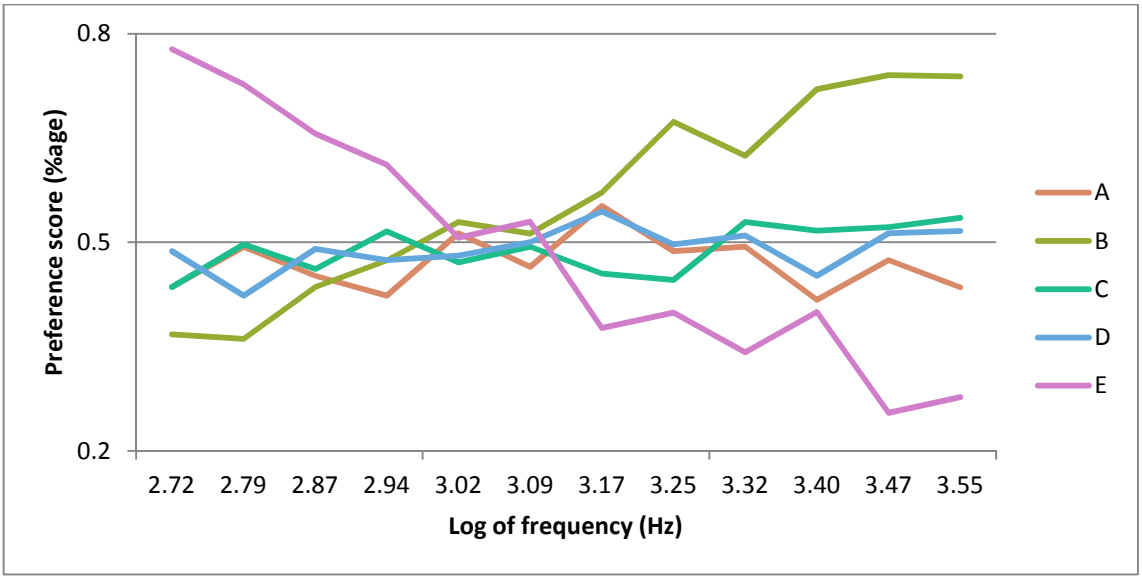


Figure 32: Plot of preference scores of 5 hues against log of pitch

Interestingly (and unexpectedly), Figure 32 shows an effect driven almost entirely by colours B (chartreuse; associated with higher pitches) and E (mauve; associated with lower pitches). This observation is confirmed by performing χ^2 tests on the selection frequencies for each hue: the distribution of selection frequencies is only significantly different from chance for colours B and E. The full results of these tests are contained in Table 4.

Colour	χ^2 (11) =	*p*
A	5.997	= .874 (ns)
B	55.090	< .001 (***)
C	4.227	= .963 (ns)
D	3.650	= .979 (ns)
E	105.727	< .001 (***)

Table 4: χ^2 test results for the selection frequencies of each hue

The fact that the plot in Figure 32 is monotonic and not periodic suggests that any hue-pitch relationship is driven by the raw pitch and not by any link to note identity. Since the pitches used were selected to repeat the same four note identities across three neighbouring octaves, this theory can easily be confirmed by collapsing the data by note identity and by octave membership. Once collapsed, these new datasets can be plotted in the same way as before (Figure 33 and Figure 34).

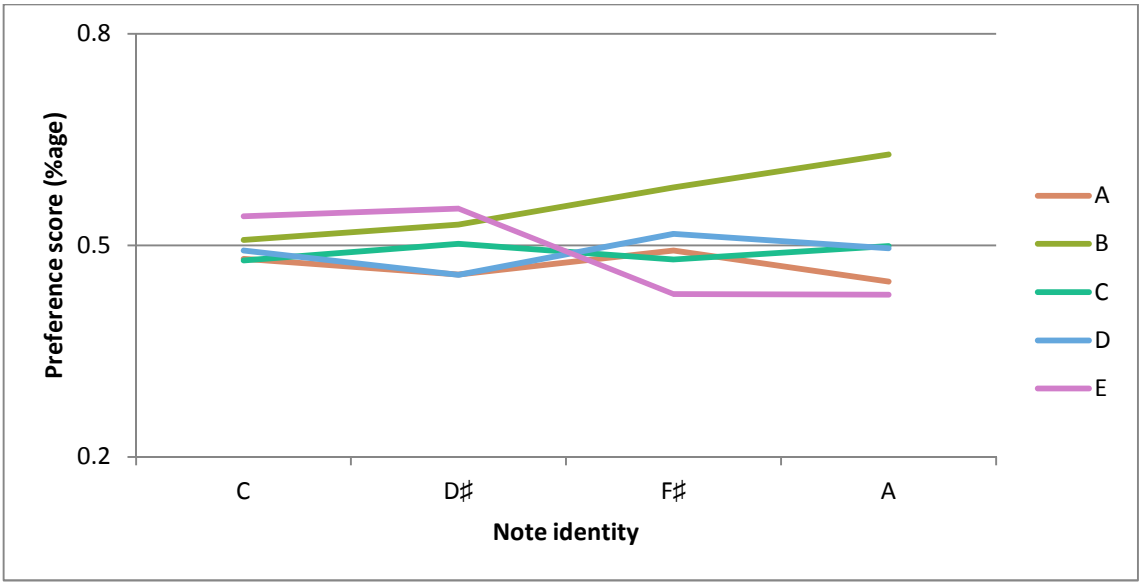


Figure 33: Preference score for each hue plotted against note identity

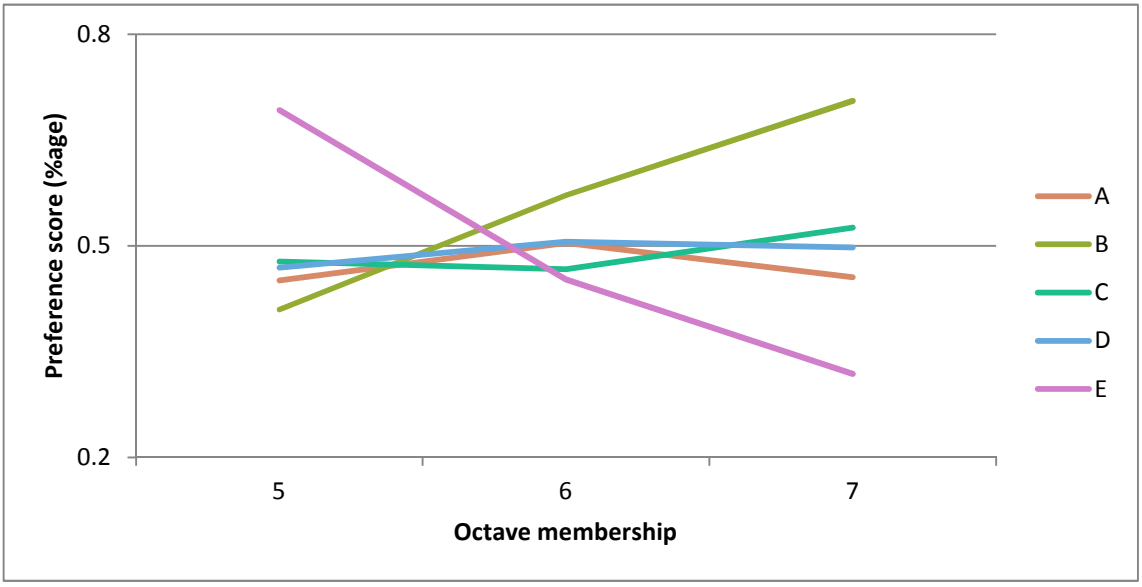


Figure 34: Preference score for each hue plotted against octave membership

Since the effect on hues B and E is much more pronounced in Figure 34 than Figure 33, the suggestion that hue may be related to note identity is further weakened. (Note also that the other three colours remain resolutely unaffected by pitch.) This difference can be quantified by comparing the χ^2 test results for each of the data reductions, as shown in Table 3.

	By note identity		By octave membership	
	χ^2 (3) =	*p*	χ^2 (2) =	*p*
A	1.305	= .728 (ns)	2.502	= .286 (ns)

B	7.083	= .069 (ns)	46.378	< .001 (***)
C	0.323	= .956 (ns)	2.347	= .309 (ns)
D	1.800	= .615 (ns)	0.892	= .640 (ns)
E	11.046	= .011 (*)	92.949	< .001 (***)

Table 5: χ^2 test results for the selection frequencies of each hue collapsed by note identity and octave membership

Note that the effect persists somewhat even when collapsed by note identity (Figure 33 and left hand side of Table 5), but is seriously reduced in comparison to collapsing by octave membership. The hypothesis that the effect is driven by note identity can still be rejected for two reasons. Firstly, the persistence of the effect may simply be caused by the overall difference in pitch between note identities: the average pitch of the C notes used in the experiment is lower than the A notes used. Secondly, the fact that the effect is present at all when the data is collapsed by octave strongly suggests that overall pitch is the cause of the effect. The fact that the effect is much stronger when collapsed by octave simply adds more weight to this argument.

Experiment 1 therefore shows that there is a cross-modal correspondence between hue and pitch. Unlike the anticipated manifestation of such a correspondence, the effect is only found for two of the five tested hues. This implies that the relationship is not linear. The second conclusion of experiment 1 is that the effect is driven by frequency rather than note identity. It is not possible to ascertain from experiment 1 whether this effect is relative or absolute. It is possible, for instance, that 500Hz is associated with purple and 3000Hz is associated with chartreuse. The more likely explanation is that participants quickly become familiar with the range of tones presented and begin to make judgements based on relative pitch. Experiment 2 seeks to clarify this ambiguity by presenting the lower half of the set of tones to one group and the higher half to another group.

5.4 Experiment 2

5.4.1 Methods

5.4.1.1 Participants

Two groups of ten students were recruited from the University of Sussex and were awarded course credits for their participation. The first group (6 female, aged between 18 and 21 years, mean age 19) were assigned to the “low” condition. The second group (8 female, aged between 18 and 21 years, mean age 19) were assigned to the “high” condition. As in Experiment 1, ethical approval was granted by the Life Sciences & Psychology Cluster-based

Research Ethics Committee at the University of Sussex and all participants reported normal hearing and normal (or corrected-to-normal) vision. None of these participants had previously taken part in Experiment 1.

5.4.1.2 Materials

As in Experiment 1, except for the splitting of the auditory stimuli into a "high" and "low" set of six notes apiece.

5.4.1.3 Procedure

After being randomly assigned to one of the two conditions, participants were briefed as in Experiment 1. Participants were blind to their assignment to an experimental condition. Procedurally, this experiment is almost identical to Experiment 1. The sole difference was the repetition of each colour-pair-note combination to compensate for the 50% reduction of notes.

5.4.2 Results

As before: the level of "missed" trials was low (3.33% in the low condition and 1.58% in the high condition); the frequencies of colours presented in non-missed trials match those predicted by chance (low: $\chi^2(4) = 0.053$, $p = 1$; high: $\chi^2(4) = 0.052$, $p = 1$); and there was no significant left-right bias in the responses (low: $\chi^2(1) = 0.043$, $p = .836$; high: $\chi^2(1) = 3.279$, $p = .070$). Unlike in Experiment 1, participants do not seem to be favouring a particular colour. Conversely, in the low condition participants seem to select *against* the colour labelled C (turquoise), whose exclusion leads to the restoration of a chance-level distribution (before: $\chi^2(4) = 11.500$, $p = .021$; after: $\chi^2(3) = 2.882$, $p = .410$). In the high condition, the distribution of responses is (borderline) non-significant without excluding any colours ($\chi^2(4) = 9.463$, $p = .051$). Indeed, in the high condition it is unclear which colour ought to be excluded, as both the maximum and minimum values deviate from the mean by an equivalent amount (1.1 standard deviations).

The same basic pattern found in Experiment 1 is found again in both conditions: colour B (chartreuse) is associated with high pitches and colour E (mauve) with lower pitches. This is illustrated by Figure 35. Significantly, the fact that this pattern is scaled and repeated (rather than divided over the two conditions), strongly suggests that the hue-pitch effect acts relative to the distribution of notes presented (i.e. is linked to polar values of high/low rather than the frequencies of the tones *per se*).

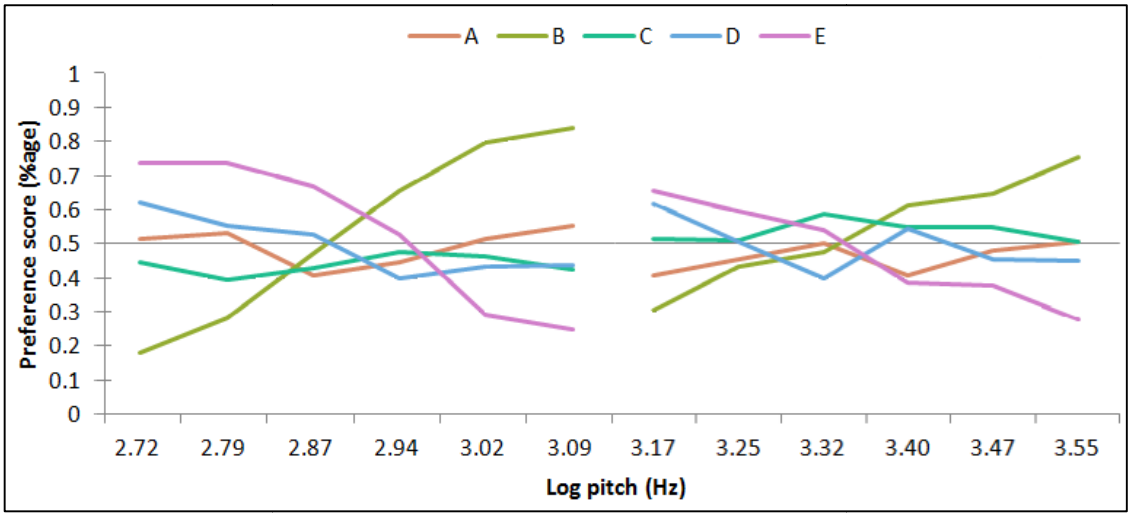


Figure 35: Low (left) and High (right) conditions both plotted for log of pitch against preference score

The fact that the pattern is stronger in the "low" condition than in the "high" condition is interesting. It is possible that this is simply down to chance, but this seems like an unsatisfactory explanation. On the other hand, it is possible that this discrepancy reflects the imperfections of the musical distribution in mirroring the subjective perception of pitch. This too, seems unsatisfactory, as the tones used were selected with these imperfections in mind: most of the deviance between perceptual and musical pitch occurs at frequencies less than 500Hz (Miśkiewicz & Rakowski, 2012; Stevens, 1937). Unfortunately, neither the significance nor the cause of this differential effect is presently well understood.

Nevertheless, as Figure 35 suggests, both B and E deviate significantly from chance in both conditions. Unexpectedly however, the colour labelled D also deviates from chance: significantly in the low condition and approaching significance in the high condition. The χ^2 statistics for these tests are presented below in Table 6.

	Low		High	
	χ^2 (5) =	*p*	χ^2 (5) =	*p*
A	5.008	= .415 (ns)	3.059	= .691 (ns)
B	106.079	< .001 (***)	39.955	< .001 (***)
C	1.415	= .923 (ns)	1.627	= .898 (ns)
D	12.553	= .028 (*)	9.727	= .083 (ns)
E	71.812	< .001 (***)	35.723	< .001 (***)

Table 6: χ^2 test results for the selection frequencies of each hue in both the "low" and "high" conditions

5.5 Post-hoc analysis

Experiment 1 demonstrates that there is a significant relationship between hue and pitch. The fact that the effect is stronger between octaves than between notes suggests that note identity does not play a role in this relationship. Experiment 2 shows that the effect is relative and scales to the range of contemporaneously presented pitches. The effect appears to be driven primarily by two of the colours. This suggests that hue itself is not the salient factor: if it were, we might expect to see intermediate effects for the colours between the two extremes. This non-linearity is not found in other crossmodal correspondences and consequently suggests that the relationship is mediated by a third factor. Given that the two other components of colour (lightness and chroma) were carefully controlled, hue must be acting so as to mediate some other attribute of colour.

We suggest that the factor driving this effect is familiarity with the geometry of perceptual colour space. All serious attempts to represent the extents of perceptual colour space in terms of hue, chroma and lightness result in irregular "colour solids". In other words, when representing these three dimensions as a solid, the resulting shape is never a recognisable – or easily describable – shape (e.g. a cube, prism or sphere). Examples of these colour solids are given in Figure 36.

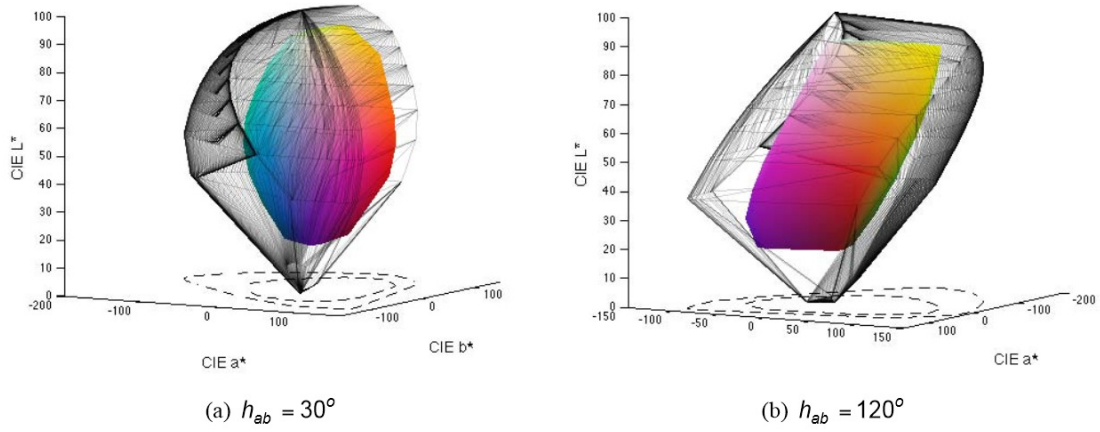


Figure 36: Examples of solid representations of colour space. Coloured solid corresponds to CIELAB. Mesh corresponds with MacAdam limits. Taken from (Heckaman & Fairchild, 2009)

Specifically, we posit that the observed effect is mediated by the lightness of the maximally saturated shade of a given hue. Physical limits determine the maximum chroma for a particular combination of hue and lightness. These are not uniform between hues and were computationally determined by David MacAdam (1935). These limits are what give hue sheets in colour atlases their distinctive shapes:

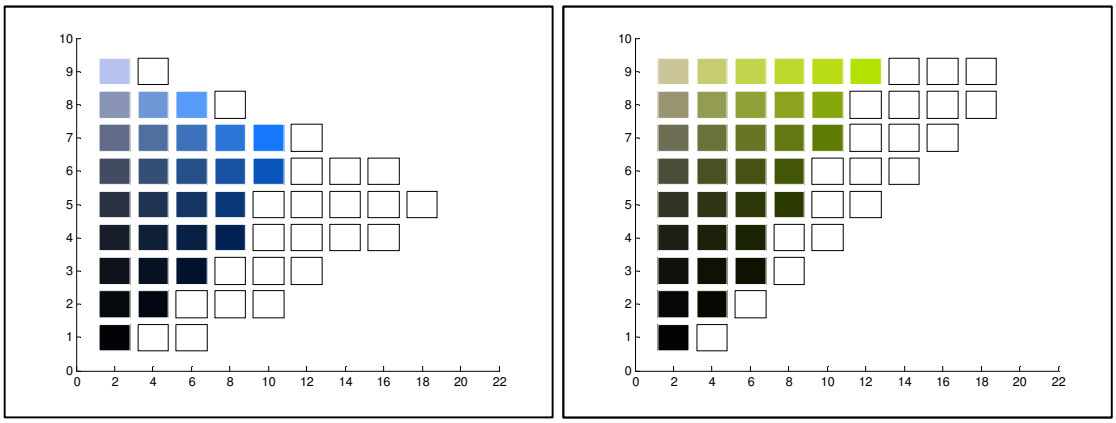


Figure 37: Munsell hue sheets for 10B and 5GY. Lightness ("value") is represented by the vertical axes and saturation ("chroma") is represented by the horizontal axes.

As shown in Figure 37, each hue has a distinctive lightness-saturation profile. These profiles tend to have a peak at a particular lightness where the maximum chroma for the hue may be obtained. We call this the "max chroma peak" and suggest that it is the lightness of this peak that drives the hue-pitch effect reported above.

The MacAdam limits were applied to the Munsell colour space during its renotation (Newhall, Nickerson, & Judd, 1943). The presentation of these limits in an accessible table makes the Munsell colour space a good platform for analysis. It does of course mean that our five colour stimuli need to be re-described in the Munsell colour space. Further, since the table only has a

"resolution" of 2.5 units of hue, our colours will need to be "rounded" to the nearest available data point. Our A becomes 2.5YR, B becomes 5GY, C becomes 10G, D becomes 10B, and E becomes 10P.

This translation is approximate, but sufficient. For a more exact MacAdam limits for our hues, we could use the recent algorithm proposed by Francisco Martínez-Verdú *et al.* (2007), which allows for precise calculation of any lightness, under one of several standard illuminants (including D65, as used in the present study). The implementation of this algorithm is beyond the scope of the present study, as the Munsell approximations are sufficient for the consequent analysis.

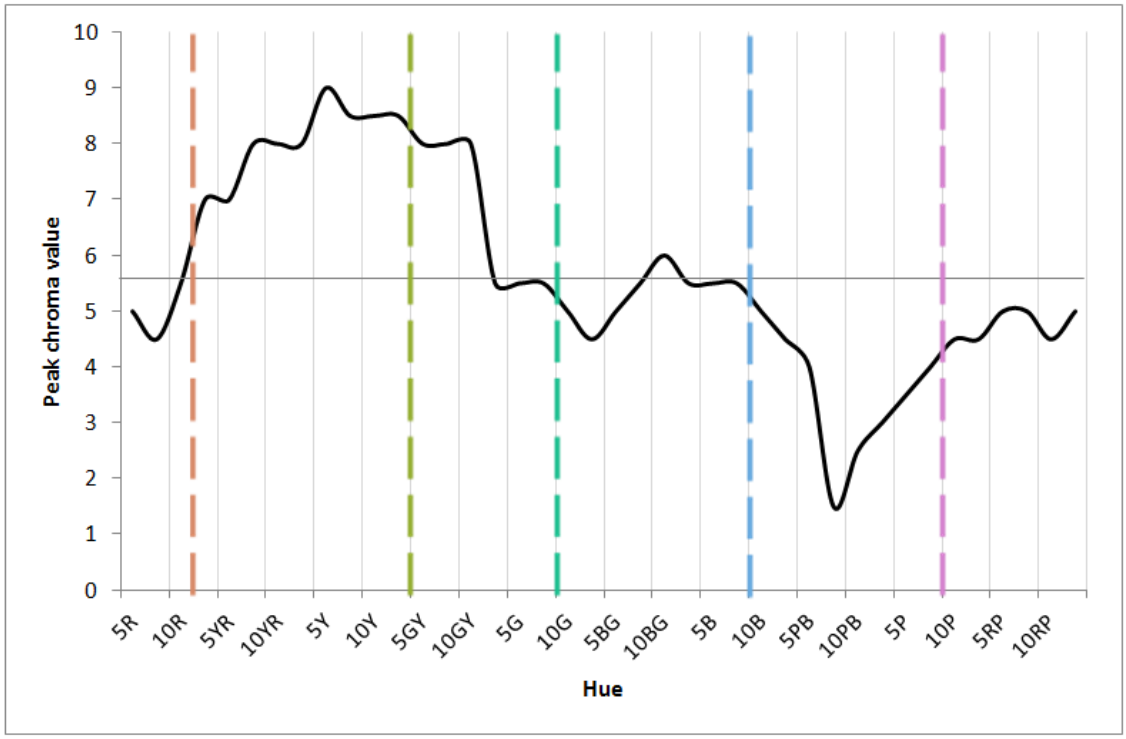


Figure 38: Value (lightness) of peak chroma across hues with experimental stimuli overlaid. The horizontal line represents the average peak chroma value (5.675)

Using the MacAdam limit data, it is possible to identify the value (lightness) at which the peak in perceivable chroma exists. For hues with flat peaks (i.e. where more than one value shares the maximum chroma) a simple mean was taken. Figure 38 shows these maxima plotted against hue and additionally overlays the colours used as experimental stimuli in the present study.

The two prominent regions in Figure 38 seem to match up approximately with the two colours identified as having driven the effect observed in both the experiments presented in the current study. It is possible to quantify this pattern by taking the average value for all chroma

peaks (5.675) and subtracting it from the value of the chroma peak for each of the experimental colours. These values are shown and ranked in Table 7. Note that colour B (chartreuse) is ranked first and colour E (purple) is ranked fifth. These rankings are a good match for the observed findings.

Colour	Munsell hue	Peak chroma value relative to mean peak chroma value	Peak value rank
A	2.5YR	+23%	2
B	5GY	+41%	1
C	10G	-12%	=3
D	10B	-12%	=3
E	10P	-21%	5

Table 7: Chroma-lightness bias factors for each colour stimuli

5.6 Discussion

In the above analysis, we argue that our results are indicative of a relationship between pitch and the lightness of the maximally saturated variant of a hue. This implies that the participants in these experiments rapidly and implicitly accessed their innate understanding of the shape of colour space. In this way, these results may be viewed more as a "top-down" effect of knowledge than a crossmodal correspondence in the strictest sense. Indeed, this effect is similar to known top-down effects in colour perception. For example, colour pairs that span a categorical boundary are judged more slowly and less accurately than an equivalently (perceptually) spaced pair of colours that are contained in a single category (Bornstein & Korda, 1984; Pilling, Wiggett, Özgen, & Davies, 2003).

Though the post-hoc analysis presents a good match between the model and the observed results, it is not perfect. One finding that is somewhat difficult to explain is the apparent presence of the hue-pitch effect for the sky blue colour stimuli labelled D. Though no such relationship was detected in Experiment 1, Experiment 2 did seem to suggest that this colour is associated with lower pitches. A significant relationship was found in the low condition and the result trended towards significance in the high condition (which also saw a less strong effect for the colours B and E). There are a number of possible causes.

The fact that the effect was only found when the number of participants was smaller suggests that it may be a false positive. Alternatively, the effect may be due to the fact that D was selected so as to align with the hue angle of focal blue. None of the other four colours are particularly close to any focal colours (see Figure 31). It is possible that there an unrelated process which links focal colours (or perhaps just focal blue) to pitch in a particular way. Finally, it is also possible that the effect is genuine and explainable within the model proposed above. This would imply that the approximate nature of our recoding into Munsell colour space had lead to an underestimation of the peak chroma value for D. An adjacent Munsell

hue (2.5PB) has the same peak chroma as E (purple), for example. It is easy to see how the coarseness of the Munsell coded MacAdam limits might lead to an error of this type. This serves to stress the importance for re-analysis with more accurate data.

An alternative explanation for the main experimental results may be that luminance corresponds with colour preferences. Palmer and Schloss (2010) found that chartreuse was universally ill-favoured, whereas violet was generally elicited more positive reactions. Could it be that the hue-pitch relationship we identify is mediated solely by preference? Do people associate chartreuse with higher pitches because they find both chartreuse and high-pitched sounds to be less pleasant than violet and low pitched sounds? Whilst this remains a possibility, the preference data is not a perfect match for the data reported in the present study: although violet is relatively highly favoured, it is beaten by blue and equal to cyan. If aesthetic preference is the primary mediator between pitch and hue, one would expect our colour D (blue) to be the most associated with lower pitches, not E (mauve).

Clearly further work is needed to verify these findings. Simply repeating the protocol with further participants would help to clear up the situation regarding colour D. As mentioned above, it would be highly interesting to compute the exact MacAdam limits for the colours used in this study: more accurate chroma-lightness bias factors may be a better fit for the results reported here.

Even based on coarse and imprecise data however, the current analysis also allows some predictions to be made. The colours at 7.5PB and 5Y have the most extreme peak chroma values at 1.5 (-74%) and 9 (+59%) respectively. If the model proposed herein is valid, these hues should result in the strongest pitch-related preference effects.

It would also be interesting to explore the inter-related effects of changing lightness and chroma values. One can imagine a protocol whereby colour stimuli of varying hue, chroma and lightness were selected in order to examine and quantify the interplay between known effects of lightness (Marks et al., 1987) and chroma (Ward et al., 2006), and that reported here.

More generally, the present study offers a novel example of a crossmodal correspondence between a pair of polar/Class I/prothetic sensory dimensions (pitch and lightness of the max chroma peak). On the other hand, perhaps the inverse is of greater theoretical interest – the present study leant weight to the non-existence of a suggested correspondence involving either of two non-polar/Class II/metathetic dimensions (hue and note identity).

In summary, although these findings are of little use to developers of sensory substitution devices, they are nevertheless of great theoretical interest. With regards to the former, visually impaired users will not benefit from the “bootstrapping” that occurs between (for instance) lightness and loudness without a distinct, comprehensive, uniform crossmodal correspondence between hue and pitch. With regards to the latter point, these findings suggest a novel cross-modal correspondence mediated by an innate appreciation for the geometry of colour space. Additionally, they seem to pose a broader question: does there exist a crossmodal correspondence genuinely driven by a metathetic dimension? Or are pairings of this nature the exclusive preserve of synaesthesia?

Chapter 6: Sensory Substitution Devices as Mediated Sensory Tools

Situating sensory substitution (and related) devices in the context of sensory tools

6.1 Abstract

Considerable effort has been devoted towards understanding sensory substitution devices in terms of their relationship to canonical sensory modalities. The approach taken in this paper is rather different, although complementary, in that we seek to define a broad conceptual space of "sensory tools" in which sensory substitution devices can be situated. This novel theoretical framework focuses on the similarities and differences between a wide range of tools (e.g. telescopes and cochlear implants) and, in doing so, provides new terminology to assist in comparisons. Additionally, by considering sensory substitution devices in the context of sensory tools, we are able to suggest some interesting questions for the field. With this approach, we hope to avoid the circularity inherent in previous attempts at defining sensory substitution and provide a better starting point to explore the effects of sensory tools, more generally, on the functioning of the nervous system.

A version of this work has been submitted for inclusion in the Proceedings of the British Academy.

6.2 Sensory substitution and sensory substitution devices

Sensory substitution is, broadly, a term used to describe the process of conveying information from one sensory modality via another. Typically the motivation is to restore some functionality of a lost or impaired sense. Indeed, since the creation of the first sensory substitution device (SSD) in the 1960s (Bach-y-Rita et al., 1969), sensory substitution has been closely associated with visual impairment: in this first device, and in many of the devices created in the years since, vision is the substituted sense.

Over the past 4 decades, sensory substitution has been defined primarily by the devices which enable it. Bach-y-Rita's first such system, called the Tactile Vision Sensory Substitution (TVSS) device, conveyed the image captured by camera to an 20-by-20 array of solenoid stimulators arranged on a dentist's chair so as to stimulate the skin on one's back (Bach-y-Rita et al., 1969). Each one of the vibrating solenoids acts as a tactile pixel (or "taxel") corresponding to the visual pixel occupying the equivalent space in the array from the camera.

Since the TVSS, there have been a number of subsequent systems that have been widely accepted as examples of Sensory Substitution Devices (SSDs). The most direct descendent of the TVSS is the Tongue Display Unit (TDU) developed by Bach-y-Rita's laboratory. Like the TVSS, the TDU operates by mapping pixels onto a two-dimensional array of taxels such that the intensity of vibration is set to the luminosity of the corresponding pixel. What sets the TDU apart is the fact that it uses electrodes instead of solenoids and stimulates the (highly conductive) surface of the tongue (Bach-y-Rita et al., 1998). The TDU has also been used to demonstrate that the substituted sense need not be limited to vision. It was successfully adapted to provide balance information to patients with bilateral vestibular damage (BVD), so as to restore a stable gait and posture (Tyler et al., 2003).

While the TVSS and TDU both use touch as the substituting sense, the vOICe (Meijer, 1992) was the first SSD to use audition and operates as a reverse spectrograph. Each image is scanned from left to right over the course of one second. As well as time, horizontal position is also conveyed using stereo panning. Frequencies are assigned along the Y axis, the amplitudes of which are based on the brightness of each pixel. A simple horizontal white line on a black background will therefore be represented as a single continuous tone, whereas a white line crossing a black background diagonally would sound like a single rising (or falling) tone. The audio representation of a complex natural scene is more difficult to describe – the authors highly recommend downloading the vOICe (http://www.seeingwithsound.com/winvoice.htm) and experiencing the sounds it generates firsthand.

Subsequent visual-auditory sensory substitution devices include the PSVA and the Vibe. As noted by the developers of the PSVA, “the time-multiplexed system of Meijer can be considered as working in real-time, but the refreshing rate (about one Hz) does not allow rapid sensory–motor interactions, as required in several tasks such as, e.g., mobility or reading” (Capelle et al., 1998). As a result of the implications of time-multiplexing, neither the PSVA nor the Vibe use time to denote any aspect of vision, but instead rely on our natural ability to localise sounds (Auvray et al., 2005; Capelle et al., 1998). The PSVA is particularly interesting due to its use of a weighted “foveal” region that gives the centre of the image as much auditory space as the periphery.

In general terms, sensory substitution devices may be defined by a tripartite component model consisting of an artificial sensor, a coupling system and a stimulator. As described by Claude Veraart: “artificial systems should include a transducer corresponding to the receptor organ, an encoder corresponding to the sensory processing system, and finally an interpreter corresponding to perccptual functions” (Veraart, 1989). This description appears to have been widely adopted: firstly in direct relation to Veraart’s initial observation (Arno et al., 1999), but later accepted *prima facie* (Lenay, Gapenne, Hanneton, Marque, & Genouëlle, 2003; Raj, Neuhaus, Moucheboeuf, Noorden, & Lecoutre, 2011; Ward & Wright, 2014). Other accounts broadly agree – one only alludes to the coupling system (Visell, 2009) and another refers to the stimulator as the “human-machine interface” (Bach-y-Rita & Kercel, 2003). This general, “tripartite model” of sensory substitution devices is illustrated in Figure 39.

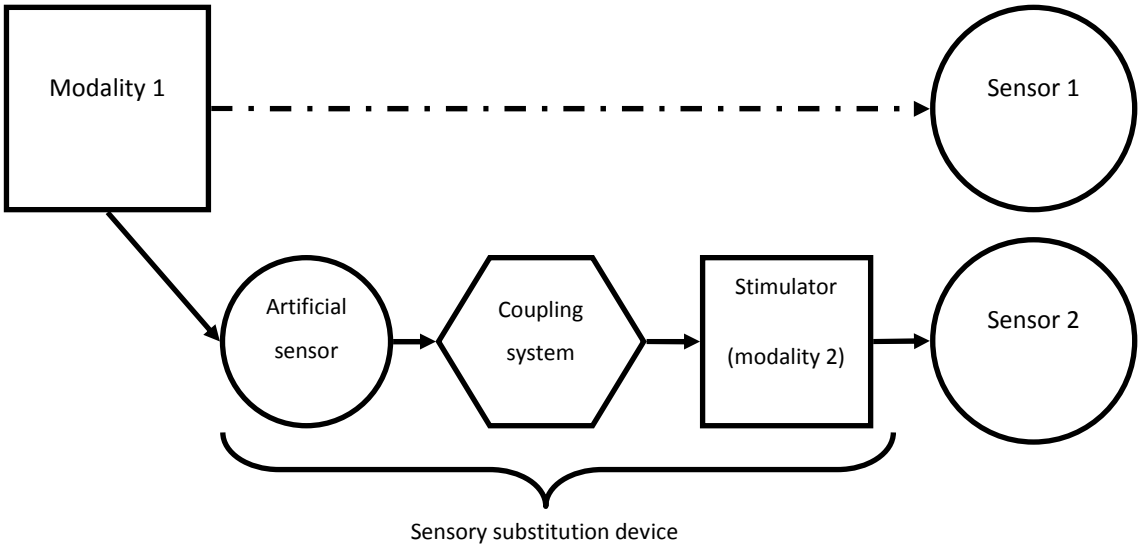


Figure 39: General schematic of the components of a sensory substitution device.

No matter how well we describe what we mean by a sensory substitution device however, this will never alleviate the need to robustly characterise the underlying phenomenon. Without placing sensory substitution in a wider context, circular definitions between the phenomenon and the devices that afford it are inevitable. Despite this, there have been relatively few attempts to pin down what sensory substitution really is.

6.2.1 Defining sensory substitution

In a 1991 review of tactile sensory substitution systems, sensory substitution was defined by Bach-y-Rita's team at the University of Wisconsin as "the use of one human sense to receive information normally received by another sense" (Kaczmarek, Webster, Bach-y-Rita, & Tompkins, 1991). Likewise, Charles Lenay and colleagues describe sensory substitution devices as "systems [that] transform stimuli characteristic of one sensory modality (for example, vision) into stimuli of another sensory modality (for example, touch)" (Lenay et al., 2003). To describe the devices mentioned above simply in these terms is to fail to capture the essence of what makes them interesting.

By way of advancing this situation, we have previously defined sensory substitution as "the artificial conveyance of rich, abstract sensory information of one sense via a different modality" (Ward & Wright, 2014). In the context of this definition we use "abstract" to mean non-symbolic. That is, the information should be in terms of physical or sensory properties, not a linguistic (or otherwise symbolic) interpretation. We likewise use "rich" to refer to the amount of information conveyed. The crucial test of these criteria is whether it is possible to make the reverse substitution and arrive back at a recognisable (excepting degradation due to bandwidth restrictions) representation of the original information; an abstract and rich representation should survive such a double-conversion.

In the current paper, we expand on this definition to contextualise sensory substitution devices as mediated sensory tools. In doing so, we aim to offer an accurate characterisation of this most interesting phenomenon.

In trying to identify the core principles of sensory substitution, it is tempting to start at the output: what sensory modality is produced by an SSD? Bach-y-Rita's group notably went so far as to title one of their papers "Seeing with the skin" (White et al., 1970). The philosopher Ned Block has suggested that the experiences reported by SSD users are primarily spatial in nature (Block, 2003). Others have described sensory substitution as a kind of "dual experience" (Humphrey, 2006, p. 58). It is undoubtedly a very interesting question, which will greatly

inform our understanding of human sensory processing. Indeed, even we are guilty of trying to answer this question (Ward & Wright, 2014).

But for all that the question is interesting, what will it tell us about the nature of a sensory substitution device? Is the answer necessary or sufficient to accurately characterise sensory substitution? Instead of this rather inward looking approach, we suggest that the best way to understand sensory substitution is by comparing and contrasting it to related phenomena.

6.2.2 The historical context

Historically, sensory substitution devices have rooted in the context of assistive technologies for the visually impaired. Indeed, by the definitions of sensory substitution provided by Bach-y-Rita and Lenay (see above) both seem to cover most systems designed to compensate for a sensory impairment. Bach-y-Rita goes further and actively identifies Braille, sign language and long canes as forms of sensory substitution (Bach-y-Rita, 1983).

In a 2003 review paper, Bach-y-Rita and Kercel reiterate their belief that Braille is "[t]he most successful sensory substitution system to the present" (Bach-y-Rita & Kercel, 2003). Alastair Haigh and colleagues stated recently that Braille is "an obvious and widely used example" of sensory substitution (Haigh et al., 2013). They go on to qualify this, by explaining that "This system only replaces a specific aspect of a modality however, namely language; substitution on a general level represents a much greater technical challenge". Whilst Braille undoubtedly uses one sense to "receive information normally received by another sense" (and so meets the definition offered by Bach-y-Rita), it does not seem to resemble devices like the TVSS in any other way. The question then, is whether Braille and sensory substitution are more than superficially conceptually related?

In the same review paper, Bach-y-Rita and Kercel also state that "[a] blind person using a cane is exhibiting another very successful simple sensory substitution system" (Bach-y-Rita & Kercel, 2003). Like Braille, a long cane is used to receive information usually received using sight. Note that Bach-y-Rita's definition deftly sidesteps the issue of the sensory modality being conveyed. This is important here, as it is not completely clear which sensory modality a cane receives. Is it visual in any way? This appears to be a hard position to defend. Is it tactile? This appears to be more accurate, but the mechanical information obtained at the tip of the cane isn't completely comparable to human touch. Could it be classed as a novel form of sense? These distinctions become important as stricter definitions of sensory substitution are applied. Lenay, for instance, stipulates that the information received must be transformed "into stimuli of another sensory modality" (Lenay et al., 2003). In other words, if a cane receives information in the

modality of touch and conveys this in the same modality, it could not be included in Lenay's definition of a sensory substitution device.

6.2.3 Related devices

In addition to the historical context, we must consider a few of the many interesting non-SSDs that have been discussed in the sensory substitution literature in recent years.

A perennially recurring example is a glove for people whose extremities have been damaged by leprosy or diabetes. This glove captures tactile sensations and refers these to tactors typically on the forehead (Bach-y-Rita & Kercel, 2003; Pax, R.A., Webster, & Radwin, 1989). It has also been suggested that similar technology could be used by astronauts. Not only could a tactile glove compensate for bulky clothing renders the wearers hands insensate, but it could also facilitate remote operation of robotic equipment (Sulzman & Wolfe, 1991).

In a much more clear-cut way than the long cane discussed above, such a glove would not be substituting any modality for another: it would merely refer the tactile information from the glove to the forehead (or wherever the tactors are situated). Clearly then, this device is not a sensory substitution device. It is however, achieving something interesting and relevant to sensory substitution research. Some classification for this phenomenon is therefore required.

And what of entirely novel sensory modalities? Our definition of sensory substitution implies that both the substituted and substituting senses are natural, but some of the most interesting devices have conveyed information from modalities of which we usually have no direct experience. This phenomenon has gained the name "sensory augmentation". One of the best known examples is the FeelSpace device, which conveys the compass bearing of the wearer to a belt of vibrating pads. The most northerly pad at any given moment is the one which is vibrates, which causes the wearer (over time) to incorporate this "sixth sense" into many aspects of everyday life (Nagel et al., 2005).

Similarly, some electrical engineering researchers have explored the insertion of small rare earth magnets into their fingertips in order to tangibly perceive magnetic fields (Hameed et al., 2010). As one implant recipient explains, even this simple interface between magnetism and touch can give rise to profoundly new experiences:

"*Each object has its own unique field, with different strength and "texture." I started holding my finger over almost everything that I could, getting a feeling for each object's invisible reach. [...] It has unlocked an entirely new world for me, one that I can touch and interact with in a very real way. While a magnet implant doesn't technically count as a "sixth sense" (it's more of*

an extension of our existing sense of touch), the way that the body internalizes these tiny magnetic vibrations feels truly foreign." – (Berg, 2012)

Another example of a device providing access to a novel sense could arguably be the Enactive Torch, which encodes proximity as the intensity of vibration. The case for calling this sensory augmentation is muddied by the fact that we can already perceive distance by touch, sight and (to some extent) sound. The creators however, argue persuasively that using the Enactive Torch, and experiencing only the distance to the nearest object, is unlike any other sense (Froese et al., 2012).

Neither FeelSpace, nor the tactile glove, nor any of the other examples can be described as sensory substitution devices, but all seem to be causing similar phenomenon. What then, should we call this grouping of interesting devices? Presently there is no agreed-upon nomenclature for the overarching class of devices into which SSDs and these other two types of device appear to fit.

6.2.4 On "substitution"

Finally, the word "substitution" implies a one-for-one replacement. That is, if we were to say that we substituted vision for audition, one may conclude that we had audition, but now have vision. Neither part of this conclusion would be correct for any current SSD. On the one hand, even the best systems provide only crude approximations of vision. On the other hand, there is no evidence to suggest that the substituting modality is lost in any way.

As we have discussed in an earlier paper (Ward & Wright, 2014, sec. 4.2), users of SSDs appear to experience both modalities simultaneously and apparently in a way mediated by allocation of attention. This phenomenon is well illustrated by the report of a long-term user of The vOICe: "*At lunch, I like to look at buildings. Some look beautiful and some sound nice – especially those with strong repeating vertical lines. The best ones though, are the ones that look <u>and</u> sound nice.*" (Unpublished interview.)

Interestingly, this conclusion is not controversial within the research community. The phenomenon has been, quite elegantly, characterised as a "complicated dual experience" (Humphrey, 2006, p. 58). Even Bach-y-Rita – from whom the term "sensory substitution" originates – is clear that users of SSDs do not lose sensation from the substituting modality (Bach-y-Rita, 2002, p. 501). Despite this, the term has stuck and it persists even among those who refute the idea that any substitution is taking place. Perhaps this persistence is due to the alliterative catchiness of the label, or perhaps it has more to do with the embodied promise of

a true substitution system. Either way, the term “substitution” is unhelpful to describe the phenomenon afforded by today’s devices.

6.3 Sensory substitution devices as sensory tools

As we have shown, the term “sensory substitution” is presently used too broadly to capture the essence of the phenomenon, yet fails to incorporate some of the most interesting examples of related technology. Moreover, we have suggested that the word “substitution” is misleading. What is clearly needed therefore is a systematic and well-defined nomenclature.

In addition to the current inadequacy of the current terminology, another issue impeding the progress of the research community is the lack of links to other areas of research. This is not an attempt to deny the inherent pluridisciplinary nature of sensory substitution research, which clearly draws on work in psychology, neuroscience, computer science, occupational therapy and mobility research, as well as many others. Instead, it is a comment on the present isolation of the devices and phenomena that are studied. It is difficult, for example, to generalise findings without the benefit of a widely understood and agreed-upon context.

We believe that the appropriate context for sensory substitution (and related) devices is that of “sensory tools”. They are tools in the sense that they are actively employed by a user to perform a task without being consumed in the process. Unlike more iconic examples of tools (e.g. a hammer) however, the operation they are used to accomplish is primarily sensory. That is, instead of mediating the way in which the user manipulates their surroundings, these tools mediate how their surroundings are perceived by the user.

We can therefore expand our earlier definition to fit sensory tools: sensory tools are devices whose primary function is to manipulate rich, abstract sensory information and present the product to a user who retains agency over the sensory experience. (In this context, we use agency to refer to real-time autonomous control.) Accordingly, examples of non-SSD sensory tools include telescopes and cochlear implants.

Another (possibly) non-SSD sensory tool is the long cane. Regardless of whether the conveyed sensory modality is visual, tactile or “spatial”, long canes easily meet our definition for being a sensory tool. They are actively employed by a user to perform a primarily sensory task. In doing so, they mediate the users perception of their surroundings.

6.3.1 Systematically classifying sensory tools

In order to meaningfully discuss the relationships we further propose that most (if not all) sensory tools may be classified on two dimensions. The first dimension describes the level of sensory change that the tool affects and has four levels:

1. **Compensatory prostheses.**
   Tools which manipulate sensory information in such a way as to restore some functionality of a sensory organ. This category includes glasses and hearing aids as well as cochlear and retinal implants. As the name suggests, the aim of these devices is to correct for some deficit, rather than add or change any natural experience.
2. **Within-sense referral (WSR) devices.**
   Tools which transform sensory information whilst retaining the modality. This could be to change the effective location of the sensor: examples of this include stethoscopes and mirrors. A WSR device could also act such as to change some other aspect of a sensor, like its orientation (i.e. prism glasses) or its intensity (i.e. night vision goggles).
3. **Between-sense referral (BSR) devices.**
   Tools which would previously have been called sensory substitution devices. Tools which afford "the artificial conveyance of rich, abstract sensory information of one sense via a different modality". As described above, these devices include the TVSS and the vOICe.
4. **Novel-sense referral (NSR) devices.**
   Tools which convey information from a novel sensory modality via an existing modality. Examples include the FeelSpace belt, the enactive torch and subdermal magnetic implants. These tools would previously have been called sensory augmentation devices.

These four categories represent increments on the scale of sensory change, but they can also group together to make coarser discriminations. The most obvious of these is the unisensory and bisensory split: compensatory prostheses and within-sense referral devices are unisensory, but between-sense referral devices and novel-sense referral devices are both bisensory. Alternatively, one might want to contrast the compensatory devices against the transformative devices (i.e. the other three).

The second dimension has just two levels: "direct" and "mediated". This distinction refers more to the implementation than the complexity. A direct sense tool is one that works primarily because of a natural physical or mechanical relationship between the source and target modalities. For instance, the magnetic implants described above would be classed as direct, because they rely on the natural relationship between magnets and magnetic fields to produce movements. A mediated sensory tool is one which requires some degree of digital processing in between the sensor and the stimulator. The TVSS and the vOICe are both

examples of mediated sensory tools, as are cochlear implants, the FeelSpace belt and thermal imaging.

As a consequence of the reliance on a natural occurring coupling system, direct sensory tools may afford a quicker path to competent use, but this is not a necessary feature. Using the same diagrammatical conventions as previously, the distinction can be represented as in Figure 40 below.

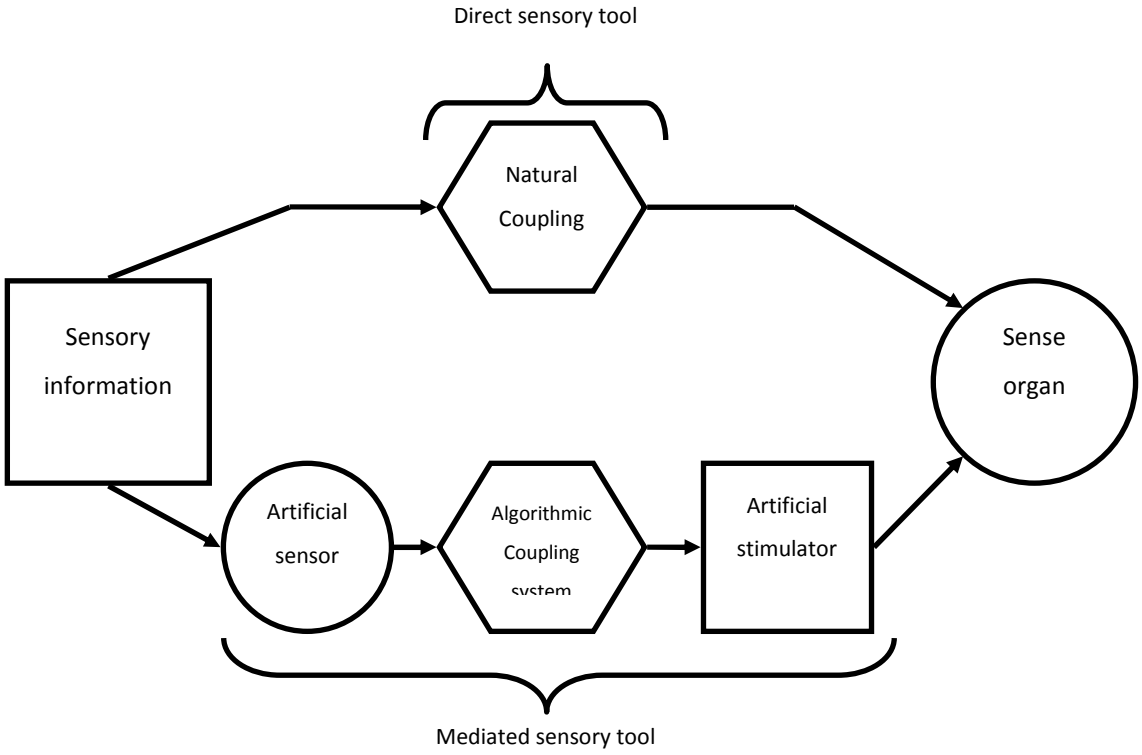


Figure 40: Comparison of direct and mediated sensory tools.

Since these two dimensions are orthogonal, there are 8 resulting combinations. It is possible to easily classify most sensory tools within this framework, but there may be grey areas. One such grey area is whether the Enactive Torch should be classed as a novel–sense referral, between-sense referral or within-sense device depending on whether the source modality is classed as proximity (i.e. a novel sense), vision or touch respectively.

Conversely, it is possible to find plenty of good examples for most of the 8 combinations, except for direct between-sense referral. It is possible that there are no natural physical or mechanical coupling systems between any of our sensory modalities, but this seems unlikely. We view this gap as a challenge: who will be able to create a direct between-sense referral device?

	Direct	Mediated
Compensatory prostheses	Glasses	Cochlear implant
Within-sense referral (WSR)	Periscope	Leprosy glove
Between-sense referral (BSR)		The vOICe
Novel-sense referral (NSR)	Magnet implants	FeelSpace

Table 8: Examples of different categories of sensory tool

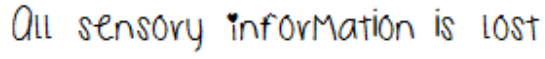
In addition to this four-by-two categorisation, sense tools must also be defined by the sensory modalities of their inputs and outputs. Compensatory prostheses and within-sense referral devices (by definition) output the same sensory modality as their input, meaning that they need only be described by one sense. For example, a pair of glasses should be described as a "direct visual compensatory prosthesis" and a glove for leprosy victims as an "mediated tactile within-sense referral device". Between-sense referral devices and novel-sense referral devices however, must be defined in terms of both a "source" and "target" modality. By existing convention, these are presented as "target-source". For example, the vOICe is a "mediated auditory-visual between-sense referral device" and magnetic implants are "direct tactile-magnetic novel-sense referral devices".

6.4 Interesting classifications

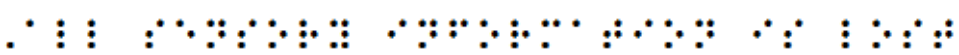
As well as classifying the obvious sensory tools like the TVSS (mediated tactile-visual BSR device) or a periscope (direct visual WSR device), we can use our proposed framework to discuss interesting examples like Braille, CCTV, oscilloscopes and virtual reality.

6.4.1 Braille is not a sensory tool

Unlike the long cane, Braille does not meet our definition of a sensory tool. Braille is a system of symbols. The purpose of symbols is to convey more meaning than the mere literal. If, in the process of translating or transliterating some meaning from one symbolic form to another, the sensory payload (or indeed modality) is changed, this new sensory form is not guaranteed to tell us anything about the original sensory form. Consider the written text below:



The letters can be transliterated to Braille cells, but these tell us nothing about the sensory experience of viewing the original letters:



In the transliteration from Latin script to Braille, we lose information about the size and shape of the letters, the font and any decorations. Conversely, we could transliterate between Latin script and Braille any number of times without losing any of the symbolic meaning.

None of this prevents the sensory information which underlies a symbol from being manipulated by a sensory tool. A mirror, for instance, is a sensory tool that has the ability to change the sensory form of written letters. Similarly, it is possible for visually impaired people to use the TDU to identify visual graphemes using an auditory signal. Accordingly, using the terms of our definition of a sensory tool, Braille might more accurately be thought of as part of "the surroundings", rather than the tool by which the perception of these are mediated. (An interesting case to consider is embossed text. Here the sensory payload is realised in two modalities. A tool that dynamically rendered text as embossed could therefore legitimately be classified as enabling between-sense referral.)

6.4.2 But CCTV and oscilloscopes can be

Some systems, such Closed Circuit TV (CCTV) or an oscilloscope, can be classed as sensory tools when used accordingly. For CCTV to count as a sensory tool, it must be live (i.e. not recorded) and controllable via a joystick or similar. Without these features, the operator is less user and more viewer; they retain no agency over the operation of the device.

An oscilloscope is interesting because it comprises only the coupling system and a visual stimulator. To make it do anything, it must be fed an electrical signal. Whether or not an oscilloscope-based system should be considered a sensory tool depends very much on the source of this electrical signal (does it, for example, give agency to the device user?). Oscilloscopes are often used to examine electronics. The visualisation of electronic waveforms in this manner could be described as a form of novel-sense referral. An alternate source of an electrical signal is a microphone. When an oscilloscope is displaying a visualisation of a sound, it acts like a reverse-vOICe and is arguably an example of a between-sense referral system. Despite this, it seems unlikely that an oscilloscope would be of any practical use to a person with a sensory impairment.

6.4.3 Anything is possible in virtual reality

Virtual reality and augmented reality are not forms of sensory tool, but can certainly act as platforms for the creation of virtual sensory tools. In a typical virtual reality set-up, the user

wears at least a set of immersive goggles and a pair of headphones. Accordingly, any sensory tool that targets either vision or audition may be simulated. Similarly, in augmented reality a user will typically view the world through a screen (perhaps mounted in goggles) that adds a computer generated layer.

Since a simulated world need not follow the physical laws of our universe, it would be possible to use virtual novel-sense referral to create entirely novel sensory modalities, devoid of any connection to any natural phenomena. Or, as demonstrated by a recent game produced at MIT, normal physical laws can be altered. “A Slower Speed of Light” allows the player to experience moving at speeds close to the speed of light (Kortemeyer, Tan, & Schirra, 2013). This manipulation represents virtual visual within-sense referral.

6.5 Theoretical implications

The point of forming a structured ontology is not merely to end up with more appropriate names for the members, but also allows us to systematically compare and contrast them. By making explicit our understanding of the relationships between these and other devices and phenomena, we don’t simply give them new labels: we form the basis for testable hypotheses that may further our understanding.

A fundamental difference between the four levels of sensory change is likely to be the degree of neural adaptation caused. One might expect that, provided that the deficit they are compensating for have not prevented normal development, compensatory prostheses cause very few changes to the brain. It is similarly reasonable to suppose that within-sense referral will require less neural adaptation than between-sense referral or novel-sense referral. Comparing these levels of sensory change will inform us about sensory processing and multisensory integration.

It would also be interesting to know whether there are any similarities in the processes by which people become accustomed to using different within-sense referral tools. In now-classic experiments, George Stratton investigated the effects of inverting his visual field by means of a pair of prism-glasses (Stratton, 1896, 1897). Generalisation studies suggest that, adaptation seems to occur primarily in the motor system, despite (or at least in addition to) phenomenological reports of perceptual adaptation (Morton & Bastian, 2004). We also know that this requires portions of the cerebellum (Morton & Bastian, 2004) and the parietal cortex (Pisella, Rode, Farnè, Boisson, & Rossetti, 2002). Would adaptation to a tactile glove involve these brain regions? Would this adaptation be subject to similar patterns of generalisation? Or

would this tactile adaptation be reliant on mechanisms more similar to the cortical remapping which causes referred sensation in amputees (Ramachandran, Stewart, & Rogers-Ramachandran, 1992)?

Another interesting area to explore could be the development of novel devices by systematically combining attributes from existing devices. As mentioned above, we see the lack of a direct between-sense referral device as a challenge. Beyond this, it would be trivial to generate a list of permutations of possible devices by taking the 4 by 2 categorical matrix and combining it with the N by N-1 matrix of available senses. What might a mediated audio-vestibular between-sense referral device be like? Could a mediated tactile-chronoception novel-sense referral device be useful to overcome jetlag?

Questions such as these allow the validity of this classification system to be tested empirically, whilst simultaneously expanding our understanding of the constituent components and the relationships between them.

6.6 The sensorimotor account

The sensorimotor account of perception holds that there is no sensation without action. That is, to gain useful information about our environment, it must be actively explored. Sensorimotor theory further states that senses are defined by the physical rules that link actions to perceptions. A sensorimotor contingency of vision, for instance, is that objects may occlude one another. This is not true for audition or touch. Conversely, audition and vision both allow distant objects to be perceived, whereas touch typically requires an object to be nearby. In their seminal paper introducing sensorimotor theory, O'Reagan and Noë describe perception as "give and take" (O'Regan & Noë, 2001b).

The sensorimotor account of perception maps well onto our proposed ontology of sense tools. In fact, if one accepts the sensorimotor account, one could assert that all sense tools are, in fact, sensorimotor tools. That is, they are all tools with which we can actively explore our sensory environment. Indeed, the four levels of sense tools described above can be accurately defined in terms of their effect on sensorimotor contingencies.

Compensatory prostheses typically alter some physical attribute of the sensory information, whilst preserving the sensorimotor contingencies. Within-sense referral devices preserve the basic structure of the sensorimotor contingencies of the source modality, but do alter specific properties of these contingencies. A periscope, for example, does not change the fact that

occlusion is a contingency of vision, but may change which objects occlude which other objects.

Between-sense referral devices are systems which take the sensorimotor contingencies of one sense and present them in a normally alien modality. When using The vOICe, for instance, contingencies such as occlusion, perspective and shadow are accessible through an auditory signal, despite not normally being a feature of audition. Novel-sense referral devices are interesting because they introduce entirely novel contingencies. Magnetism is the most obvious example source of entirely novel sensorimotor contingencies. As described above (section 6.2.3), both the FeelSpace belt and the subdermal implantation of rare earth magnets are extant examples of tactile-magnetic novel-sense referral devices – that is, they both give rise to perceptible vibrations in response to magnetic fields. Despite sharing the same broad category however, the use of these devices gives rise to two very different experiences. These differences are easily explained in terms of the sensorimotor contingencies that these devices afford.

The FeelSpace system operates by conveying the relative orientation of the geomagnetic field (Nagel et al., 2005). It does this by coupling a digital compass to a belt of tactile pads, such that only the pad facing north vibrates at any given time. In doing so, it exposes a set of novel sensorimotor contingencies. These include:

- Changing as the body rotates about an axis perpendicular to the ground.
- Remaining constant through bodily translation (i.e. movement without rotation), at least at scales encountered through most forms of transport.
- Being impervious to occlusion (except perhaps by a large magnet)
- A lack of sensitivity to (or communication of) amplitude

On the other hand, the subdermal magnetic implants lack any form of amplification and are consequently too weak to respond to magnetic north in any noticeable way. In fact, due to their simplicity and lack of moving parts, they barely respond to static fields at all. As reported by one implantee “magnetic surfaces provided almost no sensation at all” (Berg, 2012). Instead, these implants cause sensations when placed in dynamic magnetic fields, such as those commonly created by electrical devices. The result of this is that they expose a somewhat different set of novel sensorimotor contingencies. These include:

- Increasing intensity as an electronic device is approached by the finger containing the implant.
- Changing intensity as the orientation of the implant changes relative to the field.
- Vibrating with the same frequency as the source device.

- Being impervious to occlusion.

That one physical property can give rise to two such different forms of perceptual experience is perhaps comparable to mechanical vibrations, which can either give rise to sound or tactile sensation depending on their form and the sensor upon which they act. By defining sensory modalities in terms of their sensorimotor contingencies, we are able to describe these new forms of sensation in ways that would be difficult under most approaches to classifying the senses (e.g. by sensory receptor).

Thus, the sensorimotor account of perception is not merely compatible with the ontology of sense tools described in this paper, but is enhanced by and enhances it. Indeed, it may be tempting to refer to sensory tools as sensorimotor tools instead. However, since nothing proposed in the present paper relies on the sensorimotor account, we leave it to the reader to evaluate the sensorimotor account on its own merits.

6.7 Discussion

In this paper we have demonstrated that the term “sensory substitution” is of declining usefulness. We have consequently proposed a novel conceptual framework and associated nomenclature that sets sensory substitution alongside other sense-altering phenomena. This has allowed us to pose a series of questions and will undoubtedly allow others to go further. Finally, we have compared this framework with two other contemporary accounts of sensory substitution.

An unresolved aspect of our framework is the presence of grey areas, as alluded to above. Most obviously, categorising devices in our framework requires the source and target modality to be accurately identified. In cases where the source or target modality is ambiguous, the classification may consequently be unreliable. The Enactive Torch, for instance, could arguably convey vision, touch or the usually-alien modality of proximity. Since the target modality is touch, this would make the Enactive Torch a between-sense referral, within-sense referral or novel-sense referral device respectively.

If one accepts the sensorimotor account of perception, this grey area is more easily resolved. Rather than considering the modality *per se*, one can consider instead the sensorimotor contingencies conveyed by the device. For instance, the Enactive Torch may be said to convey two key sensorimotor contingencies. The first is that amplitude increases with proximity to an object. The second is that the area of sensitivity is a narrow beam extending from the front face of the device. The first contingency is shared with vision and audition, but the second is

entirely absent from the natural human experience. We can therefore conclude that the Enactive Torch is a form of novel-sense referral.

As mentioned in the section above, a systematic conceptual framework – by virtue of placing these interesting devices in a common context – will allow the research community to compare and contrast different sense tools. The accompanying nomenclature will facilitate these comparisons to be communicated clearly and unambiguously understood. We view this as an important step towards a coherent model of sensory tools and the perceptual alterations they afford.

We therefore encourage our colleagues to consider adopting our suggested nomenclature. We do not expect this to be rapid or abrupt. We fully expect any use of terms such as “between-sense referral” to coexist alongside terms like “sensory substitution” for at least the near future. The case of human echolocation provides an interesting example of such a transition in terminology. Originally known as “facial vision” (due to the belief that blind people could detect air pressure on their faces), it was suggested in 1893 to be an auditory phenomenon (Dresslar, 1893). Though this was conclusively shown to be so in the 1940s and 1950s (Cotzin & Dallenbach, 1950; Supa, Cotzin, & Dallenbach, 1944), the term “facial vision” only began to be displaced by “echo location” or “echo detection” in the 1960s (Rice, Feinstein, & Schusterman, 1965). Although the change in terminology suggested in this paper is more refinement than replacement, we hope that the precedent is sufficient for colleagues to consider rethinking the language (and associated conceptual structures) used when dealing with sensory tools. Specifically, it is our hope others in the field will make use of the greater descriptive power and reduced ambiguity offered by the classifications suggested in this paper.

Chapter 7: References

Alexander, M. S., Flodin, B. W. G., & Marigold, D. S. (2011). Prism adaptation and generalization during visually guided locomotor tasks. *Journal of Neurophysiology*, *106*(2), 860–871. doi:10.1152/jn.01040.2010

Alper, S., & Raharinirina, S. (2006). Assistive technology for individuals with disabilities: A review and synthesis of the literature. *Journal of Special Education*, *21*(2), 47–64.

Amedi, A., Stern, W. M., Camprodon, J. A., Bermpohl, F., Merabet, L. B., Rotman, S., … Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nat Neurosci*, *10*(6), 687–689. doi:10.1038/nn1912

Anderson, A. J., & Carpenter, R. H. S. (2010). Saccadic latency in deterministic environments: Getting back on track after the unexpected happens. *Journal of Vision*, *10*(14), 12. doi:10.1167/10.14.12

Arno, P., Capelle, C., Wanet-Defalque, M.-C., Catalan-Ahumada, M., & Veraart, C. (1999). Auditory coding of visual patterns for the blind. *Perception*, *28*(8), 1013 – 1029. doi:10.1068/p2607

Arno, P., De Volder, A. G., Vanlierde, A., Wanet-Defalque, M.-C., Streel, E., Robert, A., … Veraart, C. (2001). Occipital Activation by Pattern Recognition in the Early Blind Using Auditory Substitution for Vision. *NeuroImage*, *13*(4), 632–645. doi:10.1006/nimg.2000.0731

Arno, P., Vanlierde, A., Streel, E., Wanet-Defalque, M.-C., Sanabria-Bohorquez, S., & Veraart, C. (2001). Auditory substitution of vision: pattern recognition by the blind. *Applied Cognitive Psychology*, *15*(5), 509–519. doi:10.1002/acp.720

Arnott, S. R., & Alain, C. (2011). The auditory dorsal pathway: Orienting vision. *Neuroscience & Biobehavioral Reviews*, *35*(10), 2162–2173. doi:10.1016/j.neubiorev.2011.04.005

Asakawa, C., Takagi, H., Ino, S., & Ifukube, T. (2003). Maximum listening speeds for the blind. In *Proceedings of the 2003 International Conference on Auditory Display* (pp. 276–279). Boston, MA.

Auvray, M., Hanneton, S., Lenay, C., & O'Regan, K. (2005). There is something out there: distal attribution in sensory substitution, twenty years later. *Journal of Integrative Neuroscience*, *4*(4), 505–521.

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with "The vOICe." *Perception*, *36*(3), 416 – 430. doi:10.1068/p5631

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

Bach-y-Rita, P. (1983). Tactile Vision Substitution: Past and Future. *International Journal of Neuroscience*, *19*(1-4), 29–36.

Bach-y-Rita, P. (2002). Sensory Substitution and Qualia. In A. Noë & E. Thompson (Eds.), *Vision and Mind: Selected Readings in the Philosophy of Perception* (pp. 497–514). MIT Press.

Bach-y-Rita, P. (2004). Tactile Sensory Substitution Studies. *Annals of the New York Academy of Sciences*, *1013*(The Coevolution of Human Potential and Converging Technologies), 83–91. doi:10.1196/annals.1305.006

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision Substitution by Tactile Image Projection. *Nature*, *221*(5184), 963–964. doi:10.1038/221963a0

Bach-y-Rita, P., Kaczmarek, K. A., Tyler, M. E., & Garcia-Lara, J. (1998). Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note. *Journal of Rehabilitation Research and Development*, *35*(4), 427–430.

Bach-y-Rita, P., & Kercel, S. W. (2003). Sensory substitution and the human–machine interface. *Trends in Cognitive Sciences*, *7*(12), 541–546. doi:10.1016/j.tics.2003.10.013

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151–1162. doi:10.3758/BF03208371

Berg, D. (2012, March 22). I Have a Magnet Implant In My Finger. *Gizmodo*. Retrieved from http://gizmodo.com/5895555/i-have-a-magnet-implant-in-my-finger

Berk, T., Kaufman, A., & Brownston, L. (1982). A human factors study of color notation systems for computer graphics. *Commun. ACM*, *25*(8), 547–550. doi:10.1145/358589.358606

Bernstein, I. H., Eason, T. R., & Schurman, D. L. (1971). Hue-tone sensory interaction: a negative result. *Perceptual and Motor Skills*, *33*(3), 1327–1330.

Bertenthal, B. I., Banton, T., & Bradbury, A. (1993). Directional bias in the perception of translating patterns. *Perception*, *22*(2), 193 – 207. doi:10.1068/p220193

Biles, J. (1994). GenJam: A Genetic Algorithm for Generating Jazz Solos. In *Proceedings of the International Computer Music Association* (pp. 131–137). Retrieved from http://hdl.handle.net/2027/spo.bbp2372.1994.033

Bird, J., Marshall, P., & Rogers, Y. (2009). Low-fi skin vision: a case study in rapid prototyping a sensory substitution system. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (pp. 55–64). Swinton, UK, UK: British Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=1671011.1671018

Blickle, T., & Thiele, L. (1996). A Comparison of Selection Schemes Used in Evolutionary Algorithms. *Evolutionary Computation*, *4*(4), 361–394. doi:10.1162/evco.1996.4.4.361

Block, N. (2003). Tactile sensation via spatial perception. *Trends in Cognitive Sciences*, *7*(7), 285–286. doi:10.1016/S1364-6613(03)00132-3

Bologna, G., Deville, B., & Pun, T. (2010). Sonification of Color and Depth in a Mobility Aid for Blind People. In *The 16th International Conference on Auditory Displ ay (ICAD-2010)*. Washington D.C., USA. Retrieved from http://icad.org/Proceedings/2010/BolognaDevillePun2010.pdf

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological Research*, *46*(3), 207–222. doi:10.1007/BF00308884

Bridgelal Ram, M., Grocott, P. R., & Weir, H. C. M. (2008). Issues and challenges of involving users in medical device development. *Health Expectations*, *11*(1), 63–71. doi:10.1111/j.1369-7625.2007.00464.x

Brown, D., Macpherson, T., & Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception*, *40*(9), 1120 – 1135. doi:10.1068/p6952

Capalbo, Z., & Glenney, B. (2009). Hearing color: radical pluralistic realism and SSDs. Presented at the AP-CAP 2009, Tokyo, Japan.

Capelle, C., Trullemans, C., Arno, P., & Veraart, C. (1998). A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Bio-Medical Engineering*, *45*(10), 1279–1293. doi:10.1109/10.720206

Cheng, H. D., Jiang, X. H., Sun, Y., & Wang, J. (2001). Color image segmentation: advances and prospects. *Pattern Recognition*, *34*(12), 2259–2281. doi:10.1016/S0031-3203(00)00149-7

Chinn, M. D., & Fairlie, R. W. (2010). ICT Use in the Developing World: An Analysis of Differences in Computer and Internet Penetration. *Review of International Economics*, *18*(1), 153–167. doi:10.1111/j.1467-9396.2009.00861.x

Choe, C. S., Welch, R. B., Gilford, R. M., & Juola, J. F. (1975). The “ventriloquist effect”: Visual dominance or response bias? *Perception & Psychophysics*, *18*(1), 55–60. doi:10.3758/BF03199367

Cinzia, D. D., & Vittorio, G. (2009). Neuroaesthetics: a review. *Current Opinion in Neurobiology*, *19*(6), 682–687. doi:10.1016/j.conb.2009.09.001

Clark-Carter, D. D., Heyes, A. D., & Howarth, C. I. (1986). The efficiency and walking speed of visually impaired people. *Ergonomics*, *29*(6), 779–789. doi:10.1080/00140138608968314

Collignon, O., Lassonde, M., Lepore, F., Bastien, D., & Veraart, C. (2007). Functional Cerebral Reorganization for Auditory Spatial Processing and Auditory Substitution of Vision in Early Blind Subjects. *Cerebral Cortex*, *17*(2), 457–465. doi:10.1093/cercor/bhj162

Colonius, H., & Arndt, P. (2001). A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & Psychophysics*, *63*(1), 126–147. doi:10.3758/BF03200508

Congdon, N., Friedman, D., & Lietman, T. (2003). Important causes of visual impairment in the world today. *JAMA*, *290*(15), 2057–2060. doi:10.1001/jama.290.15.2057

Corneil, B. D., Wanrooij, M. V., Munoz, D. P., & Opstal, A. J. V. (2002). Auditory-Visual Interactions Subserving Goal-Directed Saccades in a Complex Scene. *Journal of Neurophysiology*, *88*(1), 438–454.

Costabile, M. F., Fogli, D., Fresta, G., Mussio, P., & Piccinno, A. (2003). Building environments for end-user development and tailoring. In *2003 IEEE Symposium on Human Centric Computing Languages and Environments, 2003. Proceedings* (pp. 31–38). doi:10.1109/HCC.2003.1260199

Costabile, M. F., Fogli, D., Mussio, P., & Piccinno, A. (2007). Visual Interactive Systems for End-User Development: A Model-Based Design Methodology. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *37*(6), 1029–1046. doi:10.1109/TSMCA.2007.904776

Cotzin, M., & Dallenbach, K. (1950). “Facial Vision:” The Role of Pitch and Loudness in the Perception of Obstacles by the Blind. *The American Journal of Psychology*, *63*(4), 485–515.

Cronly–Dillon, J., Persaud, K. C., & Blore, R. (2000). Blind subjects construct conscious mental images of visual scenes encoded in musical form. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1458), 2231–2238. doi:10.1098/rspb.2000.1273

Cronly-Dillon, J., Persaud, K., & Gregory, R. P. F. (1999). The perception of visual images encoded in musical form: a study in cross-modality information transfer. *Proceedings of the Royal Society B: Biological Sciences*, *266*(1436), 2427–2433.

Danilov, Y. P., Tyler, M. E., Skinner, K. L., & Bach-y-Rita, P. (2006). Efficacy of electrotactile vestibular substitution in patients with bilateral vestibular and central balance loss. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* (Vol. Supplement, pp. 6605–6609). doi:10.1109/IEMBS.2006.260899

Den Brikner, B. P. L. M., & Beek, P. J. (1996). Reading with magnifiers. *Ergonomics*, *39*(10), 1231–1248. doi:10.1080/00140139608964542

Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, *98*(1–3), 249–258. doi:10.1016/j.jphysparis.2004.03.011

Denham, J. (2008, November). An Evaluation of VoiceOver, the Macintosh Screen Reader. *AccessWorld*, *9*(6). Retrieved from http://www.afb.org/afbpress/pub.asp?DocID=aw090603

Dresslar, F. B. (1893). On the Pressure Sense of the Drum of the Ear and “Facial-Vision.” *The American Journal of Psychology*, *5*(3), 344–350. doi:10.2307/1410997

Dudley, N. J. (1990). Aids for visual impairment. *BMJ : British Medical Journal*, *301*(6761), 1151–1153.

Durette, B., Louveton, N., Alleysson, D., & Hérault, J. (2008). Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE. Presented at the Workshop on Computer Vision Applications for the Visually Impaired.

Earl, C., & Leventhal, J. (1999). A survey of windows screen reader users: results and recommendations. *Journal of Visual Impairment and Blindness*, *93*(3).

Farcy, R., & Damaschini, R. M. (2001). Guidance-assist system for the blind (Vol. 4158, pp. 209–214). Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Retrieved from http://adsabs.harvard.edu/abs/2001SPIE.4158..209F

Fischer, G., Giaccardi, E., Ye, Y., Sutcliffe, A. G., & Mehandjiev, N. (2004). Meta-design: a manifesto for end-user development. *Commun. ACM*, *47*(9), 33–37. doi:10.1145/1015864.1015884

Fletcher, H., & Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation. *The Journal of the Acoustical Society of America*, *5*(2), 82–108. doi:10.1121/1.1915637

Fotouhi, A., Hashemi, H., Mohammad, K., & Jalali, K. H. (2004). The prevalence and causes of visual impairment in Tehran: the Tehran Eye Study. *British Journal of Ophthalmology*, *88*(6), 740–745. doi:10.1136/bjo.2003.031153

Frens, M. A., & Opstal, A. J. V. (1995). A quantitative study of auditory-evoked saccadic eye movements in two dimensions. *Experimental Brain Research*, *107*(1), 103–117. doi:10.1007/BF00228022

Froese, T., McGann, M., Bigge, W., Spiers, A., & Seth, A. K. (2012). The Enactive Torch: A New Tool for the Science of Perception. *IEEE Transactions on Haptics*, *5*(4), 365–375. doi:10.1109/TOH.2011.57

Gilbert, C. E., Anderton, L., Dandona, L., & Foster, A. (1999). Prevalence of visual impairment in children: A review of available data. *Ophthalmic Epidemiology*, *6*(1), 73–82.

Gitau, S., Marsden, G., & Donner, J. (2010). After access: challenges facing mobile-only internet users in the developing world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2603–2606). New York, NY, USA: ACM. doi:10.1145/1753326.1753720

Gomez, J. D., Bologna, G., & Pun, T. (2010). Color-audio encoding interface for visual substitution: see color matlab-based demo. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility* (pp. 245–246). New York, NY, USA: ACM. doi:10.1145/1878803.1878853

Greated, M. (2011). The nature of sound and vision in relation to colour. *Optics & Laser Technology*, *43*(2), 337–347. doi:10.1016/j.optlastec.2009.06.005

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*(1), 36–41. doi:10.1016/S1364-6613(00)01571-0

Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of Sound Localization in Mammals. *Physiological Reviews*, *90*(3), 983–1012. doi:10.1152/physrev.00026.2009

Guarniero, G. (1974). Experience of tactile vision. *Perception*, *3*(1), 101 – 104. doi:10.1068/p030101

Haigh, A., Meijer, P., & Proulx, M. J. (2013). How well do you see what you hear? The acuity of visual-to-auditory sensory substitution. *Frontiers in Cognitive Science*, *4*, 330. doi:10.3389/fpsyg.2013.00330

Hameed, J., Harrison, I., Gasson, M. N., & Warwick, K. (2010). A novel human-machine interface using subdermal magnetic implants. In *2010 IEEE 9th International Conference on Cybernetic Intelligent Systems (CIS)* (pp. 1–5). doi:10.1109/UKRICIS.2010.5898141

Haupt, R. L., & Haupt, S. E. (2004). *Practical Genetic Algorithms* (2nd ed.). Wiley-Blackwell.

Hauskeller, M. (2012). My brain, my mind, and I: some philosophical assumptions of mind-uploading. *International Journal of Machine Consciousness*, *04*(01), 187–200. doi:10.1142/S1793843012400100

Havey, G. D., Gibson, P. L., Seifert, G. J., & Kalpin, S. (2007, December 11). Method and apparatus for sensory substitution, vision prosthesis, or low-vision enhancement utilizing thermal sensing. Retrieved from http://www.google.co.uk/patents?id=TiieAAAAEBAJ

Heath, M. (2012). NAudio (Version 1.6). Retrieved from http://naudio.codeplex.com/

Heckaman, R. L., & Fairchild, M. D. (2009). G0 and the gamut of real objects. In *Proceedings of the 11th Congress of the International Colour Association, Sydney, Australia*. Sydney, Australia. Retrieved from http://www.coloursociety.org.au/csa/aic/papers/0909015Final00212.pdf

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10–20. doi:http://dx.doi.org/10.1037/h0032606

Hopper, E., & Turton, B. (1999). A genetic algorithm for a 2D industrial packing problem. *Computers & Industrial Engineering*, *37*(1-2), 375–378. doi:16/S0360-8352(99)00097-2

Hubbard, T. L. (1996). Synesthesia-like Mappings of Lightness, Pitch, and Melodic Interval. *The American Journal of Psychology*, *109*(2), 219–238. doi:10.2307/1423274

Hull, J. M. (1992). *Touching the Rock: An Experience of Blindness* (Reprint.). Vintage Books.

Humphrey, N. (2006). *Seeing Red: A Study in Consciousness*. Harvard University Press.

Hunt, R. W. G., & Pointer, M. R. (2011). *Measuring Colour* (4th ed.). Hoboken: Wiley.

Ivanov, R. (2012). Real-time GPS track simplification algorithm for outdoor navigation of visually impaired. *Journal of Network and Computer Applications*, *35*(5), 1559–1567. doi:10.1016/j.jnca.2012.02.002

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the “ventriloquism” effect. *Perceptual and Motor Skills*, *37*(3), 967–979.

Jacobson, H. (1950). The Informational Capacity of the Human Ear. *Science*, *112*(2901), 143–144. doi:10.1126/science.112.2901.143

Jacobson, H. (1951). The Informational Capacity of the Human Eye. *Science*, *113*(2933), 292–293. doi:10.1126/science.113.2933.292

Jared. (2008, September 29). *How can you program if you’re blind? Stackoverflow*. Retrieved April 1, 2013, from http://stackoverflow.com/revisions/148880/3

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 1–10.

Jay, M. F., & Sparks, D. L. (1987). Sensorimotor integration in the primate superior colliculus. I. Motor convergence. *Journal of Neurophysiology*, *57*(1), 22–34.

Johnston, V. S., & Franklin, M. (1993). Is beauty in the eye of the beholder? *Ethology and Sociobiology*, *14*(3), 183–199. doi:10.1016/0162-3095(93)90005-3

Kaczmarek, K. A. (2011). The tongue display unit (TDU) for electrotactile spatiotemporal pattern presentation. *Scientia Iranica*, *18*(6), 1476–1485. doi:10.1016/j.scient.2011.08.020

Kaczmarek, K. A., & Haase, S. J. (2003). Pattern identification as a function of stimulation on a fingertip-scanned electrotactile display. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*(3), 269–275. doi:10.1109/TNSRE.2003.816874

Kaczmarek, K. A., Tyler, M. E., & Bach-y-Rita, P. (1997). Pattern identification on a fingertip-scanned electrotactile display. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1997* (Vol. 4, pp. 1694–1696 vol.4). IEEE. doi:10.1109/IEMBS.1997.757047

Kaczmarek, K. A., Webster, J. G., Bach-y-Rita, P., & Tompkins, W. J. (1991). Electrotactile and vibrotactile displays for sensory substitution systems. *IEEE Transactions on Biomedical Engineering*, *38*(1), 1–16. doi:10.1109/10.68204

Keeley, B. L. (2002). Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. *The Journal of Philosophy*, *99*(1), 5–28. doi:10.2307/3655759

Kim, J.-K., & Zatorre, R. J. (2008). Generalized learning of visual-to-auditory substitution in sighted individuals. *Brain Research*, *1242*, 263–275. doi:doi: DOI: 10.1016/j.brainres.2008.06.038

Kirillov, A. (2012). AForge.NET Framework (Version 2.2.4.0). Retrieved from http://www.aforgenet.com/aforge/framework/

Kokjer, K. J. (1987). The Information Capacity of the Human Fingertip. *IEEE Transactions on Systems, Man and Cybernetics*, *17*(1), 100–102. doi:10.1109/TSMC.1987.289337

Kortemeyer, G., Tan, P., & Schirra, S. (2013). A Slower Speed of Light: Developing Intuition about Special Relativity with Games. In *Proceedings of the International Conference on the Foundations of Digital Games* (pp. 400–402). ACM New York. Retrieved from http://www.fdg2013.org/program/festival/openrelativity.pdf

Kuhn, G., & Benson, V. (2007). The influence of eye-gaze and arrow pointing distractor cues on voluntary eye movements. *Perception & Psychophysics*, *69*(6), 966–971. doi:10.3758/BF03193934

Kumar, S., Forster, H. M., Bailey, P., & Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *The Journal of the Acoustical Society of America*, *124*, 3810. doi:10.1121/1.3006380

Kupers, R., Chebat, D. R., Madsen, K. H., Paulson, O. B., & Ptito, M. (2010). Neural correlates of virtual route recognition in congenital blindness. *Proceedings of the National Academy of Sciences*, *107*(28), 12716 –12721. doi:10.1073/pnas.1006199107

Kupers, R., Fumal, A., de Noordhout, A. M., Gjedde, A., Schoenen, J., & Ptito, M. (2006). Transcranial magnetic stimulation of the visual cortex induces somatotopically organized qualia in blind subjects. *Proceedings of the National Academy of Sciences*, *103*(35), 13256–13260. doi:10.1073/pnas.0602925103

Lane, D. ., McNicholas, J., & Collis, G. . (1998). Dogs for the disabled: benefits to recipients and welfare of the dog. *Applied Animal Behaviour Science*, *59*(1–3), 49–60. doi:10.1016/S0168-1591(98)00120-8

Lenay, C., Gapenne, O., Hanneton, S., Marque, C., & Genouëlle, C. (2003). Sensory substitution: Limits and perspectives. In Y. Hatwell, A. Streri, & E. Gentaz (Eds.), *Touching for Knowing: Cognitive psychology of haptic manual perception (Advances in Consciousness Research)*. John Benjamins Publishing.

Leventhal, J. (2005, September). Not What the Doctor Ordered: A Review of Apple's VoiceOver Screen Reader. *AccessWorld*, *6*(5). Retrieved from http://www.afb.org/afbpress/pub.asp?DocID=aw060505&select=11

Leventhal, J., & Holborow, R. (2001, May). Who Are the Players: Reviews of Hardware and Software Digital Talking Book Players. *AccessWorld*, *2*(3). Retrieved from http://www.afb.org/afbpress/pub.asp?DocID=aw020303

Lisetti, C. L., & Nasoz, F. (2004). Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Advances in Signal Processing*, *2004*(11), 1672–1687. doi:10.1155/S1110865704406192

Loomis, J. M., Golledge, R. G., & Klatzky, R. L. (2006). Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence*, *7*(2), 193–203.

Loveday, K. (2013). MicroTimer (Version 10). Retrieved from http://www.codeproject.com/Articles/98346/Microsecond-and-Millisecond-NET-Timer

MacAdam, D. L. (1935). Maximum Visual Efficiency of Colored Materials. *Journal of the Optical Society of America*, *25*(11), 361–367. doi:10.1364/JOSA.25.000361

Manish. (2011, March 28). *How can you program if you're blind? Stackoverflow*. Retrieved April 1, 2013, from http://stackoverflow.com/revisions/5461220/1

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving Similarity and Comprehending Metaphor. *Monographs of the Society for Research in Child Development*, *52*(1), i–100. doi:10.2307/1166084

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 517–534. doi:http://dx.doi.org/10.1037/0096-1523.12.4.517

Martínez-Verdú, F., Perales, E., Chorro, E., de Fez, D., Viqueira, V., & Gilabert, E. (2007). Computation and visualization of the MacAdam limits for any lightness, hue angle, and light source. *Journal of the Optical Society of America A*, *24*(6), 1501–1515. doi:10.1364/JOSAA.24.001501

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: tests of the semantic coding hypothesis. *Perception*, *28*(7), 903 – 923. doi:10.1068/p2866

Matlab (Version 2013b). (2013). Natick, Massachusetts, United States: The MathWorks, Inc. Retrieved from http://www.mathworks.co.uk/products/matlab/

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *, Published Online: 23 December 1976; | doi:10.1038/264746a0*, *264*(5588), 746–748. doi:10.1038/264746a0

McMorrow, G., Wang, X., & Whelan, P. F. (1997). Color-to-speech sensory substitution device for the visually impaired, 272–281. doi:10.1117/12.285572

Meers, S., & Ward, K. (2004). A vision system for providing 3D perception of the environment via transcutaneous electro-neural stimulation. In *Eighth International Conference on*

Information Visualisation, 2004. IV 2004. Proceedings (pp. 546–552). doi:10.1109/IV.2004.1320198

Meijer, P. (1992). An experimental system for auditory image representations. *Biomedical Engineering, IEEE Transactions on*, *39*(2), 112–121.

Meijer, P. (2013). *The vOICe User Group: The world's largest sensory substitution network*. *seeingwithsound.com*. Retrieved from http://www.seeingwithsound.com/subscribe.htm

Melara, R. D., & O'Brien, T. P. (1987). Interaction Between Synesthetically Corresponding Dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323–336.

Merabet, L. B., Battelli, L., Obretenova, S., Maguire, S., Meijer, P., & Pascual-Leone, A. (2009). Functional recruitment of visual cortex for sound encoded object identification in the blind. *NeuroReport*, *20*(2), 132–138. doi:10.1097/WNR.0b013e32832104dc

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*(3), 640–662.

Mesz, B., Trevisan, M. A., & Sigman, M. (2011). The taste of music. *Perception*, *40*(2), 209.

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: a class of displays on the reality-virtuality continuum. In *Proc. SPIE 2351, Telemanipulator and Telepresence Technologies* (Vol. 2351, pp. 282–292). doi:10.1117/12.197321

Miśkiewicz, A., & Rakowski, A. (2012). A psychophysical pitch function determined by absolute magnitude estimation and its relation to the musical pitch scale. *The Journal of the Acoustical Society of America*, *131*(1), 987. doi:10.1121/1.3651094

Montandon, A. (n.d.). *Colourblind Eyeborg Colours to Sound*. *adammontandon.com*. Retrieved from http://www.adammontandon.com/neil-harbisson-the-cyborg/

Mørch, A. I., Stevens, G., Won, M., Klann, M., Dittrich, Y., & Wulf, V. (2004). Component-based technologies for end-user development. *Communications of the ACM*, *47*(9), 59. doi:10.1145/1015864.1015890

Morgan, G. A., Goodson, F. E., & Jones, T. (1975). Age Differences in the Associations between Felt Temperatures and Color Choices. *The American Journal of Psychology*, *88*(1), 125–130. doi:10.2307/1421671

Morton, S. M., & Bastian, A. J. (2004). Prism Adaptation During Walking Generalizes to Reaching and Requires the Cerebellum. *Journal of Neurophysiology*, *92*(4), 2497–2509. doi:10.1152/jn.00129.2004

Moulster, A., & Stockman, T. (2011). On the Road to Design: Developing a Sonified Route Navigator for Cyclists. In *Proceedings of the 17th International Conference on Auditory Display (ICAD2011), Budapest, Hungary. 20-23 June, 2011. International Community for Auditory Display, 2011.* Budapest, Hungary.

Müller, J. (1826). *Zur vergleichenden Physiologie des Gesichtssinnes des Menschen und der Tiere*. Leipzig: C. Knobloch.

Nagel, S. K., Carl, C., Kringe, T., Märtin, R., & König, P. (2005). Beyond sensory substitution—learning the sixth sense. *Journal of Neural Engineering*, *2*(4), R13–R26. doi:10.1088/1741-2560/2/4/R02

Nestor, A., & Tarr, M. J. (2008). The segmental structure of faces and its use in gender recognition. *Journal of Vision*, *8*(7).

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final Report of the O.S.A. Subcommittee on the Spacing of the Munsell Colors. *Journal of the Optical Society of America*, *33*(7), 385–411. doi:10.1364/JOSA.33.000385

Norrsell, U., Finger, S., & Lajonchere, C. (1999). Cutaneous sensory spots and the “law of specific nerve energies”: history and development of ideas. *Brain Research Bulletin*, *48*(5), 457–465. doi:10.1016/S0361-9230(98)00067-7

Nuckolls, J. B. (1999). The Case for Sound Symbolism. *Annual Review of Anthropology*, *28*, 225–252. doi:10.2307/223394

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *46*(3), 461–488. doi:10.1037/h0084327

O'Regan, J. K., & Noë, A. (2001a). A Sensorimotor Account of Vision and Visual Consciousness. *Behavioral and Brain Sciences*, *24*(05), 939–973. doi:10.1017/S0140525X01000115

O'Regan, J. K., & Noë, A. (2001b). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, *129*(1), 79–103. doi:10.1023/A:1012699224677

Ortiz, T., Poch, J., Santos, J. M., Requena, C., Martínez, A. M., Ortiz-Terán, L., … Pascual-Leone, A. (2011). Recruitment of Occipital Cortex during Sensory Substitution Training Linked to Subjective Experience of Seeing in People with Blindness. *PLoS ONE*, *6*(8), e23264. doi:10.1371/journal.pone.0023264

Owen, C. G., Fletcher, A. E., Donoghue, M., & Rudnicka, A. R. (2003). How big is the burden of visual loss caused by age related macular degeneration in the United Kingdom? *British Journal of Ophthalmology*, *87*(3), 312–317. doi:10.1136/bjo.87.3.312

Palmer, S. E., Langlois, T., Tsang, T., Schloss, K. B., & Levitin, D. J. (2011). *Color, Music, and Emotion*. Presented at the Vision Sciences Society 11th Annual Meeting, Naples, FL, USA. Retrieved from http://socrates.berkeley.edu/~plab/CME.html

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, *107*(19), 8877–8882. doi:10.1073/pnas.0906172107

Parise, C. V., & Spence, C. (2009). "When Birds of a Feather Flock Together": Synesthetic Correspondences Modulate Audiovisual Integration in Non-Synesthetes. *PLoS ONE*, *4*(5), e5664. doi:10.1371/journal.pone.0005664

Parver, L. M. (1986). Eye trauma: The neglected disorder. *Archives of Ophthalmology*, *104*(10), 1452–1453. doi:10.1001/archopht.1986.01050220046022

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, *18*(1), 49–65. doi:10.1016/0167-8760(84)90014-X

Patterson, R. D. (1986). Spiral Detection of Periodicity and the Spiral Form of Musical Scales. *Psychology of Music*, *14*(1), 44–61. doi:10.1177/0305735686141004

Pax, R.A., J., Webster, J. G., & Radwin, R. G. (1989). A conductive polymer sensor for the measurement of palmar pressures. In *Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century., Proceedings of the Annual International Conference of the IEEE Engineering in* (pp. 1483–1484 vol.5). doi:10.1109/IEMBS.1989.96300

Petrazzini, B., & Kibati, M. (1999). The Internet in developing countries. *Commun. ACM*, *42*(6), 31–36. doi:10.1145/303849.303858

Pilling, M., Wiggett, A., Özgen, E., & Davies, I. R. L. (2003). Is color “categorical perception” really perceptual? *Memory & Cognition*, *31*(4), 538–551. doi:10.3758/BF03196095

Pisella, L., Rode, G., Farnè, A., Boisson, D., & Rossetti, Y. (2002). Dissociated long lasting improvements of straight-ahead pointing and line bisection tasks in two hemineglect patients. *Neuropsychologia*, *40*(3), 327–334. doi:10.1016/S0028-3932(01)00107-5

Poirier, C., De Volder, A., Tranduy, D., & Scheiber, C. (2007). Pattern recognition using a device substituting audition for vision in blindfolded sighted subjects. *Neuropsychologia*, *45*(5), 1108–1121. doi:10.1016/j.neuropsychologia.2006.09.018

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, *13*(3), 278–285. doi:10.1037/h0072651

Ptito, M., Moesgaard, S. M., Gjedde, A., & Kupers, R. (2005). Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. *Brain*, *128*(3), 606–614. doi:10.1093/brain/awh380

Rahi, J. S., & Cable, N. (2003). Severe visual impairment and blindness in children in the UK. *The Lancet*, *362*(9393), 1359–1365. doi:10.1016/S0140-6736(03)14631-4

Raj, A. K., Neuhaus, P. D., Moucheboeuf, A. M., Noorden, J. H., & Lecoutre, D. V. (2011). Mina: A Sensorimotor Robotic Orthosis for Mobility Assistance. *Journal of Robotics*, *2011*. doi:10.1155/2011/284352

Ramachandran, V. S., Stewart, M., & Rogers-Ramachandran, D. C. (1992). Perceptual correlates of massive cortical reorganization. *Neuroreport*, *3*(7), 583–586.

Renier, L., Collignon, O., Poirier, C., Tranduy, D., Vanlierde, A., Bol, A., … De Volder, A. G. (2005). Cross-modal activation of visual cortex during depth perception using auditory substitution of vision. *NeuroImage*, *26*(2), 573–580. doi:10.1016/j.neuroimage.2005.01.047

Renier, L., Laloyaux, C., Collignon, O., Tranduy, D., Vanlierde, A., Bruyer, R., & Volder, A. G. D. (2005). The Ponzo illusion with auditory substitution of vision in sighted and early-blind subjects. *Perception*, *34*(7), 857 – 867. doi:10.1068/p5219

Resnikoff, S., Pascolini, D., Etya'ale, D., Kocur, I., Pararajasegaram, R., Pokharel, G., & Mariotti, S. (2004). Global data on visual impairment in the year 2002. *Bulletin of the World Health Organization*, *82*(11), 844–851.

Resnikoff, S., Pascolini, D., Mariotti, S. P., & Pokharel, G. P. (2008). Global magnitude of visual impairment caused by uncorrected refractive errors in 2004. *Bulletin of the World Health Organization*, *86*(1), 63–70. doi:10.1590/S0042-96862008000100017

Rice, C. E., Feinstein, S. H., & Schusterman, R. J. (1965). Echo-detection ability of the blind: Size and distance factors. *Journal of Experimental Psychology*, *70*(3), 246–251. doi:10.1037/h0022215

Richter, H., Magnusson, S., Imamura, K., Fredrikson, M., Okura, M., Watanabe, Y., & Långström, B. (2002). Long-term adaptation to prism-induced inversion of the retinal images. *Experimental Brain Research*, *144*(4), 445–457. doi:10.1007/s00221-002-1097-6

Robinson, D. W., & Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, *7*(5), 166. doi:10.1088/0508-3443/7/5/302

Robinson, R., Deutsch, J., Jones, H. S., Youngson-Reilly, S., Hamlin, D. M., Dhurjon, L., & Fielder, A. R. (1994). Unrecognised and unregistered visual impairment. *British Journal of Ophthalmology*, *78*(10), 736–740. doi:10.1136/bjo.78.10.736

Rojas, M., Masip, D., Todorov, A., & Vitria, J. (2011). Automatic Prediction of Facial Trait Judgments: Appearance vs. Structural Models. *PloS One*, *6*(8), e23323.

Ruegg, C., & Cuda, M. (2012). *Math.Net Numerics*. Retrieved from http://numerics.mathdotnet.com/

Sajka, J. (2003, March). The Open Source Course: An Overview of Linux. *AccessWorld*, *4*(2). Retrieved from http://www.afb.org/afbpress/Pub.asp?DocID=aw040206

Sampaio, E., Maris, S., & Bach-y-Rita, P. (2001). Brain plasticity: “visual” acuity of blind persons via the tongue. *Brain Research*, *908*(2), 204–207. doi:doi: DOI: 10.1016/S0006-8993(01)02667-1

Santiago, J., Lupáñez, J., Pérez, E., & Funes, M. J. (2007). Time (also) flies from left to right. *Psychonomic Bulletin & Review*, *14*(3), 512–516. doi:10.3758/BF03194099

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225–239. doi:http://dx.doi.org/10.1037/h0070931

Saqib. (2009, January 17). *How can you program if you’re blind? Stackoverflow*. Retrieved April 1, 2013, from http://stackoverflow.com/revisions/453758/1

Scherer, M. J. (1996). Outcomes of assistive technology use on quality of life. *Disability and Rehabilitation*, *18*(9), 439–448.

Schwarz, M. W., Cowan, W. B., & Beatty, J. C. (1987). An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. *ACM Trans. Graph.*, *6*(2), 123–158. doi:10.1145/31336.31338

Scott, R. A. (1969). *The making of blind men*. Transaction Publishers.

Seo, H.-S., Arshamian, A., Schemmer, K., Scheer, I., Sander, T., Ritter, G., & Hummel, T. (2010). Cross-modal integration between odors and abstract symbols. *Neuroscience Letters*, *478*(3), 175–178. doi:10.1016/j.neulet.2010.05.011

Shah, S. G. S., & Robinson, I. (2007). Benefits of and barriers to involving users in medical device technology development and evaluation. *International Journal of Technology Assessment in Health Care*, *23*(1), 131–137. doi:10.1017/S0266462307051677

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147–152. doi:10.1016/S0926-6410(02)00069-1

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, *89*(4), 305–333. doi:http://dx.doi.org.ezproxy.sussex.ac.uk/10.1037/0033-295X.89.4.305

Shoval, S., Ulrich, I., & Borenstein, J. (2003). NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired]. *IEEE Robotics Automation Magazine*, *10*(1), 9–20. doi:10.1109/MRA.2003.1191706

Simpson, R. H., Quinn, M., & Ausubel, D. P. (1956). Synesthesia in Children: Association of Colors with Pure Tone Frequencies. *The Journal of Genetic Psychology*, *89*(1), 95–103. doi:10.1080/00221325.1956.10532990

Sims, K. (1991). Artificial evolution for computer graphics. *Computer Graphics*, *25*(4), 319–328.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*(1), 99–142. doi:10.1016/0010-0285(92)90004-L

Soddu, C. (2002). New Naturality: A Generative Approach to Art and Design. *Leonardo*, *35*(3), 291–294.

Song, H. J., & Beilharz, K. (2008). Aesthetic and auditory enhancements for multi-stream information sonification. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts* (pp. 224–231). New York, NY, USA: ACM. doi:10.1145/1413634.1413678

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. doi:10.3758/s13414-010-0073-7

Spence, C., & Gallace, A. (2011). Tasting shapes and words. *Food Quality and Preference*, *22*(3), 290–295. doi:10.1016/j.foodqual.2010.11.005

Steinberg, J. C. (1937). Positions of Stimulation in the Cochlea by Pure Tones. *The Journal of the Acoustical Society of America*, *8*, 176. doi:10.1121/1.1915891

Stent, A., Syrdal, A., & Mishra, T. (2011). On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (pp. 211–218). New York, NY, USA: ACM. doi:10.1145/2049536.2049574

Steven, S. S. (1957). On the psychological law. *The Psychological Review*, *64*(3), 153 – 181.

Stevens, S. S. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185. doi:10.1121/1.1915893

Stevens, S. S., & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*, *53*(3), 329–353. doi:10.2307/1417526

Stratton, G. M. (1896). Some preliminary experiments on vision without inversion of the retinal image. *Psychological Review*, *3*(6), 611–617. doi:10.1037/h0072918

Stratton, G. M. (1897). Vision without inversion of the retinal image. *Psychological Review*, *4*(4), 341–360. doi:10.1037/h0075482

Striem-Amit, E., Guendelman, M., & Amedi, A. (2012). “Visual” Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution. *PLoS ONE*, *7*(3), e33136. doi:10.1371/journal.pone.0033136

Sulzman, F. M., & Wolfe, J. W. (1991). Neurosciences research in space Future directions. *Acta Astronautica*, *23*, 289–293. doi:10.1016/0094-5765(91)90130-W

Supa, M., Cotzin, M., & Dallenbach, K. M. (1944). “Facial Vision”: The Perception of Obstacles by the Blind. *The American Journal of Psychology*, *57*(2), 133–183.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences*, *5*(5), 211–215. doi:10.1016/S1364-6613(00)01626-0

Thylefors, B. (1998). A global initiative for the elimination of avoidable blindness. *Community Eye Health*, *11*(25), 1–3.

Tkalcic, M., & Tasic, J. F. (2003). Colour spaces: perceptual, historical and applicational background. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8* (Vol. 1, pp. 304–308 vol.1). doi:10.1109/EURCON.2003.1248032

Tumulty, G., & Resler, M. M. (1984). Eye Trauma. *The American Journal of Nursing*, *84*(6), 740–744. doi:10.2307/3463716

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, *23*(4), 515–557. doi:10.1016/0010-0285(91)90005-9

Tyler, M., Danilov, Y., & Bach-y-Rita, P. (2003). Closing an open-loop control system: vestibular substitution through the tongue. *Journal of Integrative Neuroscience*, *2*(2), 159–164.

Unger, R., & Moult, J. (1993). Genetic Algorithms for Protein Folding Simulations. *Journal of Molecular Biology*, *231*(1), 75–81. doi:10.1006/jmbi.1993.1258

Veraart, C. (1989). Neurophysiological approach to the design of visual prostheses: a theoretical discussion, *13*(1-2), 57–62.

Visell, Y. (2009). Tactile sensory substitution: Models for enaction in HCI. *Interacting with Computers*, *21*(1-2), 38–53. doi:10.1016/j.intcom.2008.08.004

W Schneider, A Eschman, & A Zuccolotto. (2002). E-Prime: User's guide. Psychology Software Tools, Inc.

Wade, G. (2005, January 19). Seeing things in a different light. *BBC Local: Devon*. Retrieved from http://www.bbc.co.uk/devon/news_features/2005/eyeborg.shtml

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal Infants' Sensitivity to Synaesthetic Cross-Modality Correspondences. *Psychological Science*, *21*(1), 21 –25. doi:10.1177/0956797609354734

Walker, P., & Smith, S. (1985). Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception*, *14*(6), 729 – 736. doi:10.1068/p140729

Walker, R., Walker, D. G., Husain, M., & Kennard, C. (2000). Control of voluntary and reflexive saccades. *Experimental Brain Research*, *130*(4), 540–544. doi:10.1007/s002219900285

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-Colour Synaesthesia: to What Extent Does it Use Cross-Modal Mechanisms Common to us All? *Cortex*, *42*(2), 264–280. doi:10.1016/S0010-9452(08)70352-6

Ward, J., & Meijer, P. (2010). Visual experiences in the blind induced by an auditory sensory substitution device. *Consciousness and Cognition*, *19*(1), 492–500. doi:10.1016/j.concog.2009.10.006

Ward, J., Moore, S., Thompson-Lake, D., Salih, S., & Beck, B. (2008). The aesthetic appeal of auditory-visual synaesthetic perceptions in people without synaesthesia. *Perception*, *37*(8), 1285–1296.

Ward, J., & Wright, T. D. (2014). Sensory substitution as an artificially acquired synaesthesia. *Neuroscience & Biobehavioral Reviews*, *41*, 26–35. doi:10.1016/j.neubiorev.2012.07.007

White, B. W., Saunders, F. A., Scadden, L., Bach-y-Rita, P., & Collins, C. C. (1970). Seeing with the skin. *Perception & Psychophysics*, *7*(1), 23–27. doi:10.3758/BF03210126

World Health Organisation. (2012, June). *Visual impairment and blindness fact sheet*. *WHO media centre*. Retrieved September 8, 2013, from http://www.who.int/mediacentre/factsheets/fs282/en/

World Health Organization. (2010). Chapter VII - Diseases of the eye and adnexa. In *International Statistical Classification of Diseases and Related Health Problems* (10th ed.).

Wright, T. D. (2013). *Ployglot Framework for Sensory Substitution Devices*. *Github*. Retrieved from http://tdwright.github.io/Polyglot/

Wright, T. D., & Ward, J. (2013). The Evolution of a Visual-to-Auditory Sensory Substitution Device using Interactive Genetic Algorithms. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1620–1638. doi:10.1080/17470218.2012.754911

Wright, T. D., Ward, J., Simonon, S., & Margolis, A. (2012). Wheres Wally? Audio-visual mismatch directs ocular saccades in sensory substitution. *Seeing and Perceiving*, *25*(s1), 61–61. doi:10.1163/187847612X646820

Wulf, V., & Jarke, M. (2004). The economics of end-user development. *Communications of the ACM*, *47*(9), 41. doi:10.1145/1015864.1015886

Yao, L., & Peck, C. K. (1997). Saccadic eye movements to visual and auditory targets. *Experimental Brain Research*, *115*(1), 25–34. doi:10.1007/PL00005682

Yesilada, Y., Harper, S., Goble, C., & Stevens, R. (2004). Screen Readers Cannot See. In N. Koch, P. Fraternali, & M. Wirsing (Eds.), *Web Engineering* (pp. 445–458). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezproxy.sussex.ac.uk/chapter/10.1007/978-3-540-27834-4_55

Yip, A., & Sinha, P. (2002). Role of color in face recognition. *Journal of Vision*, *2*(7), 596–596. doi:10.1167/2.7.596

Zambarbieri, D., Schmid, R., Magenes, G., & Prablanc, C. (1982). Saccadic responses evoked by presentation of visual and auditory targets. *Experimental Brain Research*, *47*(3), 417–427. doi:10.1007/BF00239359

Zeki, S. (1999). *Inner vision: an exploration of art and the brain*. Oxford University Press.

Chapter 8: List of figures

Chapter 9: List of tables