On: 02 January 2013, At: 10:36
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

The Evolution of a Visual-to-Auditory Sensory Substitution Device using Interactive Genetic Algorithms

Thomas Wright [a] & Jamie Ward [a]

[a] School of Psychology & Sackler Centre for Consciousness Science, University of Sussex, U.K

Accepted author version posted online: 03 Dec 2012.

To cite this article: Thomas Wright & Jamie Ward (2012): The Evolution of a Visual-to-Auditory Sensory Substitution Device using Interactive Genetic Algorithms, The Quarterly Journal of Experimental Psychology, DOI:10.1080/17470218.2012.754911

To link to this article: http://dx.doi.org/10.1080/17470218.2012.754911

The Evolution of a Visual-to-Auditory Sensory Substitution Device using Interactive Genetic Algorithms

Thomas Wright[1] & Jamie Ward[1]

[1] School of Psychology & Sackler Centre for Consciousness Science, University of Sussex, U.K.

Address correspondence to :-

Thomas Wright,

School of Psychology,

University of Sussex,

Falmer, Brighton,

BN1 9QH, U.K.

Tel. : +44 (0)1273 876652

Fax. : +44 (0)1273 678058

E-mail : t.d.wright@sussex.ac.uk

Running head: "EVOLUTION OF SSDS"

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), 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.

Keywords

visual impairment, sensory substitution, genetic algorithms, aesthetics, blindness, polyglot

Article

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 comprised of a sensor (e.g. a camera), a coupling process (the software) and a stimulator (e.g. headphones, vibro-tactile array). Within a few hours of training novice participants have some ability to localise and recognise objects (Auvray, Hanneton, & O'Regan, 2007; Brown, Macpherson, & Ward, 2011) and generalise 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.

Sensory Substitution

Tactile based systems continue the tradition of Bach-y-Rita and his original Tactile Vision Sensory Substitution (TVSS) device (Bach-y-Rita, Collins, Saunders, White, & Scadden, 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, Kaczmarek, Tyler, & Garcia-Lara, 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, Hammeal, Bornstein, & Smith, 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 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 localisation 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; i.e. the image is heard piecemeal over time. The PVSA (Prosthesis Substituting Vision for Audition) uses pitch to encode position in both the horizontal and vertical axes. The PVSA 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, Persaud, & Gregory, 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 regards to the latter, one may be able to develop better conversion algorithms that are more intuitive to use because they are optimised with respect to human perceptual abilities. Whilst one can base a judgment 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 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 (4x4x4x4x2x2x8x8 or 2^{16}) 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 8-way interaction generated by the ANOVA! An alternative way of approaching this involves the use of Interactive Genetic Algorithms.

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 genomes 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 formalised 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 utilised 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 that which 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).

The Present Studies

The sensory substitution algorithms that we explore can be considered as variants of "the vOICe" (Meijer, 1992) or, rather, "the vOICe" (the capitalised 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 ask 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 must determine which image a soundscape was derived from. In the third task, the fitness function is 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 (Exp. 3), link sound to vision (at least in those with residual vision or prior visual experience; Exp. 2), and not find them aversive (Exp. 1).

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.

Genome Design

The genome for the SSD consisted of 8 different 'traits' coded by a total of 16 bits (i.e. generating 2^{16} unique genomes). The 8 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 8 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 8 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 4 levels: 250Hz, 500Hz, 750Hz or 1,000Hz.

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

v) The distribution of frequencies between the floor and ceiling was calculated in 4 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 which 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 (semi-tones) 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 1 illustrates the transfer function in each case.

<<FIGURE 1 HERE>>

Figure 1: 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 4 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 3 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:

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

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 (vi) to

$a = 1 - v$ and $a = 1 - \frac{1}{1+e^{-((v-m)-w)}}$

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 (v) this achieved by incrementing *i* either up from 0 to YRes-1 or down from YRes-1 to 0.

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 judgments 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 optimise objective performance on discriminating or identifying the soundscape? In some theories, aesthetic judgment is underpinned by the same mechanisms 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).

Methods

Participants

Twenty students (15 female, aged between 18 and 39) 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.

Materials

The stimulus material consisted of 30 natural images of everyday indoor and outdoor scenes. 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.

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 (337mm x 270mm) 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 one or sound two. 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.

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 8 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 – i.e. 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 1 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 0.25).

<<FIGURE 2 HERE>>

Figure 2. Proportion of each frequency allocation mode over 15 generations in Experiment 1 as selected by 20 participants. The trait of "Western (musical)" 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 2. In this instance, one trait is selected against (medium contrast adjustment) and another is selected for (small contrast adjustment). This is confirmed by post-hoc t-tests (small 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 3 HERE>>

Figure 3: 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 3, 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, 10kHz, 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 4.

<<FIGURE 4 HERE>>

Figure 4: 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 5 HERE>>

Figure 5: 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.

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 optimised 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 semi-tones 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 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 2500-5000Hz range (Kumar, Forster, Bailey, & Griffiths, 2008). There is some evidence consistent with this in our study – a ceiling of 2500 Hz was selected for (and a very high ceiling selected against). This has been linked to the fact that sounds in the 2,500-5,000Hz range are perceived as subjectively louder (ISO 226:2003), but may also depend on an interaction with other acoustic features (e.g. temporal modulation; Kumar et al., 2008).

Experiment 2: Audio-visual 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 (Pratt, 1930; Melara & O'Brien, 1987), loudness and luminance (Marks et al., 1987), and pitch and shape (Parise & Spence, 2009; Marks et al., 1987). Many of these are

present from a very early age suggesting that they are not learned (e.g. Walker et al., 2010). However, this literature is based either on preference measures for audio-visual 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 judgments for bound relative to unbound audio-visual stimuli (Parise & Spence, 2009). By showing that these associations are selected for in an audio-visual matching task, we aim to demonstrate that these associations may also enhance accuracy-based performance when congruently paired.

Methods

Participants

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

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.

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 (337mm x 270mm) at a comfortable viewing distance and wore headphones (Sony, MDR-XD100). The genetic algorithm used the same parameters as 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 minutes to complete.

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 generation 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 (χ^2(3, N = 200) = 29.60, p < .001). The highest frequency of 10KHz is selected against (t(19) = -7.393, p < .001) with others not differing from baseline. This is shown in Figure 5. 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 6 HERE>>

Figure 6: 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 audio-visual interactions; namely pitch-space (χ^2(1, N = 200) = 19.22, p < .001) and loudness-luminance (χ^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 6. (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 7 HERE>>

Figure 7: Proportion of pitch-space genomes (left) and luminosity-loudness genomes (right) over 10 generations in Experiment 2 as selected by 20 participants.

Discussion

In addition to selecting against a 10,000Hz 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.

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.

Participants

Twenty sighted participants (11 female, aged between 18 and 28) were recruited from the University of Sussex and 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 requires skill – it is not possible to cheat or to bias the outcome.

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.

Procedure

As in Experiment 1 they heard two soundscapes derived from the images. Participants were asked to press a button (marked "play") and 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 aspect of the genetic algorithm were as described for Experiment 2. The experiment took approximately 45 minutes to complete.

Results

Once again, regressing the mean score or 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 generation 1 and 10 were 50.4% and 59.5%. Four traits showed evidence of selection when assessed in the final generation.

Figure 7 shows that there is a distinct advantage conferred by the musical types of frequency allocation (χ^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 logaritghmic 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 audio-visual matching even though the auditory soundscape was task-relevant in all three experiments.

<<FIGURE 8 HERE>>

Figure 8: 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 8 shows that the frequency range floor (i.e. the lowest frequency in a soundscape) of 750Hz is strongly selected for in the final generation (χ^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 9 HERE>>

Figure 9: Proportion of genomes containing a given frequency range floor (in Hz) over 10 generations in Experiment 3 as selected by 20 participants. Note that 750Hz 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 9 illustrates this collapsing the 8 resolutions into 4 bins. When looking at this 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 10 HERE>>

Figure 10: 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 10. 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 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 11 HERE>>

Figure 11: 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.

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 20Hz). However, other findings are unexpected. Firstly, 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 (R. L. 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.

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.

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. (2011) explored different settings within the "vOICe" device; for instance, comparing contrast settings (bright=loud v. bright=quiet) and the length of the soundscape (1s v. 2s) in a 2x2 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 are 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 summarised in Table 1.

<<TABLE 1 HERE>>

Table 1: Summary of the results from all three experiments. +/- denotes traits that are selected for or against respectively.

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 which operates in the lower frequency range (up to 2,500Hz) 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. Resolving whether how to represent 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 extend the research to the blind and visually impaired. Although visually impaired people tend to perform better (Arno 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; i.e. 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 (O'Regan, 1992; Brown et al., 2011). 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 is 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 (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, Hammeal, Bornstein, & Smith, 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.

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 itself, 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 and 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) is 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 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/judgments of participants, offers a significant advance in this field.

References

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

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., 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

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

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.

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.

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

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

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

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

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

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.

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

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

International Organization for Standardization (2003). Normal equal-loudness-level contours. ISO 226:2003 Acoustics. Geneva, Switzerland.

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.

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

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

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

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

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

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

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

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

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

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

Ortiz, T., Poch, J., Santos, J. M., Requena, C., Martínez, A. M., Ortiz-Terán, L., Turrero, A., et al. (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

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

Pollok, B., Schnitzler, I., Stoerig, P., Mierdorf, T., & Schnitzler, A. (2005). Image-to-sound conversion: experience-induced plasticity in auditory cortex of blindfolded adults. Experimental Brain Research, 167(2), 287–291. doi:10.1007/s00221-005-0060-8

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

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/p5219Rojas, M., Masip, D., Todorov, A., & Vitria, J. (2011). Automatic Prediction of Facial Trait Judgments: Appearance vs. Structural Models. *PloS one*, *6*(8), e23323.

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

Smith, J. R. (1991). Designing biomorphs with an interactive genetic algorithm (pp. 535-538). Presented at the Proceedings of the Fourth International Conference on Genetic Algorithms, Morgan Kaufmann Publishers.

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. *Proceedings of the 3rd international conference on Digital*

Interactive Media in Entertainment and Arts, DIMEA '08 (pp. 224–231). New York, NY, USA: ACM. doi:10.1145/1413634.1413678

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

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

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

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

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.

Zeki, S. (1999). *Inner vision: an exploration of art and the brain*. Oxford University Press.

End

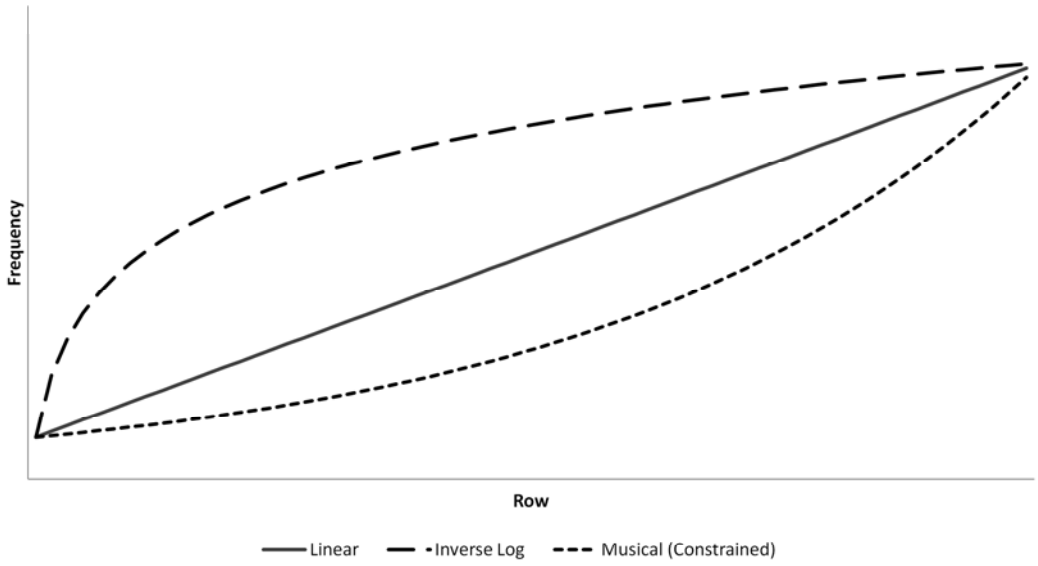

Figure 3: Comparison of frequency allocation modes: Linear, inverse log, musical (constrained)

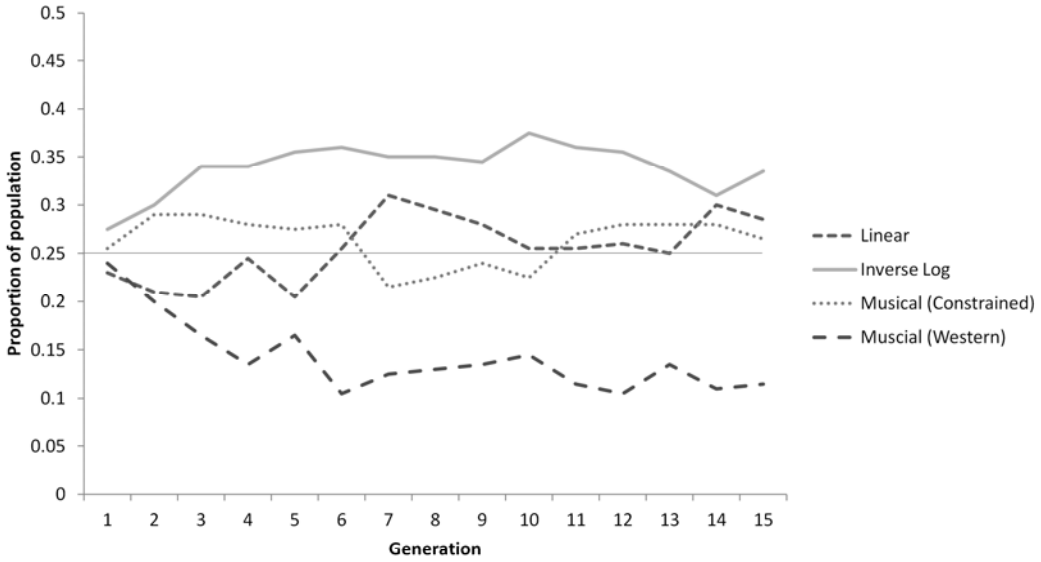

Figure 4. Proportion of each frequency allocation mode over 15 generations in Experiment 1 as selected by 20 participants. The trait of "Western (musical)" is selected against.

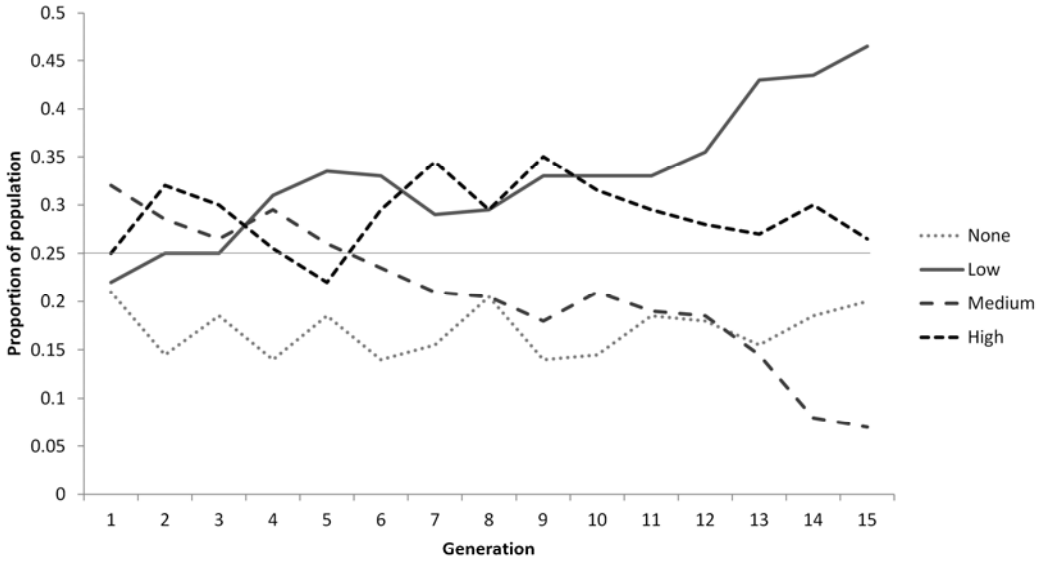

Figure 3: Proportion of each contrast enhancement mode over 15 generations in Experiment 1 as selected by 20 participants. Whereas a small contrast enhancement is selected for, a medium contrast enhancement is selected against.

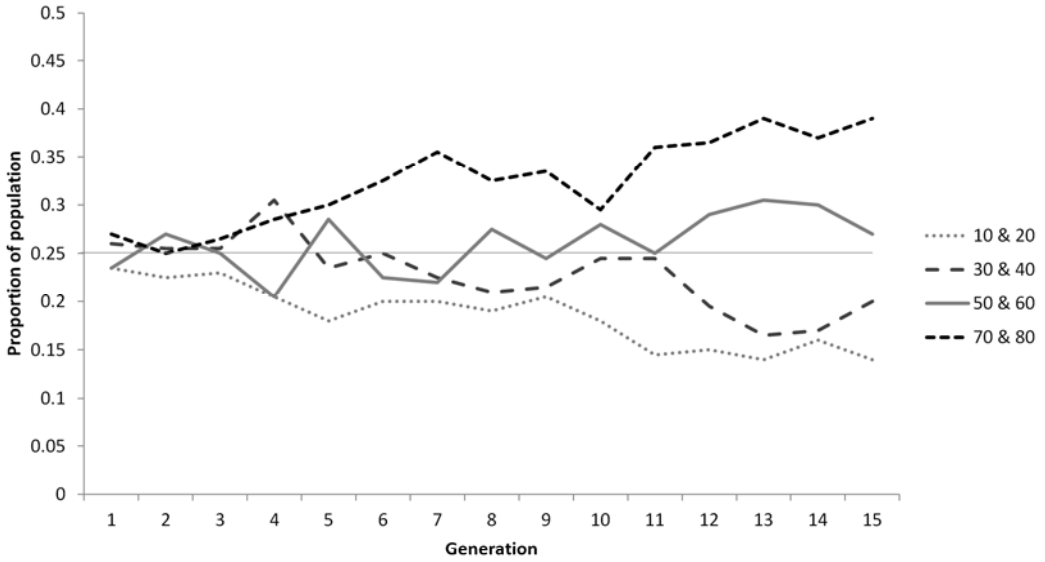

Figure 4: Proportion of genomes containing a given Y-resolution (number of discrete frequencies) over 15 generations in Experiment 1 as selected by 20 participants. There is a monotonic relationship between resolution and prevalence in the final generation.

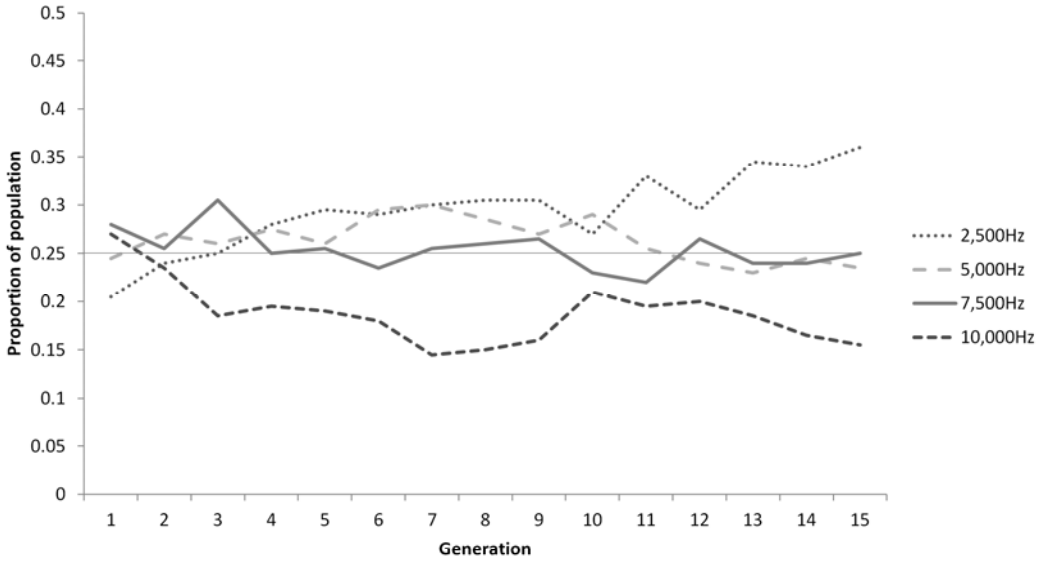

Figure 5: Proportion of frequency range ceilings over 15 generations in Experiment 1 as selected by 20 participants. Note that 2500Hz is selected for and 10,000Hz is selected against.

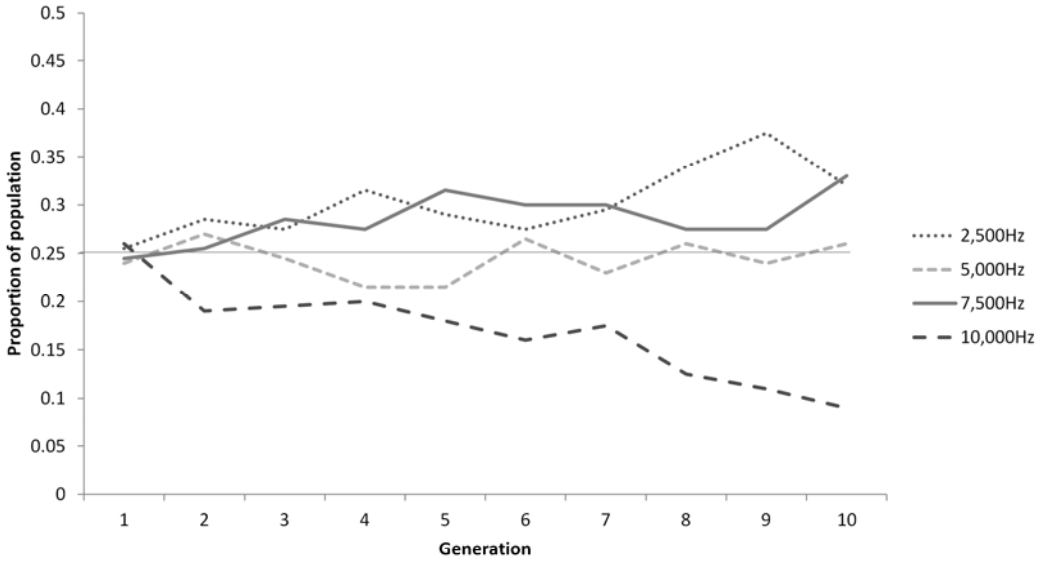

Figure 6: Proportion of frequency range ceilings (in Hz) over 10 generations in Experiment 2 as selected by 20 participants. Note that 10,000Hz is selected against.

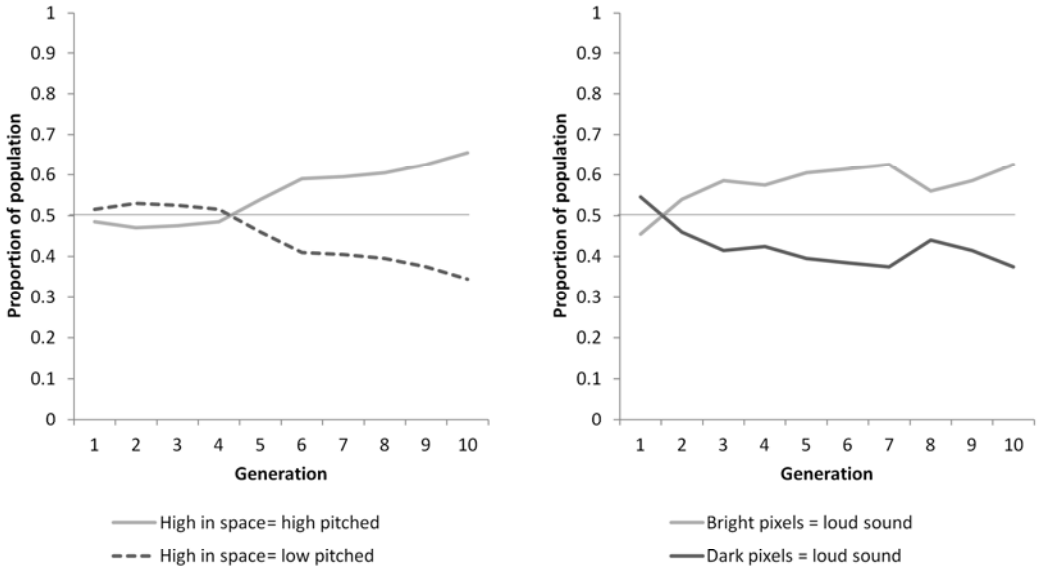

Figure 7: Proportion of pitch-space genomes (left) and luminosity-loudness genomes (right) over 10 generations in Experiment 2 as selected by 20 participants.

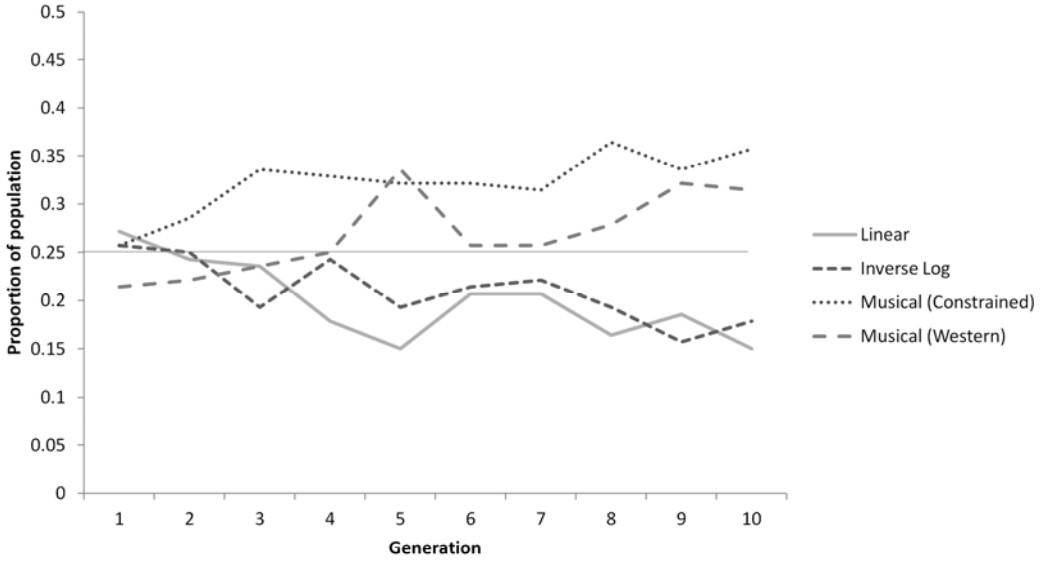

Figure 8: Proportion of genomes containing a given frequency allocation modes over 10 generations in Experiment 3 as selected by 20 participants. Note that musical (i.e. logarithmic) distributions of discrete frequencies are selected for.

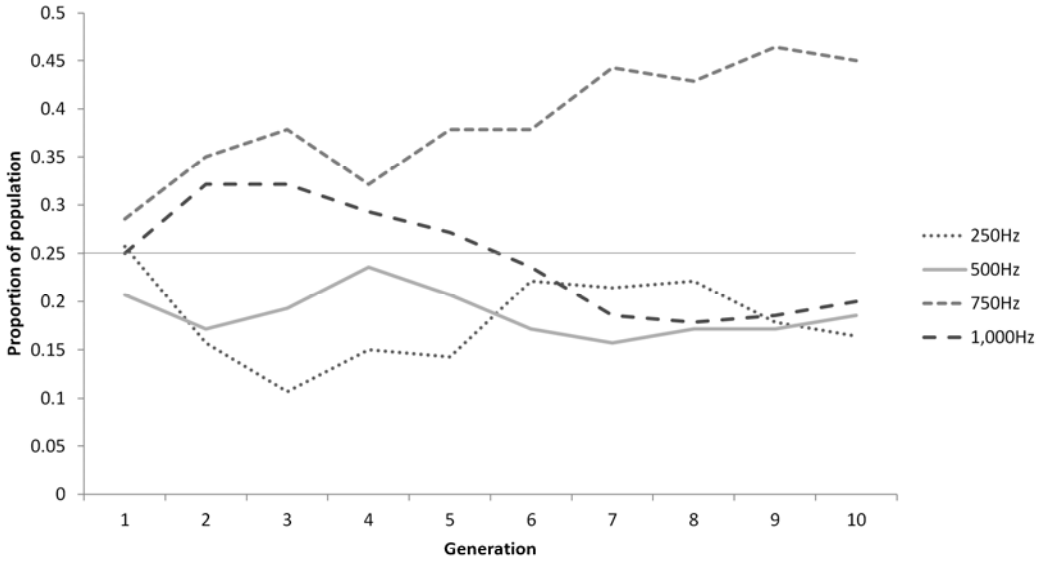

Figure 9: Proportion of genomes containing a given frequency range floor (in Hz) over 10 generations in Experiment 3 as selected by 20 participants. Note that 750Hz is selected for.

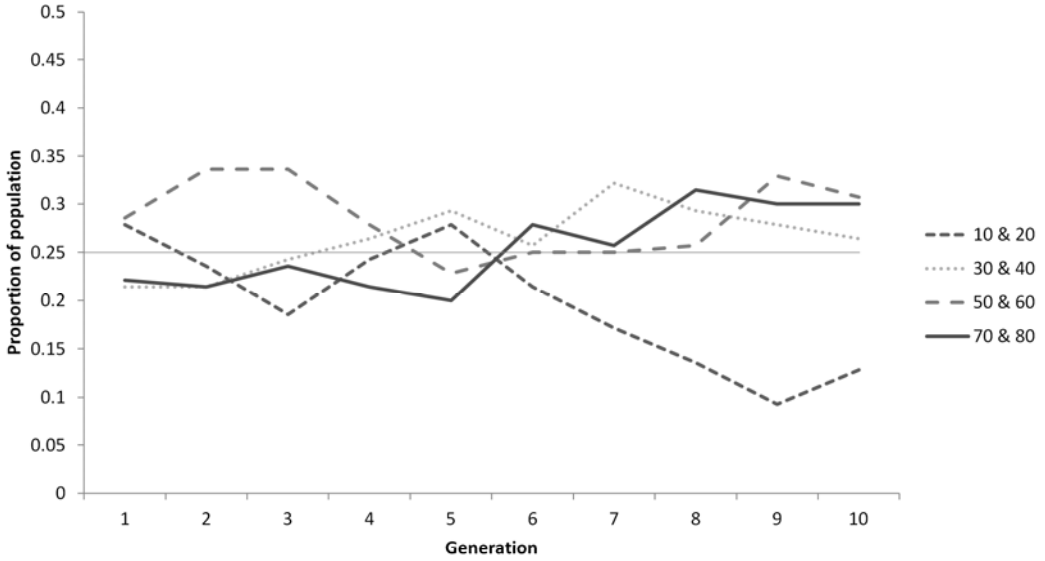

Figure 10: Proportion of genomes containing a given X-resolution (number of separate time points) over 10 generations in Experiment 3 as selected by 20 participants.

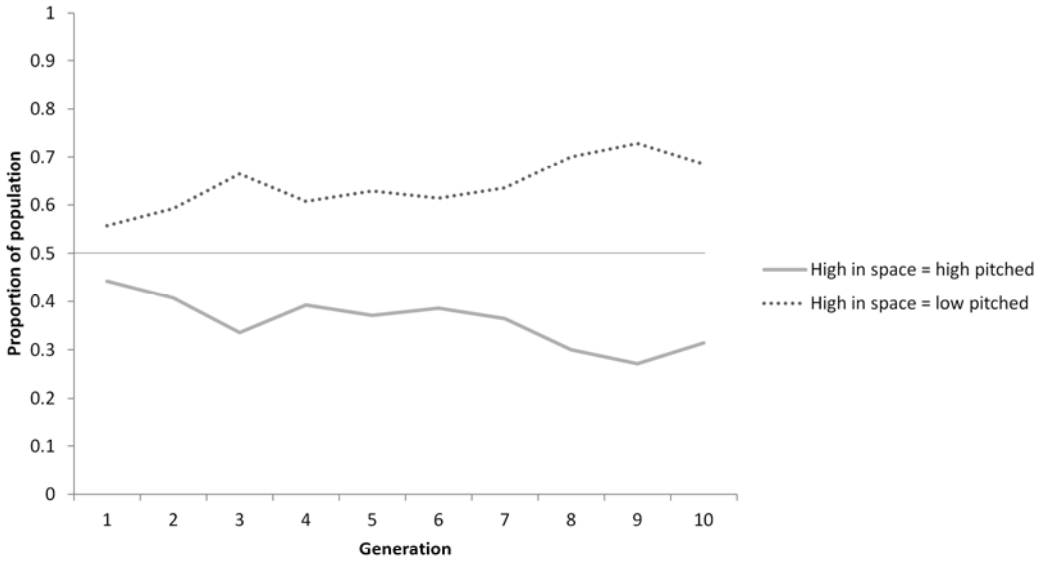

Figure 11: Proportion of genomes containing pitch-space inversions over 10 generations in Experiment 3 as selected by 20 participants. Note that high space = low frequency is selected for.

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

Table 2: Summary of the results from all three experiments. +/- denotes traits that are selected for or against respectively.