The Creation, Presentation and Implications of Selected Auditory Illusions

Scott Flinn
flinn@cs.ubc.ca

Kellogg S. Booth
ksbooth@cs.ubc.ca

Department of Computer Science
The University of British Columbia

Abstract

This report describes the initial phase of a project whose goal is to produce a rich acoustic environment in which the behaviour of multiple independent activities is communicated through perceptually distinguishable auditory streams. While much is known about the perception of isolated auditory phenomena, there are few general guidelines for the selection of auditory elements that can be composed to achieve a display that is effective in situations where the ambient acoustic conditions are uncontrolled. Several auditory illusions and effects are described in the areas of relative pitch discrimination, perception of auditory streams, and the natural association of visual and auditory stimuli. The effects have been evaluated informally through a set of demonstration programs that have been presented to a large and varied audience. Each auditory effect is introduced, suggestions for an effective demonstration are given, and our experience with the demonstration program is summarized. Implementation issues relevant to the reproduction of these effects on other platforms are also discussed. We conclude by describing several experiments aimed at resolving issues raised by our experience with these effects.

Keywords

Auditory illusions; pitch discrimination; auditory scene analysis; auditory streaming; visual capture; auditory displays; auditory environments.

Contents

List of Figures

1 Introduction

The work described in this report represents the first phase of a project whose ultimate objective is to provide a workstation with an ambient acoustic environment. Ideally the auditory display will convey useful information regarding the state of the workstation, the devices and services on which it depends, and the behaviour of ongoing computations.

Eventually specific auditory elements and techniques must be selected as the components of a complex display. These are likely to include auditory icons in a number of possible varieties [18, 2, 20]. The general idea of generating informative sounds in response to user actions, demonstrated effectively by Gaver's Sonic Finder [19], will be reflected in the design.

However, a complete solution will also include elements that are independent of user interaction, intended to convey information regarding the behaviour of autonomous or externally controlled processes. Gaver's ARKola simulation [21], which conveys information concerning the interactions and individual states of a set of bottling machines by simulating the sounds generated by their operation, is perhaps the most literal demonstration of the utility of a rich acoustic environment. However the possibilities range from the complex internal workings of a single object, as in the program auralization of Digiano and Baecker [10], to the coordination of logically independent and heterogeneous streams in a unified auditory display, as exemplified by the auditory window system of Ludwig et al. [24]. Ultimately the richness and complexity of familiar environmental sounds [1, 28, 37] should be exploited, with an emphasis on those that are both pleasant and easily distinguished [32].

1.1 Design of Auditory Display Elements

Choosing suitable auditory elements remains a considerable challenge. The difficulty lies primarily in the coding of information. A mapping must be designed that links the semantics of the domain of information to be communicated to the characteristics of the sound used to represent it. Introductory experiments with simple mappings [30, 39] suggested that the obvious approach of mapping each dimension of a data set to a distinct characteristic of a tone, such as its pitch, loudness, attack or decay, had the potential to produce effective auditory renderings of complex, high dimensional data sets. Bly systematically pursued this idea with a set of careful experiments whose results were again promising [5, 4]. Encouraged by this success, others explored variations of the approach. For example, Bly continued to pursue the problem with a number of collaborations [3, 7]. Rabenhorst et al. used a similar approach to complement the visualization of scalar fields in a semiconductor [31], and other similar experiments were conducted.

However it soon became apparent that while simple mappings produced auditory displays in which patterns and other distinct artifacts could be easily perceived, the cognitive task of reversing the mapping (extracting meaning from the display based on an understanding of the forward mapping) remained extremely difficult. For this reason, emphasis began to shift toward sophisticated mappings between the underlying data and more familiar properties of sounds and musical presentation. For example, Evans created a visual presentation and a complementary *sonic map* (or "sonic score") from the same set of data [12, 13]. While intended to reinforce the visual presentation, the sonic map had the simultaneous objective of achieving the aesthetic goals of traditional musical accompaniment. Lunney and Morrison explored the transformation of instrumental measurements into musical patterns to aid blind scientists and science students [25]. The sounds Gaver created for the Sonic Finder were based on the observation that extremely complex sounds, such as a car door closing or a jar being filled with water, can be recognized and distinguished effortlessly if they are sufficiently familiar and can be associated with the physical actions that produce them.

Publications in this area generally indicate that their authors are aware of the enormous body of psychology literature devoted to auditory perception. Nevertheless very few projects have directly incorporated results from psychophysics or perceptual cognition. Codification of information remains ad hoc (even if based on a strong understanding of the relevant psychological principles) or domain specific. A good representative of the latter category is the acoustic operating room monitor of Fitch [15]. In this case an effective system has been designed by choosing a mapping that is very appropriate to the problem, but the solution contributes little to a general procedure for identifying suitable mappings.

The essential difficulty is that results from the psychology literature are rarely applicable in practice because the psychologists have carefully eliminated confounding factors from their experiments that cannot be eliminated from the environments in which interactive software is intended to operate. Frysinger has adapted and refined some of the experimental techniques of traditional psychoacoustics with the objective of producing more directly applicable results [17]. This is a step in the right direction, but a more general set of guidelines for the effective coding of information in a typical workstation environment would be of tremendous benefit to the field.

To achieve the goals set out earlier in the section, auditory elements must be chosen that, when composed in a complex acoustic display, are capable of delivering a variety of types of information from sources that are naturally perceived as logically distinct by persons having little musical or other relevant training. In the absence of a mature set of guidelines for selecting these elements other grounds for the decision are needed, as a better initial choice will increase the chances of success in subsequent formal evaluations. The auditory effects described in this report were implemented primarily to address that need, although a number of additional benefits were also anticipated. First, it is often difficult to acquire a complete understanding of an auditory phenomenon simply by reading about it. The implementation effort provided both practical experience in working with various effects and a deeper appreciation of the subtle interactions between auditory phenomena. Second, the implementations facilitated informal evaluation of a number of individual effects, as well as many of their interactions, in a typical workstation environment and with a large and diverse audience. Third, the development process provided a good measure of the ability of the workstation to generate auditory displays with sufficient precision, complexity and flexibility.

1.2 Overview

The auditory illusions and effects included in this report can be divided into three distinct categories: pitch discrimination, auditory streaming, and visual capture. The next section defines these terms and discusses a number of specific phenomena chosen from each category. It has been organized as a demonstration manual, providing for each effect or illusion both a general description of the phenomenon and a guide to using our specific implementation to present the effect to a general audience. The guide is written to enable a presenter to anticipate a wide range of reactions and to build on them in a constructive way. However, this organization also serves as a concise summary of the things we have learned concerning the way in which the various phenomena are perceived by a wide range of people under equally diverse conditions.

While Section 2 deals with specific implementations of each effect, the discussion is easily generalized to a similar implementation on any suitable platform. Section 3 discusses implementation details that are specific to the NeXTStep platform we have used. Both strengths and weaknesses of the platform are considered, and a profile of the minimal system requirements for effectively generating auditory displays of this type is suggested. Section 4 concludes with a discussion of the implications of these investigations and of the continuing work that they suggest.

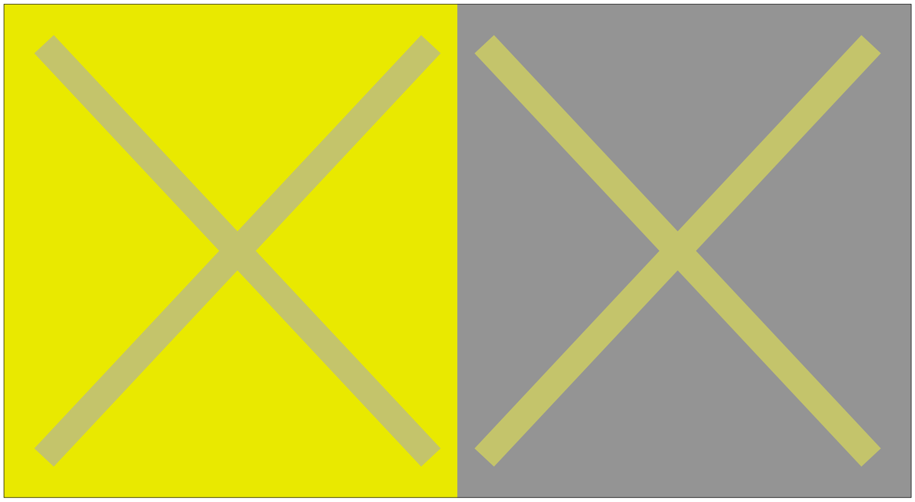


Figure 1: Yellow-Gray Contrast Illusion

2 Illusions and Effects

2.1 Visual Illusions

There is not yet a widely accepted way of including auditory material in an electronic document, and there is essentially no way at all of doing so in a printed one. It is therefore difficult to convey many of the subtleties of auditory effects and illusions through text and graphic illustration. We begin this section with two visual illusions that can be presented more effectively. While these illusions are not meant to be analogous to any particular auditory effect, they share the same objective of illuminating specific characteristics of human perception and their inclusion will establish the context in which the auditory effects should be considered.

2.1.1 Yellow-Gray Contrast

Introduction

Sun et al. have studied the presentation of a number of visual illusions using the X Windows system [36]. Figure 1 illustrates one of the colour contrast illusions they describe. The figure shows two large X shapes, one on a gray background and the other on a yellow background. The two Xes are drawn using the same colour: an equal mixture of the yellow and gray background colours. However, most people perceive the two Xes as having different colours; the X on the yellow background should appear slightly gray, and the X on the gray background should appear slightly yellow. The effectiveness of the illusion will depend on the fidelity with which this document has been rendered from its colour PostScript source. A high quality reproduction will appear to refute the claim that the Xes are the same colour,

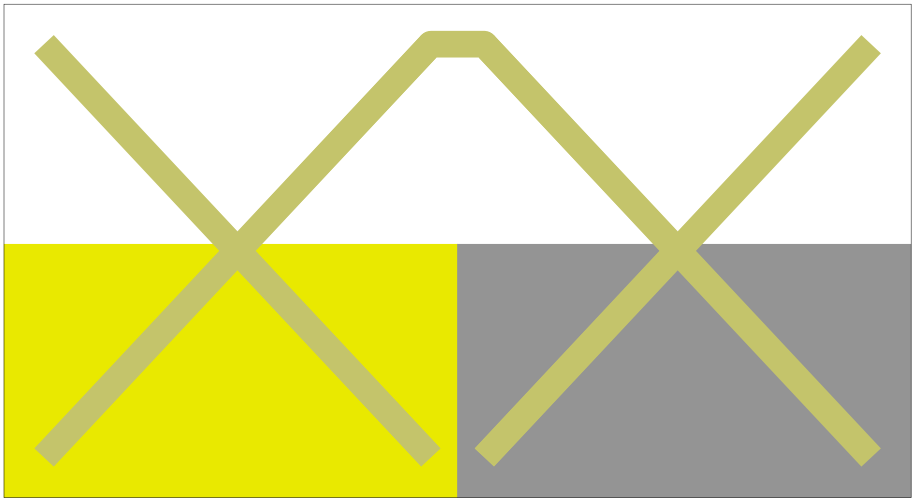


Figure 2: Partial Yellow-Gray Contrast Illusion

even upon close inspection. Figure 2 defends the claim by showing half of each X on the same white background. The illusion can be alternately created and destroyed by covering the top and bottom half of this figure respectively.

The light emitted or reflected from the two Xes is essentially identical in its spectral distribution. The illusion is created by the visual system after light from each X reaches the retina, and is dependent on the spectral content of the light reaching adjacent areas of the retina. When one looks at the yellow-gray X against the yellow background, the effect is the same as if the same X is viewed under conditions in which the ambient lighting is predominantly yellow. In such an environment, one would expect objects illuminated by the yellow ambient light to appear more yellow than when viewed under white light (since other colours are not available to be reflected). The visual system attempts to correct the appearance of objects in the environment, based on its perception of the ambient lighting conditions, so that they will appear to have roughly the same colour in a wide variety of lighting conditions. In the case of the yellow-gray X on a yellow background, therefore, the visual system effectively removes the yellow from the X in response to its detection of the yellow background, creating a perception that is distinctly gray.

One observation suggested by this illusion will be especially relevant to the later discussion of auditory illusions. In particular, statements regarding the relative colour of the two Xes must be made carefully. They are the same colour in the sense that the light they emit or reflect has the same spectral content. On the other hand, we know the names of colours only because we have learned to associate particular names with unique perceptual experiences. For example it is possible for two sources of light having very different spectral distributions to be perceived as the same colour (the colours produced by such spectra are called *metamers*). They are considered to be the same colour in spite of their differing spectra because they give rise to the same perceptual experience. In the same way, the perceptual experiences produced by the two Xes in the yellow-gray contrast illusion are the same as those with which we associate the names yellow and gray. In this sense, they are certainly not the same colour.

The illusion demonstrates that we do not always see what we are shown: we are shown two Xes of the same colour, but we see two Xes of different colours. The designer of a display intended for human consumption must therefore be concerned not only with what the display contains but with how it is perceived. This requires an understanding of the perception of both individual phenomena and their interactions.

Presentation

The interactive implementation of this illusion involves a window whose contents are identical to Figure 1 and a second window containing four simple controls. Two of the controls adjust the intensities of the yellow and gray backgrounds. As these values are modified the two X figures are redrawn using an equal mixture of the new background colours. An effective way to begin the demonstration is to set the yellow background to about 75% of maximum intensity, and the gray background to about 25%. The values producing the most effective illusion depend somewhat on the ambient lighting conditions; the demonstrator should select these values, then make small adjustments to achieve the strongest illusion.

Once the background intensities have been set, the audience should be asked for an opinion on the colours of the two Xes. It is common for viewers to attempt to second guess the demonstrator's intentions, or to be dishonest in the hopes of avoiding embarrassment (many will admit this as they come to believe that it is not the demonstrator's objective to make fools of them). Others will remain adamant that the two Xes really do *appear* to be the same colour. By far the majority of viewers, however, will concede that they perceive the colours of the two Xes as being different. Members of larger groups will often argue amongst themselves as to the verity of the situation. For those who believe the Xes to be of different colours, the remaining two interactive controls can be used to further confuse the issue. One control allows the thickness of the X figures to be adjusted, providing more or fewer pixels to examine. The final control is a toggle button that inserts an additional segment joining the Xes at the top (as in Figure 2). Portions of the foregoing explanation and discussion can then be delivered verbally with a level of detail suitable to the audience. It has been found that the full explanation is appreciated by most who have reached or passed the high school level.

2.1.2 The Café Wall Illusion

Introduction

Figure 3 shows the *café wall* illusion. It was first studied by Gregory [22] who was prompted to investigate the phenomenon when a member of his laboratory observed rows of alternating white and black ceramic tiles, each row staggered horizontally by the width of half a tile

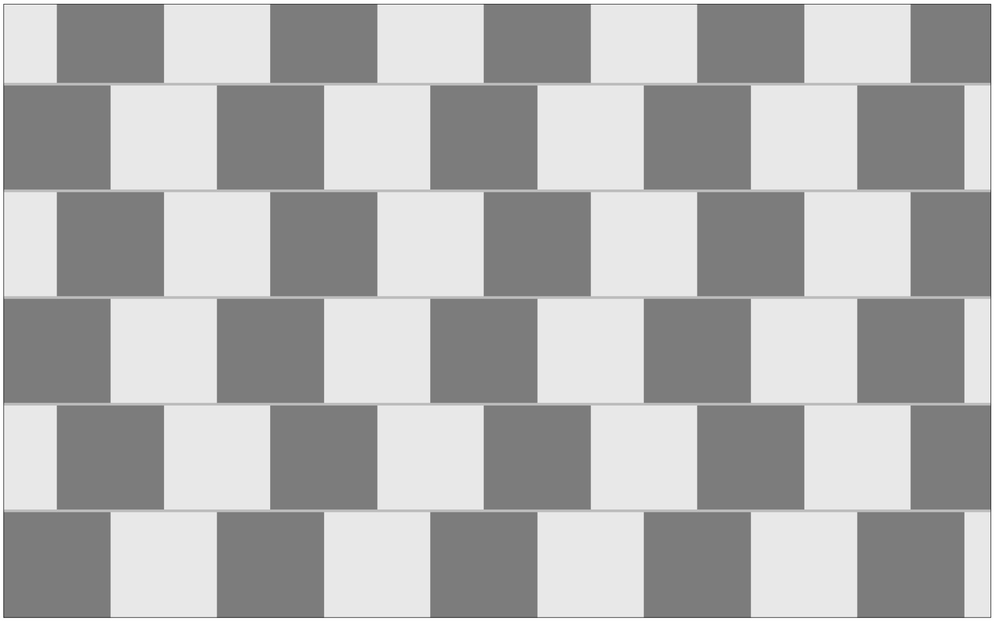


Figure 3: The Café Wall Illusion

and with rows separated by horizontal lines of mortar, on the wall of a nearby café. This arrangement gives rise to the perception of lines of mortar that are neither horizontal nor parallel and of rows of tiles that are taller at one end than at the other. The lines are in fact both horizontal and parallel, and all tiles are square and of the same size.

The effect has a straight forward explanation. The mortar boundary between two dark squares is visually apparent, as is the boundary between two light squares. However, normal visual acuity is not sufficiently sharp to resolve the mortar between light and dark squares when viewed from a typical distance. Although the mortar in these regions is not perceived as a separate object, the area it occupies must still be interpreted in some fashion by the visual system. The result is that where a dark block borders light blocks, the lines of mortar above and below are perceived as being part of that portion of the block, making it appear thicker on that side than on the other. Exactly the same reasoning applies to light blocks at their boundaries with dark ones.

This effect alone should give the appearance of a succession of wedge shaped blocks separated by a jagged or sawtooth boundary. However, the higher cognitive levels of the visual system tend to select the simplest geometric models that are consistent with visual stimulus. In the case of the lines of mortar, a straight line model is sufficiently close to be accepted (recall that the lines are in fact both straight and parallel). When faced with the contradictory evidence of wedge shaped blocks separated by straight lines, the visual system reaches a compromise that involves rows of blocks whose horizontal boundaries are straight, but not parallel. These phenomena have been studied in detail, and much more is known about the way in which the compromise develops [26, 11], but this simplified view will be sufficient for our purposes.

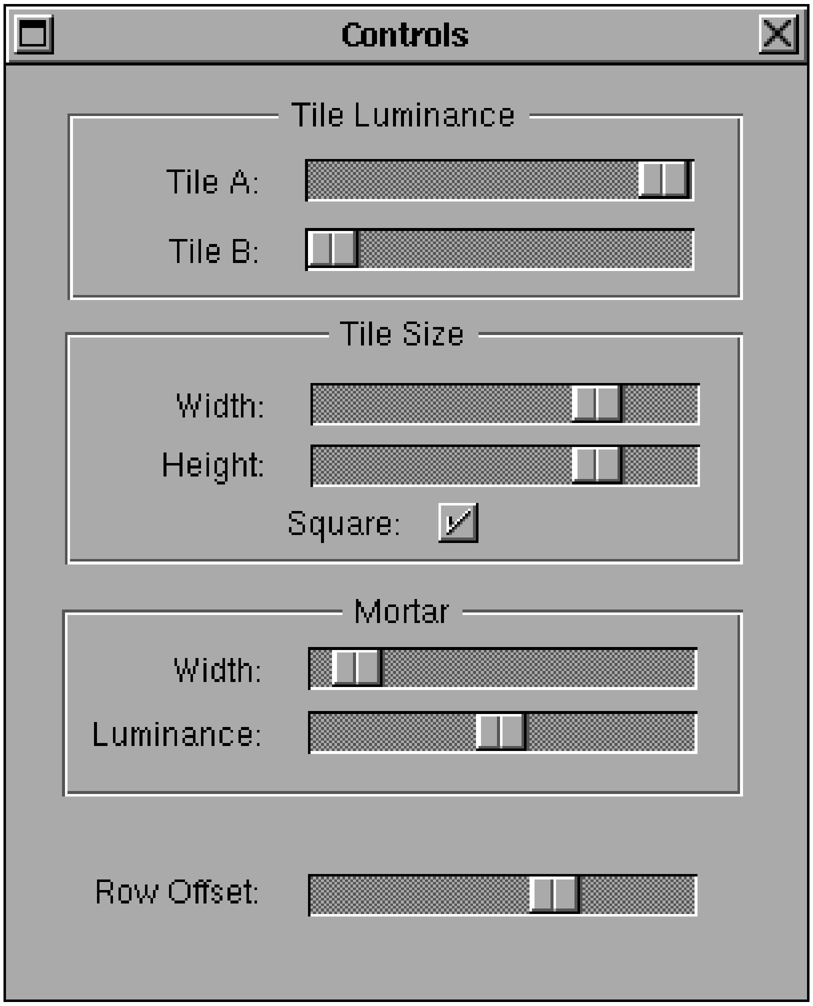


Figure 4: Controls for the Café Wall Illusion

Presentation

The interactive implementation of this illusion can be used to give a compelling demonstration. As for the yellow-gray contrast illusion, a single window is used to present the café wall display. A separate control panel, shown in Figure 4, is used to demonstrate different aspects of the explanation given above.

First, there are a number of ways to demonstrate the robustness of the illusion. The display window size can be adjusted using the standard window sizing control. A small window showing between two and three rows of tiles can be used to demonstrate that the illusion is strong even when there are few tiles to create it. Two slider controls can be used to adjust the width and height of the tiles. By default the tiles are constrained to be square (adjusting one slider causes the other to be adjusted to match); the switch labelled *Square* toggles this constraint. Displays with large, small, tall and narrow, or short and wide tiles all produce similar, though not identical effects.

The slider control labelled *Row Offset* is used to adjust the horizontal offset of the even numbered rows with respect to the odd numbered ones (numbering from the bottom and starting at one). It can be used to produce a checker board pattern in which square tiles appear to be separated by parallel, horizontal mortar lines. Similarly, the configuration in which like coloured tiles are aligned vertically produces no illusion. Most audiences are interested in seeing that the second row from the bottom, which originally appears narrower at the right end than at the left, reverses in this regard when the row offset is moved left by one full tile width.

The luminance and width of the mortar lines can each be manipulated using a slider control. Removing the mortar lines (adjusting to zero width) destroys the illusion immediately, while increasing their thickness slowly weakens it. Viewers of a display with thick mortar lines typically report seeing right angles and parallel edges, although many also report apparent variations in the mortar luminance or a three dimensional "woven" effect. Since the illusion depends on mortar lines that are too narrow to be separately resolved, the illusion tends to weaken more rapidly for those viewing from a closer distance.

The two remaining slider controls affect the luminance of the light and dark tiles. Most viewers report that the illusion is slightly strengthened when the black tiles are made somewhat lighter and the white ones somewhat darker, as long as the mortar luminance remains intermediate between them. Beginning with this arrangement, the demonstrator can slowly adjust the mortar luminance control to darken the mortar lines. The illusion fades as the mortar luminance approaches that of the darker tiles, disappearing completely as the mortar reaches black. In the other direction, the illusion slowly grows as the mortar luminance approaches an intermediate value, then diminishes again as the mortar approaches white.

It is usually more effective to carry out these manipulations before providing an explanation of the effect, revisiting some of the procedures as required to emphasize parts of the explanation that may remain unclear. For example, the mortar width can again be adjusted while viewers pay close attention to the boundary between a dark and light block. Skepticism is often expressed concerning the compromise between straight mortar lines and wedge shaped blocks that leads to the appearance of straight but converging lines. When a sufficiently small block size is chosen (a third of the original size, say), the display presents too much information contradicting the straight line hypothesis for the illusion to be maintained. Viewers may again report a three dimensional "woven" effect, but it is impossible to focus on any adjacent pair of mortar lines and perceive them as straight but convergent.

2.1.3 Discussion

The number of pixels used to display the mortar in this illusion is a tiny fraction of the total number in the scene. Nonetheless, the illusion can be convincingly created and destroyed by simply altering the luminance of the mortar gray within a modest range. In other words, the perception of the figure's geometry is completely dependent on the relative luminance of a small number of its pixels. Like the yellow-gray contrast illusion, the café wall demonstrates clearly that the perception of a stimulus may be very different from the stimulus itself: lines that appear neither horizontal nor parallel are in fact both.

The implementations of these two visual illusions also tell us something important about the platform on which they are implemented. The phenomena that create visual illusions often occur at the limits of human perception (detecting the mortar between light and dark blocks in the café wall illusion, for example). An effective implementation of the illusion on a workstation therefore demonstrates that the workstation display is capable of operating at those limits. This issue will be raised again in conjunction with several of the auditory effects described below.

2.2 Pitch Discrimination

An auditory display that conveys information must encode that information in a suitable way. Pitch is often suggested as an obvious candidate: a high pitch would mean one thing and a low pitch another. To say that one pitch is higher than another is to suggest that pitch can be represented as a one dimensional scale along which tones of different pitch can be uniquely ordered (i.e., for any pair of tones, one can be identified as being higher in pitch than the other). However pitch is more properly regarded as multi-dimensional [35], requiring that greater care be taken when using pitch to encode information. This section (2.2) describes two related auditory illusions that illustrate the potential difficulty of relative pitch discrimination.

2.2.1 Shepard's Tones

Introduction

The artist M. C. Escher has produced many drawings of perspective paradoxes and impossible figures [14]. One of his best known illustrations, *Ascending and Descending*, features a set of stairs having four flights that are connected in a cycle. Drone like people are depicted endlessly ascending or endlessly descending. Figure 5 illustrates the cyclic arrangement. The illusion is created by bending the rules of perspective projection in a figure that apparently follows them. The stair case appears to be the two dimensional projection of a three dimensional object, but it is impossible to construct such an object in three dimensions.

In 1964, Roger Shepard constructed an auditory effect that is analogous to the eternally ascending staircase [34]. The effect is composed of twelve carefully constructed tones, which have come to be known as *Shepard's Tones*, played in a cycle (1 2 3 ... 11 12 1 2 3 ...). As the cycle plays, each tone appears to be higher in pitch than the last. In particular, the first tone appears to be higher in pitch than the twelfth that immediately preceded it. Since relative pitch is normally considered to be transitive (if pitch C is higher than pitch B, which in turn is higher than pitch A, then pitch C is higher than pitch A), it follows that every tone appears higher in pitch than all those that preceded it, and lower in pitch than all those that follow. However, since they are played in a cycle, each tone eventually both precedes and follows every other. In other words, each tone appears to be both higher than

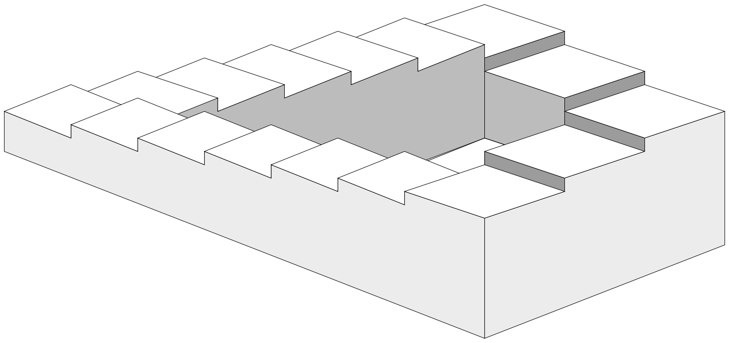


Figure 5: The Eternally Ascending/Descending Staircase

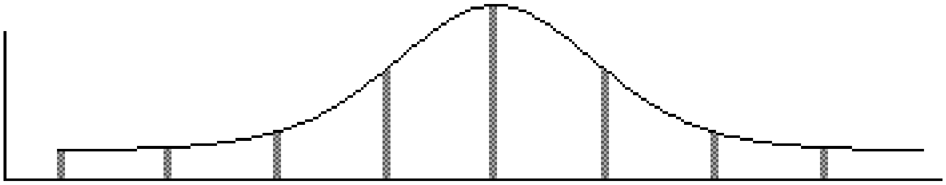


Figure 6: Composition of First Shepard's Tone

every other (when it follows them) and lower than every other (when it precedes them). This contradiction suggests that the assumption of transitivity should be rejected.

The illusion depends on a careful distribution of harmonics (or partials) in each tone. Figure 6 illustrates the composition of the first tone in the sequence. Each vertical bar in the figure represents a separate frequency component; all eight components are combined to produce a single tone with a shrill timbre. Frequency increases logarithmically along the horizontal axis (a linear scale for pitch), and the vertical axis measures amplitude. The components are each separated by exactly one octave (a factor of two in frequency), so the components of the composite tone span a range of seven octaves. The second tone in the sequence is obtained from the first by shifting each component to the right by one semitone (a factor of $\sqrt[12]{2}$ in frequency) and adjusting the amplitude of each according to the stationary amplitude envelope represented in the figure by a smooth curve. The remaining tones in the sequence are produced in the same way.

Since the components are separated by octaves, the pitch of the composite tone is easily named. If the base frequency is 110.0 Hz, a pitch the western musical scale gives the name A, then the pitch of every component is A. The pitch of the shrill composite tone is then surely A. When every component is shifted a semitone higher, the pitch of the resulting tone is $B^\flat$, and so on. It is therefore not an illusion that the pitch appears to increase. The illusion arises instead from the way in which the component amplitudes change. As the components march towards the right, the fixed amplitude envelope causes the higher frequency components to diminish in their contribution and the lower components to become more dominant. In this way the average spectral content remains roughly constant as the pitch clearly increases with each step. If a thirteenth tone were to be generated (with every component shifted twelve semitones, or an octave, to the right), it would look identical to the first tone with two exceptions: the lowest component of the first tone would be absent in the thirteenth, which in turn would have a component an octave higher than the highest in the first. However, since these components are at the tails of the amplitude envelope, they are both extreme in frequency and low in amplitude. They are therefore below the *psychoacoustical threshold* of hearing: a pitch that quiet and high (or low) is simply inaudible. The transition from the twelfth tone to the first at the moment the sequence begins to repeat is therefore very difficult to detect. This, in combination with the roughly constant spectral content of each tone, generates the illusion.

Presentation

This illusion, which has been reproduced on a NeXT workstation using a digital signal processor (DSP) for tone synthesis, has proven to be remarkably robust under a wide range of ambient auditory conditions. We have presented it successfully to individuals in an otherwise empty laboratory and to relatively large groups in a lab filled with enthusiastic high school students.

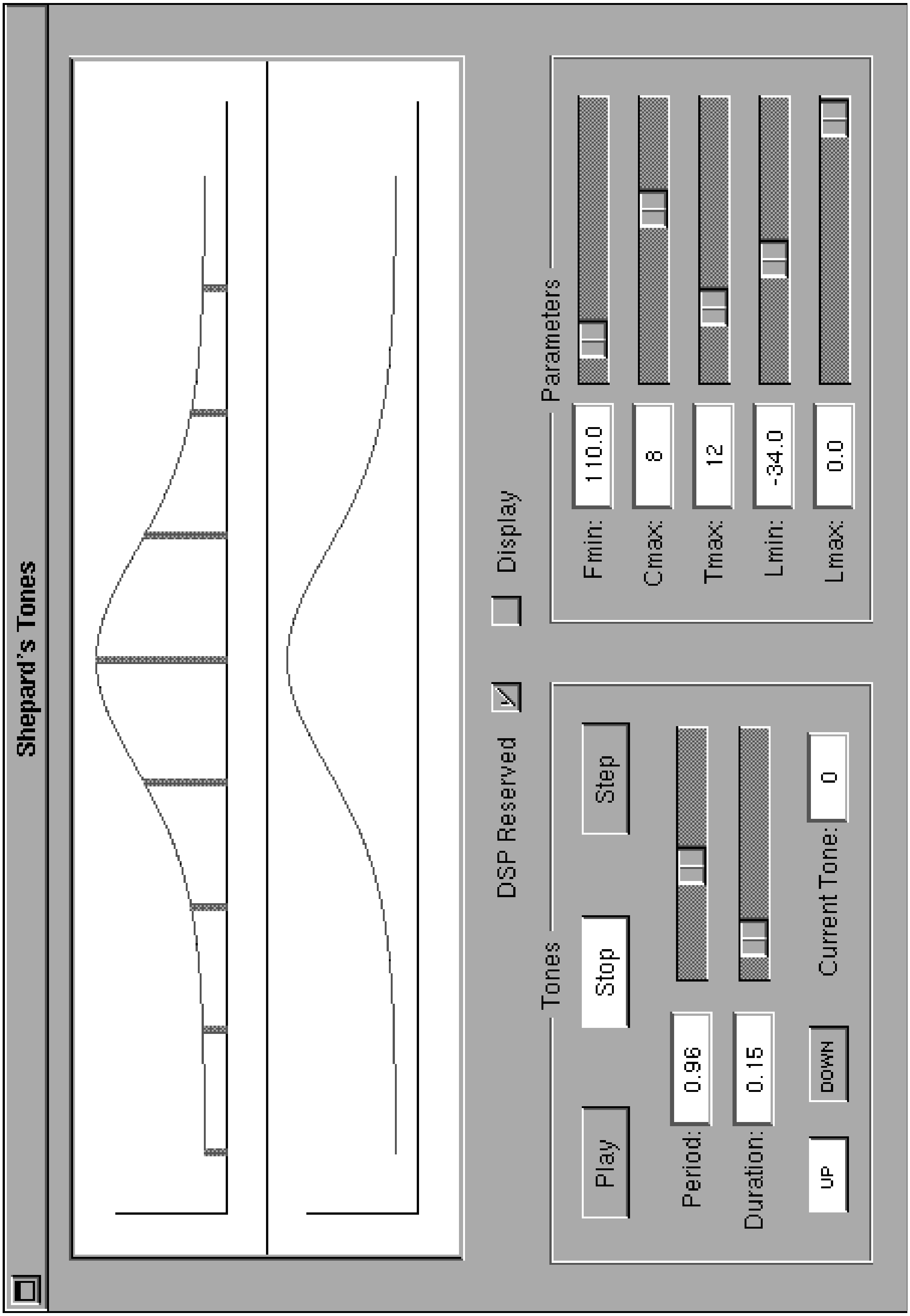
Figure 7 shows the controls and display used for the demonstration. The top panel shows the composition of the first tone, as in Figure 6. When the toggle switch labelled *Display* is selected, the panel below shows the composition of the current tone in the sequence. These two displays are very useful in explaining the cause of the illusion. The toggle switch is provided so that the sequence can be played without giving clues as to the cause of the illusion.

Timed rendering of the sequence is controlled by the *Play* and *Stop* buttons and the values of the *Period* and *Duration* fields. When *Play* is selected, the twelve tones are played in sequence, one every *period* seconds and each for the given duration. When the *Display* switch is selected, the second display panel is updated each time a tone is played and the value of the *Current Tone* field is changed to reflect the position of the current tone in the sequence. The *Step* button can be used to play just the next tone in the sequence, which is useful when explaining the contents of the display panels. The *up* and *down* buttons control the direction in which the sequence is traversed and affect both the *Play* and *Step* modes.

The period and duration can be modified either by editing the text field or adjusting the corresponding slider. These values can be used to explore an interesting phenomenon. By default the tones are separated by 800 milliseconds of silence. When this value is reduced, either by shortening the period or lengthening the duration, the illusion begins to break down [33].

The parameters in the lower right part of the display are given the names originally used by Shepard. *Fmin* refers to the frequency of the lowest component of the first tone; all other components are calculated relative to this value.

Cmax indicates the number of components in each tone, and can be used to demonstrate



Period: 0.96

Duration: 0.15

Current Tone: 0

Fmin: 110.0

Cmax: 8

Tmax: 12

Lmin: -34.0

Lmax: 0.0

Figure 7: Shepard's Tones Implementation

how the illusion develops. With only two components (which requires that $Fmin$ be increased to make the lowest component audible), the transition from the twelfth tone to the first is relatively slight whereas it is clear as the intermediate tones are played that the lower of the two components comes to dominate the higher. With three components, however, neither transition is apparent; with eight components, the transitions are masked completely. When this value is set to n, the frequency of the highest component is

$$\frac{2^n Fmin}{\sqrt[12]{2}} = 2^{n-1/12} Fmin.$$

If this frequency exceeds half the digital audio sample rate, an effect called *foldover* will result in lower frequency artifacts. $Cmax$ is therefore clamped to ensure that this maximum frequency is not exceeded.

$Tmax$ determines the number of tones in the sequence and is usually left at twelve. With n tones in the sequence, components are shifted in frequency by a factor of $\sqrt[n]{2}$ at each step to ensure equal spacing in an octave range.

$Lmin$ and $Lmax$ are used in the calculation of the amplitude envelope. These values can be adjusted to reverse its shape, making it low in the middle and high at the ends. In this configuration, the low and high components are accentuated and the transition from the twelfth to the first tone is apparent. This can be used to emphasize the importance of the extreme components being below the psychoacoustical threshold.

2.2.2 The Tritone Paradox

Introduction

Relative pitch discrimination cues have been carefully removed from Shepard's Tones, but proximity or gestalt cues remain. If the first tone is an A and the second a $B^\flat$, then the transition between the two can be regarded as either an increase of one semitone or a decrease of eleven. The former is the more natural choice and so the pitch appears to increase. Suppose, however, that the second tone is an $E^\flat$, the pitch exactly half way between the initial A and the A an octave higher (the *tritone* of the octave). If the relative pitch discrimination cues have been completely eliminated, the decision as to which tone is higher in pitch should be arbitrary. This arrangement gives rise to the *tritone paradox* first described by Deutsch [8].

The situation is analogous to the well known Necker cube (Figure 8), an illusion first published by the Swiss crystalographer in 1831. Although the figure is drawn as a network of lines in two dimensions, it is most naturally perceived as the projection of a three dimensional cube. Because the projection is orthographic rather than perspective, there are no depth cues to inform the decision as to which face is nearer. One interpretation has corner A at the lower left of the nearer face while the other has corner B at the upper right of the nearer one. An individual can typically switch between these interpretations at will.

The pitch ambiguity of the tritone paradox, like the orientation of the Necker cube, can be resolved in either direction. However, the tritone effect is unique in that a single individual

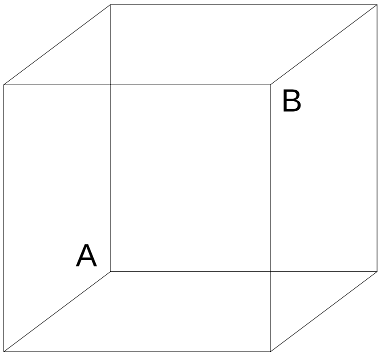


Figure 8: the Necker Cube

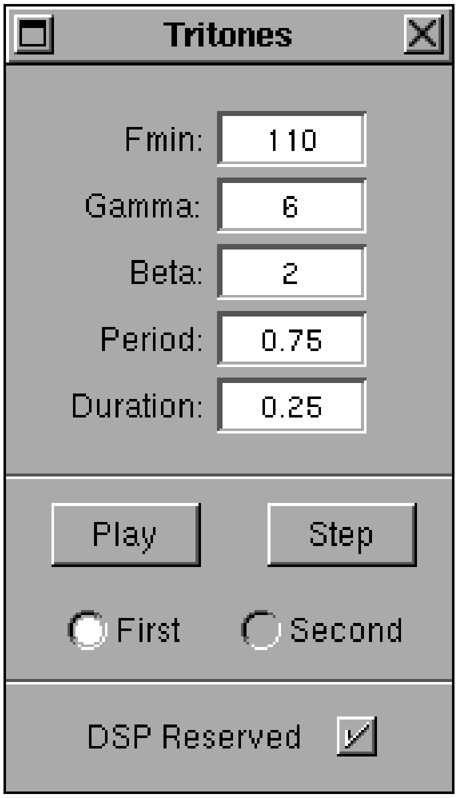
is typically unable to resolve it in both directions. A person who hears one tone as higher than the other will *always* hear that tone as higher. Furthermore, the population is divided between those who hear one tone as higher and those who hear the other as higher. Deutsch has reported finding a systematic difference in this respect between subjects from California and those from Great Britain [9].

Even more interesting is that the direction in which the ambiguity is resolved depends on the absolute pitches of the two tones involved. Although few people have "perfect pitch" (the ability to properly name the pitch of a tone played in isolation), this observation lends support to the claim that most people are able to use absolute pitch information to at least a limited extent.

Presentation

Depending on the audience, it is sometimes useful to begin a demonstration of the tritone paradox with a brief introduction to the Necker cube. Our implementation provides an interactive Necker cube display in which the cube can be rotated about a vertical axis. When the figure is stationary, one can easily switch between the two interpretations of orientation. A *Motor* button engages an automatic rotation mode whose speed is controlled by adjustable text field and slider values. In this mode, the cube appears to be spinning in a well defined direction, but that direction can be made to reverse in the same way that the orientation of the static image reverses. Some people find it difficult to cause a direction reversal without glancing away a number of times. The corner which appears closer when the gaze is re-established will determine the apparent direction of rotation.

The tones are controlled using the panel shown in Figure 9. The *Step* button allows the tones to be played one at a time while the *Play* buttons causes them to alternate repeatedly. The labels below these buttons (*First* and *Second*) are used to identify the tone currently playing. The white dot is always beside the label of the tone that was most recently dispatched. These labels allow the presenter, as the tones are playing, to ask the audience



Fmin: 110

Gamma: 6

Beta: 2

Period: 0.75

Duration: 0.25

DSP Reserved ☑

Figure 9: Controls for the Tritone Paradox Implementation

which tone is perceived as higher in pitch, the first or the second.

The parameters are similar to those for Shepard's Tones, since the tritones are constructed in a similar fashion. *Fmin* is the base frequency of the first tone. The base frequency of the second is computed as $Fmin\sqrt{Beta}$, so the tritone is generated only when $Beta = 2$. *Gamma* determines the number of individual components in each tone. As for Shepard's Tones, this value is clamped to ensure that the highest components do not produce lower frequency artifacts as a result of *foldover*. The *Duration* determines the length of each tone, and *Period* controls the timing of the presentation.

We have found that with a base frequency of 110.0 Hz, most people perceive the second tone as higher in pitch. This is not surprising since the second tone is derived from the first by increasing the frequency of each of its components, producing an average spectral content that is in fact higher. However, larger groups will usually contain a few individuals who disagree. It is interesting to watch these few as they develop the courage to voice their disagreement. This courage arises from their inability to force a reversal in their perception of relative pitch. In fact, very few observers are able to accomplish such a reversal. However, when *Fmin* is changed to a value of 160.0 Hz, the reversal is impossible to avoid: those who heard the second tone as higher now perceive it as lower, and vice versa.

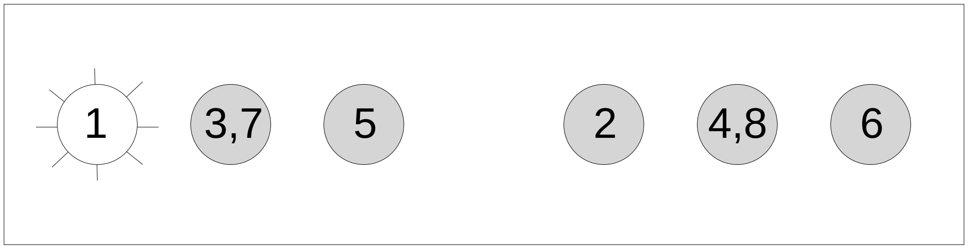


Figure 10: Visual Streaming

2.3 Auditory Streaming

One of the more formidable challenges in creating a rich acoustic environment is to generate independent auditory streams that are naturally perceived as distinct. When sitting at a workstation one has little difficulty in distinguishing the whir of the CPU or hard disk fan from the click of the keys from voices in the room, even though these may all have frequencies in common. The auditory system effortlessly processes the extremely complex auditory stream that reaches the ear into logically distinct components. This process is known as *auditory scene analysis*.

Bregman offers a good introduction to this area and has compiled a comprehensive summary of related work [6]. The first chapter provides a good overview using a number of effects that are representative of the area. The remainder of the book is devoted to the many studies that relate to various aspects of these basic effects. In order to gain a deeper understanding of the issues involved, as well as experience in creating auditory scenes that can be easily analyzed, we have implemented three of the effects.

2.3.1 Pitch separation

Introduction

This effect involves the creation of two *auditory streams* that can be made to fuse into a single percept and to separate again into two apparently distinct streams by controlling simple parameters such as pitch, loudness and presentation speed. Figure 10 illustrates an analogous visual phenomenon. The figure shows a panel in which there are six lamps in a row, only one of which is lit at any given time. The lamps are repeatedly lit in the sequence indicated in the figure. When the rate at which the sequence progresses is slow, it is natural to perceive a complex pattern that jumps back and forth between the left and right groups of three. When the sequence is repeated very quickly, however, it is more natural to perceive two lit lamps, one moving smoothly back and forth through the left group of three and the other doing the same in the right group. At an intermediate presentation speed, the behaviour of the display is unclear.

The auditory analog of this effect simply replaces the lamps with tones of increasing pitch, with the pitch separation being equal except between the third and fourth tones where it is considerably larger. The tones are played in the same order in which the lamps are lit,

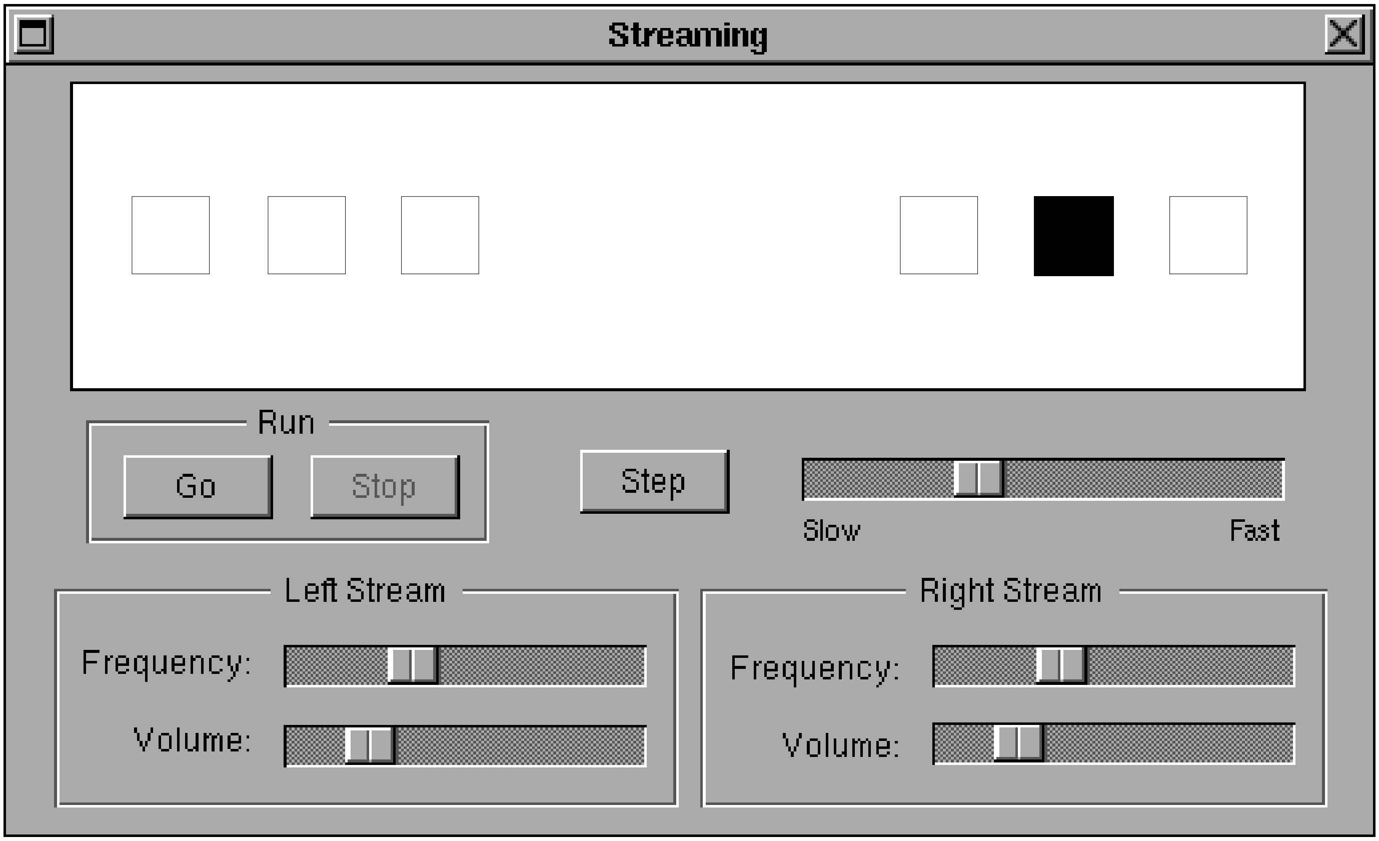


Figure 11: Visual and Auditory Streaming Demonstration

alternating between tones from the lower pitched group and tones from the higher one. At slow presentation speeds it is natural to hear a single tone that oscillates in pitch with each step. At higher speeds, however, the *streaming* effect first seen with the lamps becomes apparent as it is more natural to perceive two tones, one moving back and forth through the lower pitched tones and the other moving back and forth through the higher group. The two separate streams appear to be half as fast as the original and are out of phase by 45 degrees.

Presentation

The implementation of these effects synchronizes the presentation of the visual and auditory versions, playing each tone as the corresponding lamp is lit. However, it is often good to begin by muting the sound and showing just the visual effect (a mute button is not provided since the physical volume knob is quite effective).

Figure 11 shows the display panel and controls available for this demonstration. The presentation is started simply by clicking the *Go* button. The default speed is sufficiently slow that it should be easy to track the progress of the single lamp (box) as it moves between the six fixed positions. The *Slow/Fast* slider can be used to gradually increase the speed of presentation (note that, for technical reasons, speed adjustments take affect only at the beginning of the next cycle). As the speed increases, it should become more natural to interpret the display as two boxes moving in tandem.

The same procedure can be followed, once the volume is restored, to produce the auditory effect. The remaining controls are used to adjust the base frequency and volume of the three lower tones or the three higher tones. After selecting a speed at which the two sets of tones are easily perceived as distinct streams, the base frequencies of the two sets can be adjusted so that they overlap in their range of pitch. In this configuration there is no perceptual basis on which to divide the tones into two or more sets, so the streams again merge, this time at a higher presentation rate. The volumes of the two sets of tones can then be adjusted (using the sliders, not the knob) so that one set is fairly loud while the other is fairly subdued. This again provides a basis for logical separation and most listeners are able to again hear two distinct streams.

Other combinations of pitch, volume and presentation rate are possible, although many are not convincing. The sequence of manipulations outlined above has proven successful in persuading listeners that the number of perceived auditory streams can be controlled with these parameters. Experience with this demonstration suggests that the desired effects are difficult to achieve, even when the display meets all requirements of timing and tone quality. That the demonstration is effective when presented carefully, however, is promising.

2.3.2 Trill threshold

Introduction

The concept of a *trill threshold* was used as the basis for a more formal investigation of the use of pitch and presentation speed to control auditory streaming. In musical parlance, a trill is a pair of notes, separated in pitch by one or two semitones, that are played rapidly in alternating succession. It is natural to perceive a trill as an atomic musical object rather than the composition of individual elements. The *trill threshold* is the point at which the difference in pitch between the two notes is sufficiently large that the trill appears to *break*. At greater pitch differences the alternation of notes is no longer perceived as an atomic musical object. This effect is essentially a two note version of the pitch separation effect described in Section 2.3.1.

The investigation of this effect involved determining for each of a number of presentation rates the pitch separation at which the trill appeared to break. Two measurements were actually made: the pitch at which the trill appeared to break as the separation was increased, and the pitch at which it appeared to be re-established as the separation was subsequently decreased.

The graph in Figure 12 is a qualitative reproduction of the results. The steeper of the two curves represents the threshold at which the trill was perceived to break. It shows that as the alternation speed of the two notes slows, the trill can be sustained with greater pitch separations. The lower curve, which is nearly flat, represents the threshold at which the trill was re-established. These curves divide the graph into three regions. The points in the top region represent combinations of presentation speed and pitch separation at which it is

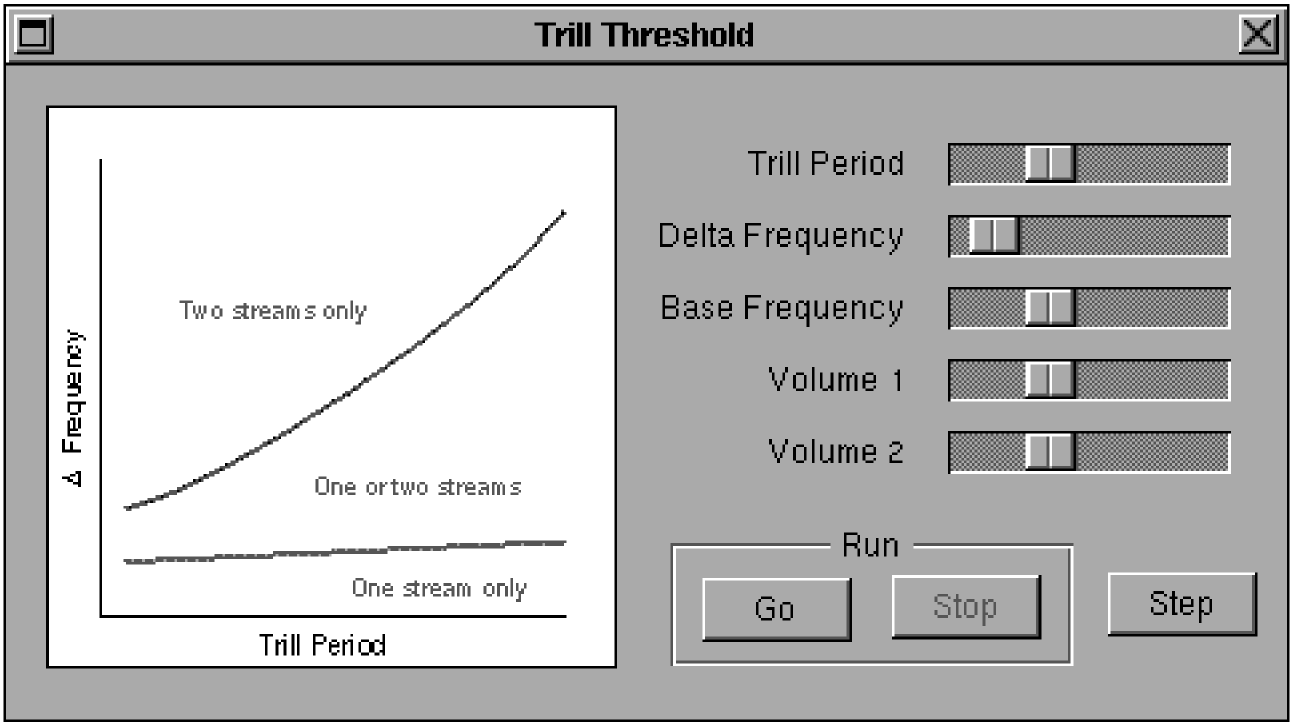


Figure 12: Trill Threshold Demonstration

impossible to perceive the trill as an atomic object (a single stream). Similarly, combinations in the bottom region cannot be perceived as distinct notes. The middle region represents combinations that were perceived as a single stream during the increasing pitch separation condition, but as two streams during the decreasing condition. In other words, there is a large area for which perception of the trill as an atomic object or otherwise is a matter of conscious attention. The existence of such an area emphasizes the difficulty of precisely controlling stream separation in an auditory display.

Presentation

Figure 12 shows the controls for the Trill Threshold demonstration. The presentation is started simply by clicking the *Go* button. Slider controls are available to control the base frequency of the trill (i.e., the frequency of the lower tone) and the volume of each tone. The pitch separation is controlled by the *Delta Frequency* slider and the *Trill Period* slider determines the presentation rate. It is usually unnecessary to adjust the base frequency or either volume control for an effective demonstration. The graph, described in the introduction above, is convenient as a presentation aid but can also be used to set the trill period and frequency separation. Simply clicking on the graph will adjust these two parameters to the combination indicated, and an X is marked at the selected position. This is convenient for exploring the trade off between the two parameters. It can be misleading, however, because the graph is only an approximation, so the trill will not necessarily break precisely where the frequency separation value crosses the curve.

2.3.3 Rhythmic streaming

Introduction

The rhythmic streaming effect is similar to the previous two in that two separable streams are interleaved to produce a complex pattern. It differs, however, in two important ways. First, the tones within a component stream have the same pitch. The pitch of each stream can be adjusted to achieve separation between the two, but the tones within each stream remain identical. Second, each stream has four tones, but the even numbered tones in the second stream are silent. An equivalent interpretation is that the second stream has the same period but half as many tones and half the presentation rate. In either case, the second stream lags the first by 45 degrees.

The top panel of Figure 13 illustrates the arrangement, where time progresses horizontally. The squares represent the tones of the first stream, the triangles the audible tones of the second stream, and the circles the silent tones of the second. When this pattern is sounded with a period of about one second, it produces a rhythm similar to the that of a galloping horse. The two logical streams are *perceived* as one since there is no criterion by which to distinguish them. At slower presentation speeds the galloping effect is less pronounced, but the two streams are still perceived as one.

When the pitch of the second stream is adjusted upward or downward, the streams gradually become distinct and begin to diverge perceptually. There is a difference in pitch, analogous to the trill threshold, beyond which it is difficult to perceive the two streams as one. In this condition the rhythmic galloping effect disappears completely. As the pitch separation of the two streams is again reduced, the streams begin to combine and the rhythmic effect re-emerges.

Presentation

Figure 13 shows the panel used to control the rhythmic streaming demonstration. As for the previous two auditory streaming effects, the presentation is started simply by clicking the *Go* button. Similarly, the *Slow/Fast* slider controls the rate of presentation.

The pitch and volume of the first stream can be adjusted using the controls in the *Fixed Tone* area. The *Moving Tone* controls refer to the second stream, since its pitch moves in relation to the first. As the frequency of the moving tone is gradually increased, the pitch separation is depicted graphically by drawing its triangles and circles at a higher position than the squares. When the separation is sufficient, the rhythmic effect should break down.

Recall that the trill threshold experiment identified a large region in which the number of perceived streams (one or two) is a matter of conscious attention. The corresponding region for this effect appears to be at least as large. Most listeners are able to hear the rhythm at pitch separations as large as an octave, and are able to ignore it until the separation is nearly zero. The situation was improved with the addition of the *Animate* switch. This control

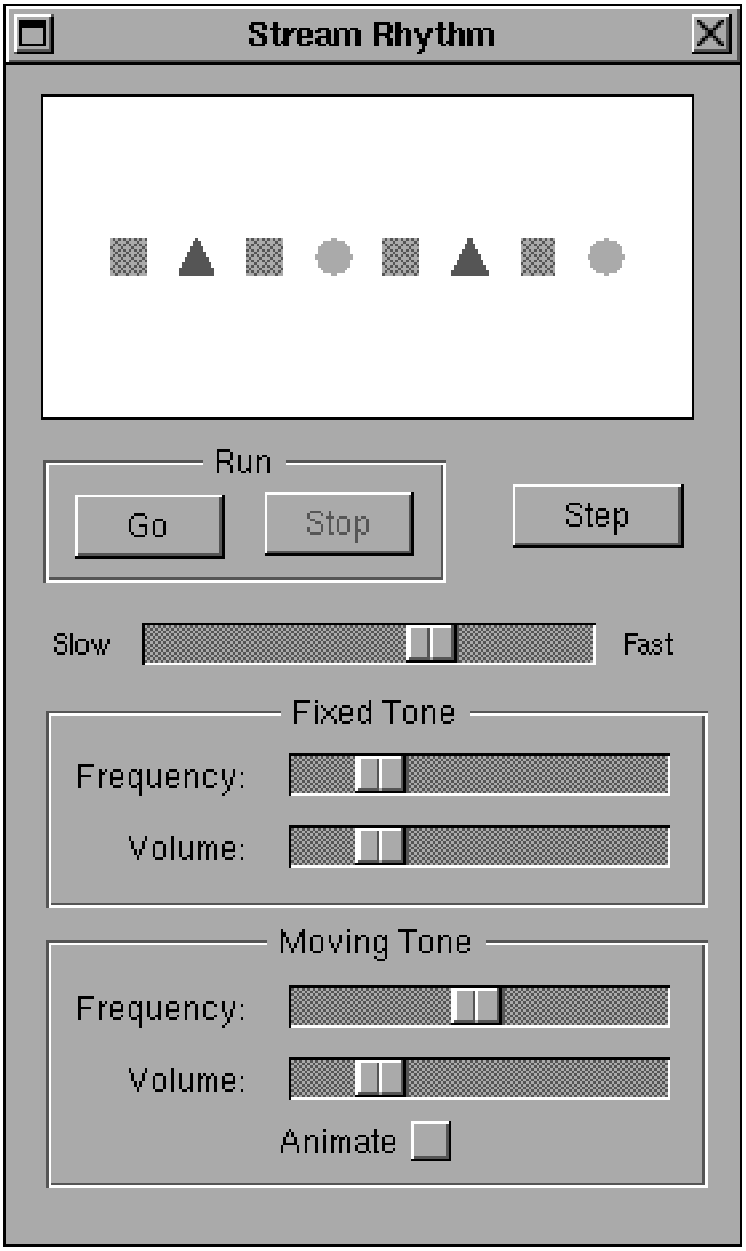


Figure 13: Rhythmic Streaming Demonstration

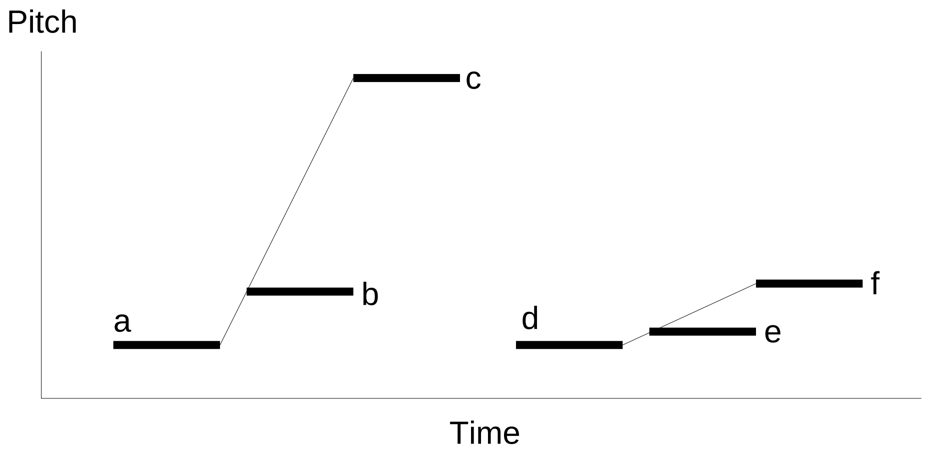


Figure 14: Time vs. Pitch in Auditory Scene Analysis

toggles a mode in which the pitch of the moving sequence is increased by one semitone each cycle (until it reaches the maximum at which time its direction reverses). With this controlled presentation, most listeners are able to retain the rhythmic effect until a separation of ten of eleven semitones (nearly but not quite an octave) is reached. At this point one is typically surprised to discover that the rhythmic effect has suddenly vanished. As the separation is decreased again, the rhythm reappears more gradually.

The original implementation of this effect had a small imperfection in the timing of the tone generation, caused by the need to update the display. When a delay of less than ten milliseconds was incurred at the beginning of each cycle, the rhythmic effect was almost totally destroyed. When the drawing code was optimized, the effect became much more apparent. A slight delay remains, however, and listeners often remark that it is distracting.

2.3.4 Discussion

Each of these effects is governed by the same trade off between pitch separation and presentation speed. It has been postulated that the normal process of auditory cognition interprets tones as belonging to the same stream unless it is unable to *track* the change between them. Figure 14 illustrates two sets of tones, each tone represented by a thick horizontal line. In the first set, the transition from tone a to tone b is relatively steep, even though their pitch separation is small. This is because they are presented in quick succession. The transition from tone a to tone c is just as steep, even though there is more time between them, because the pitch separation is large. The timing of the second set of tones is identical to that of the first, but the transitions from d to e and from d to f are more gradual because the pitch separations are smaller. Although the actual mechanism is not yet well understood, it is believed that the auditory system will track transitions that do not exceed a certain slope. Tones are partitioned into sets that can be connected by transitions below this threshold, thereby defining auditory streams. The ability to track steep transitions appears to be enhanced by conscious attention, accounting for the large middle region of the trill threshold graph (Figure 12).

Early versions of these demonstrations were not particularly effective. Timing problems such as the one described for the rhythmic effect caused numerous difficulties. A flaw somewhere in the tone synthesis software causes audible clicks to be produced at the end of each tone as the decay portion of the amplitude envelope is processed improperly. At high presentation rates these clicks detract significantly from the desired effects.

Even with careful tuning, a convincing presentation is still difficult to achieve. In a workstation environment where independent applications are producing auditory streams that are meant to be distinct, a high degree of cooperation would be required to ensure separation based on properties such as pitch. It may therefore be better to rely on the richness and ease of recognition of a timbral vocabulary to achieve the desired separation.

2.4 Visual Capture

Introduction

Visual capture refers to the process by which an auditory stimulus comes to be cognitively associated with a visual one. It is a natural process that allows us to easily match sounds with the objects or events that caused them. When the events are distant, the relationship is sometimes strained by the fact that light travels much faster than sound.

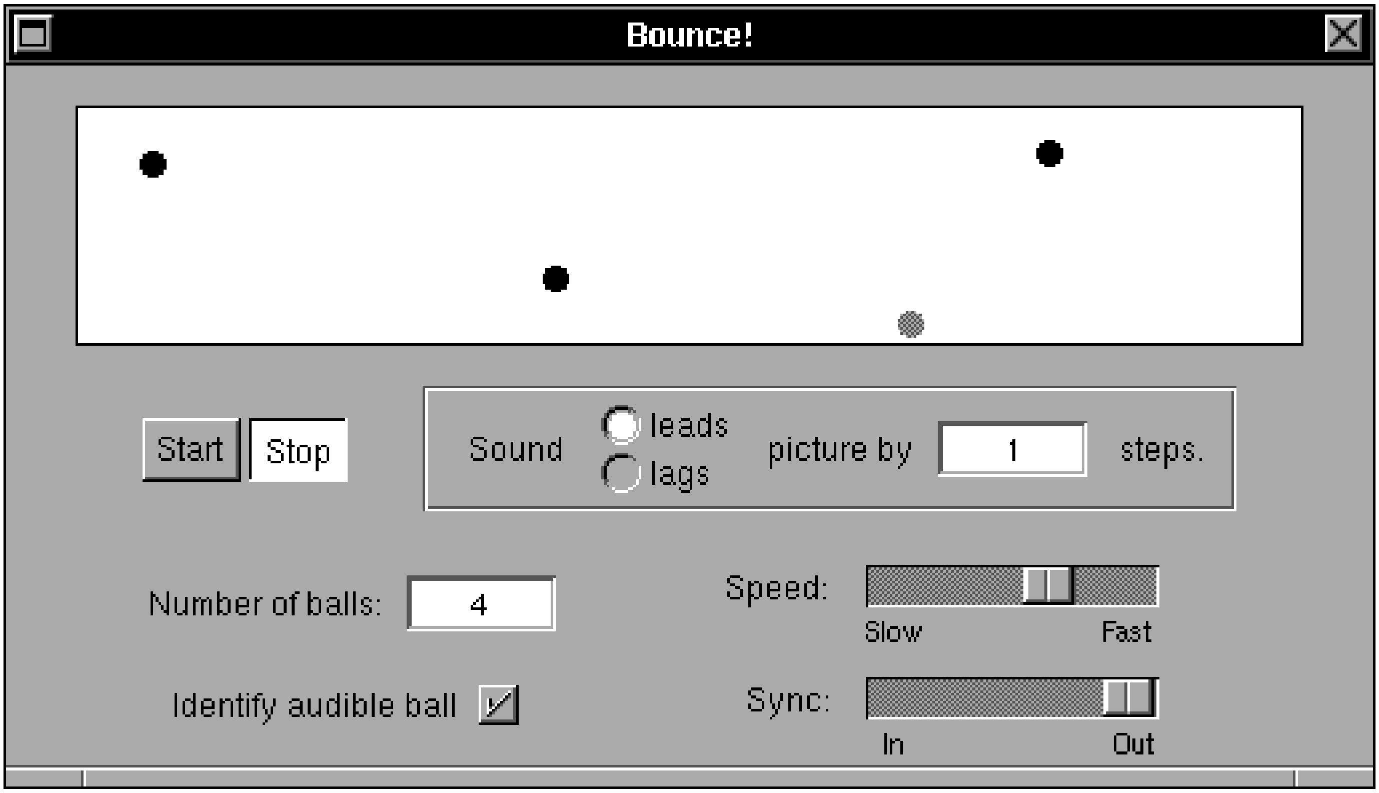
The psychology literature contains some relevant material, such as Warren's study of the auditory perception of breaking and bouncing events [38], but the phenomenon is not yet well understood. The synchronization of sound and image is especially relevant in film. Metz offers several enlightening observations concerning the way in which one naturally thinks of "aural objects" in terms of the physical or visual objects that created them [27] (giving rise, for example, to the peculiar phrase "off-screen voice"). Paul has studied the effects of poorly synchronized speech in film [29], and the film scoring guide of Karlin and Wright [23] has a number of sections discussing the limits of effective synchronization, which range from a few milliseconds to a few seconds, depending on the situation.

In the domain of computer software, the designers of video games have developed a fairly sophisticated set of guidelines for creating sound effects that are useful as well as entertaining. These are often proprietary, however, and have not been subjected to systematic evaluation.

In order to gain a somewhat deeper insight as well as some practical experience in these matters, we implemented a simple application in which a bouncing ball is accompanied by a sound that begins softly shortly before contact with the ground, reaches maximum loudness at the moment of contact, then briefly fades. Several distractors can be added to the scene, and the synchronization between sound and image can be adjusted.

Presentation

Figure 15 shows the system we have developed to explore the phenomenon of visual capture. The top panel shows a number of bouncing balls. They move horizontally at a constant speed as they accelerate along a parabolic trajectory toward the ground then rebound from



Sound ○ leads ○ lags picture by 1 steps.

Number of balls: 4

Speed:

Identify audible ball ☑

Sync:

Figure 15: Visual Capture Demonstration

it. The vertical extent of their trajectory can be increased by increasing the height of the entire window. The animation is started and stopped with a pair of radio buttons.

The number of balls in the display is controlled through an editable text field (the default is one, but it is more interesting with between three and six balls). Only one ball makes an audible bounce, even when others are present and actively bouncing. A typical demonstration involves configuring the various parameters, then asking observers to identify the ball with which the sound is associated. Selecting the *Identify audible ball* switch causes this ball to be drawn in bright red, making it visually salient and allowing quick confirmation of a guess.

The *Sync* slider affects the synchronization of ball trajectories and has nothing to do with the sound. When set to the *in sync* position, the balls bounce with the same period and phase, causing them to remain aligned horizontally. When completely out of sync, the motions of the balls are completely unrelated. As the use of a slider control suggests, there is a continuum between these extremes.

The remaining controls affect the synchronization of the bouncing noise with the trajectory of the corresponding ball. They allow the demonstrator to compose a sentence that describes the amount by which the sound leads or lags the image. With only one ball, it does not seem unusual to associate the bouncing sound with it even when the sound and image are

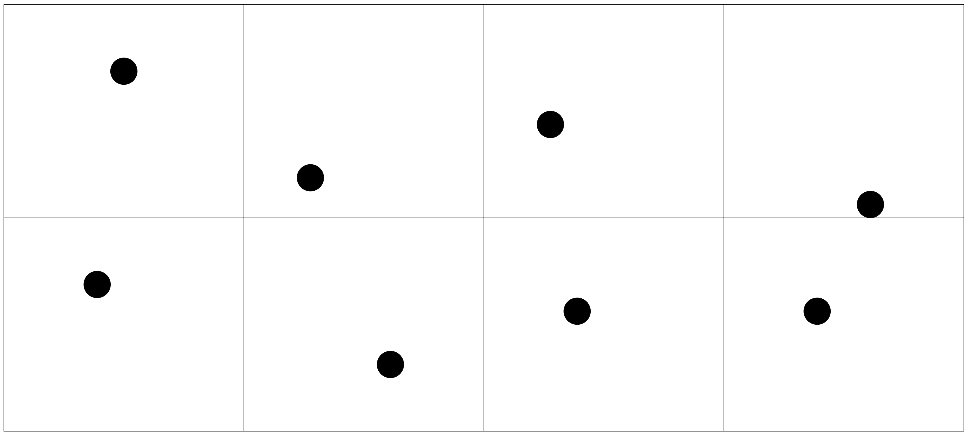


Figure 16: Configuration for Visual Capture Experiment

poorly synchronized. The situation becomes more interesting, however, when multiple balls are shown. When the sound neither leads nor lags the image, it is generally fairly easy to identify the "audible ball". It becomes considerably more challenging as the synchronization degrades.

One would expect that it would be natural to associate an image with a sound that lags it, since the discrepancy between the speeds of light and sound commonly produces this situation. However, informal evaluation using this program has suggested the opposite. Observers appear to be more successful at identifying the correct ball when the sound leads the image, rather than lagging it. This is in accord with well established guidelines for film editing.

A set of simple but useful experiments could easily be derived from this demonstration, aimed at answering questions concerning the magnitude and parity of acceptable asynchrony under various circumstances. A subject would be presented with a number of trials with various degrees of synchronization and numbers of distractors, and would be asked to identify the audible ball. If the balls were numbered, keyboard input would provide an easy means of response. A potentially better arrangement is shown in Figure 16. In this case, each ball occupies its own portion of the display. The choice of which ball is producing the noise could be indicated simply by clicking on the appropriate cell with the mouse pointer.

3 Implementation

The introductions of the previous section are intended to be independent of any particular implementation. The discussion of presentation is necessarily tied to the specific demonstration programs we have implemented, but the descriptions would equally apply to programs on any platform providing the same functionality through a similar interface. Here we describe the implementation environment and system resources that were employed to produce our specific implementations, and brief instructions are given for starting each demonstration. Once a particular program has been launched, the discussion of its presentation from the previous section should guide an effective demonstration.

3.1 Implementation Environment

All of the demonstration programs described here were implemented under version 2.1 of the NeXTStep operating system. They have all been tested under version 3.0 and should work on any subsequent system release. The InterfaceBuilder application and the Application Kit class libraries were exploited fully to produce effective interfaces and visual aids with modest effort.

The Music Kit is also an integral part of those implementations that produce sound (i.e., all but the demonstration of visual illusions). It is a sophisticated software library that controls DSP synthesis at many levels of abstraction, from the hardware level to that of orchestras, performers, instruments and conductors. Because the lower levels of the kit are hardware dependent, these implementations are currently restricted to original NeXT hardware that has a digital signal processor as a standard and integral component. Additional drivers have been produced to allow the Music Kit to function on Intel based platforms with suitable sound cards, but none of these applications have been tested in that environment.

Since they use only standard NeXTStep software components, it should be straight forward to port these programs to any platform that supports NeXTStep and the Music Kit (the latter is now a separate public domain offering maintained by CCRMA at Stanford University). As implementations of the newer OpenStep architecture become available (on such platforms as Solaris and Windows NT), the porting effort should again be straight forward although it is unlikely that any of the programs will compile without modification. Of greater concern is that OpenStep, unlike NeXTStep, does not require that Mach provide the underlying operating system kernel. The implementations of the three auditory streaming effects and of the visual capture demonstration all use a sound server, described in Section 3.2, that uses the Mach Interface Generator (Mig) to build its client and server RPC stubs. The server also uses Mach ports rather than IP sockets for communication. To adapt these programs to an OpenStep platform that does not use Mach, one would have to either write replacement stubs for the server and its client library and replace the Mach port implementation with TCP/IP sockets, or abandon the sound server and implement the sound generation directly within the demonstration program. The latter is almost certain to be the easier choice.

3.2 Sound Server

The ultimate goal of this work, as stated earlier, is to produce a rich acoustic environment in which independent applications and services generate informative auditory streams that do not destructively interfere with each other. The Music Kit, on the other hand, is designed with musical performance in mind. In order to guarantee precise timing of an application's performance, it requires that the application take complete control of the DSP synthesis hardware. A demonstration of the combination of many effects would therefore require a single monolithic application that can produce them all. This in fact is the historical reason why the three auditory streaming effects are all part of the same demonstration program.

To provide greater flexibility in this regard, a *sound server* was built that assumed exclusive control of the DSP hardware, using it to multiplex the synthesis requests of multiple clients. A single DSP is capable of sustaining between fifteen and twenty simultaneous voices, depending on the mixture of synthesis algorithms requested. The synthesis requirements of the programs described here all involve short tones with intervals of silence. An efficient mechanism for rapidly sharing DSP resources during these silent intervals resulted in a server that can support a large number of clients, combining to produce acoustic displays of extremely high complexity.

The sound server was eventually redesigned to separate scheduling from synthesis facilities, and extended to allow synthesis to be combined with the presentation of other media [16]. Its API, on the other hand, was extended to support new media types but was otherwise unchanged. An interactive modeller, also described in the document cited above, is used to construct timed sequences of synthesis events that are saved in files in human readable form. The demonstration programs are linked with a set of class libraries that allow *sequence objects* to be initialized from these files and dispatched to the sound server with appropriate synchronization instructions. The details of contacting and communicating with the server are encapsulated in these classes, making the sound generation portion of the demonstration programs extremely simple.

As mentioned above, the server uses the Mach Interface Generator to build its server and client library RPC stubs, and communicates via Mach ports rather than IP sockets. The remainder of the implementation however uses standard Unix system resources and is implemented in ANSI C. It should therefore be a straight forward project to replace the Mach dependencies with equivalent Unix facilities to produce a server that runs on any Unix platform having suitable synthesis hardware.

3.3 Application Specifics

The lessons we have learned through showing these demonstrations to a large and diverse audience are summarized in the discussions of Section 2, and the most important ones are reiterated in the conclusion. The remainder of this section outlines technical difficulties that were encountered in attempting to produce effective demonstrations, and also provides platform specific instructions for launching each program.

3.3.1 Shepard's Tones

The Shepard's Tones demonstration program has the name *Shepard* and is completely self contained. It is started simply by double clicking its icon.

Both the Shepard's Tones and the Tritone implementations originally made use of the sound server described earlier to allow simultaneous access to the DSP synthesis resources. However, because of their popularity, they were both made to be self contained (and therefore easier to install and use) by extending them to control the Music Kit directly. As a result, both programs require exclusive control of the DSP. To allow them to work together, they each have one additional control that is unrelated to the perception of the illusions. When either program starts, it attempts to claim the DSP. If it is unsuccessful, an alert panel is displayed and the program continues, but with the tone generation controls disabled. One can regain control by finding the application that currently has the DSP (the file `/tmp/dsp.who` contains this information) and either terminating it or instructing it to relinquish the DSP. Clicking on the *DSP Reserved* switch will then claim the DSP for use by either auditory illusion program. If both programs are run together, the DSP can be relinquished by one and claimed by the other simply by clicking on the *DSP Reserved* switch in first the former, then the latter.

Although sound synthesis requires exclusive control of the DSP hardware, applications can use it simultaneously to play digital audio samples. (In system versions earlier than 3.0, digital samples cannot overlap but can be interleaved by independent applications without additional cooperation.) An attempt was therefore made to free the Shepard's Tones program completely from the resource contention problem by precomputing digital samples for each tone, then playing them back. However, the computation of a new set of twelve tones required over a minute on a NeXT machine with a 25 MHz 68040 processor. This effectively disabled the portion of the demonstration in which the various tone parameters are adjusted to expose the elements that produce the illusion. The power of the DSP hardware and the flexibility of the Music Kit software that controls it are essential for an interactive demonstration of these aspects of the illusion.

Computation of the tones proceeds exactly as described in Shepard's original paper [34], with the exception that a higher base frequency was required to produce a convincing effect. Choosing a value of 110.0 Hz for *Fmin* and retaining Shepard's original values for the remaining parameters was found to produce a very robust illusion. Since a large frequency range is involved, the best configuration may depend strongly on the frequency response of the physical speaker involved. Since all parameters are widely adjustable, however, this should not present a significant obstacle.

3.3.2 Tritone Paradox

The Tritone Paradox demonstration program has the name *Tritone* and, like the Shepard's Tones program, is completely self contained. It is started simply by double clicking its icon.

The preceding discussion of DSP resource contention also applies to the Tritone demonstration. It is also worth noting that the spinning Necker cube is achieved with a hard coded orthographic projection of cube vertices to produce two dimensional PostScript line drawing instructions. It does not require the use of a sophisticated three dimensional graphics library. Computation of the tones proceeds exactly as described in Deutsch's original paper [8].

3.3.3 Auditory streaming

The three auditory streaming effects are combined in a single program called *Streams.app*. It requires the sound server, which can be launched either by double clicking on the *soundServer* icon or typing its name from a shell. If the Streams application cannot locate the sound server during its initialization it presents an alert panel to that effect, then terminates when the panel is dismissed.

In order to achieve a high degree of sharing of DSP resources, the sound server manipulates low level synthesis elements that are normally controlled by other Music Kit classes and whose use is only partially documented. Some aspect of their use is improperly implemented by the server with the result that the trailing portion of amplitude envelopes is ignored. Tones are therefore cut off abruptly rather than decaying, producing an audible click. At slow presentation speeds the clicks are easily ignored, but when presented in rapid succession they destroy the desired auditory effect. The clicks are quiet compared with the tones they terminate, but they are nonetheless very distracting in each of the three streaming demonstrations. This serves to emphasize the sensitivity of the auditory streaming phenomena to the quality of sound generation.

Pitch separation

Once the Streams application is launched, the pitch separation demonstration is selected using the *Streaming* entry found in the main menu.

The tone sequences used in this demonstration were constructed using the sound modeller, as described above. Once a sequence is dispatched to the sound server it plays to completion, even when the *Stop* button is pressed shortly after the sequence begins. Similarly, adjustments of the pitch, volume and speed controls only take effect at the beginning of the next sequence iteration. At higher presentation speeds the delay is less evident, but at slow speeds adjustments should be made only late in the sequence to achieve the appearance of instantaneous response. This restriction could be easily removed, but it has not proven to be a significant impediment.

The clicks generated at the end of each tone produce especially serious degradation of this effect as presentation speed is increased.

This implementation coordinates the generation of tones with the moving of the box in the display. Achieving reasonable synchrony in a self contained implementation is relatively simple, but in this case the sound server generates the tones while the application does its own drawing. The sound server can actually send a message to a specified Mach port each time it dispatches an event from a sequence. This facility is used to signal the Streams application each time another tone is produced. The message contains an integer identifying the tone that was most recently dispatched, allowing the display to be updated to show the corresponding box.

Trill threshold

The *Trill threshold* entry of the main Streams menu selects this effect. It is more forgiving than the first streaming effect, although tone termination clicks do become significant at high presentation speeds.

The graph included in the display can be misleading in two ways. First, it is only a qualitative reproduction of the original experimental data. One should therefore avoid the temptation to expect the trill to break as pitch separation crosses the curve. The temptation is increased by the ease with which combinations of pitch separation and trill speed can be selected by pointing to locations on the graph. Second, while parameter combinations can be selected from the graph, a dragging mode is not available (one was attempted, but the interference of the modal event loop with the timing of the rapid trill alternation caused the effect to degrade substantially).

Rhythmic streaming

The rhythmic streaming panel is activated using the *Rhythm* entry of the main menu. This demonstration is also more forgiving than the first streaming effect. The only difficulty that normally arises is that, as mentioned earlier, there is a very slight pause at the end of every tone sequence iteration as the display is updated. Many listeners will report that this interferes significantly with the streaming effect on which they are attempting to concentrate.

The pitch separation of the two streams is currently constrained to an octave in either direction. Listeners will sometimes report being able to hear the characteristic galloping rhythm even at this maximum separation. A greater maximum pitch separation may therefore be advisable, although we have achieved reasonable success in the current configuration.

3.3.4 Visual capture

The visual capture demonstration program is called *Capture.app* and is launched simply by double clicking its icon. Like the auditory streaming effects it also requires the sound server, which can be launched either by double clicking on the *soundServer* icon or typing its name from a shell. If the Capture application cannot locate the sound server during its initialization it presents an alert panel to that effect, then terminates when the panel is dismissed.

Coordination of sound and image is achieved using the sound server messaging mechanism described above for the pitch separation streaming effect. Animation of each ball is controlled completely by the sound server. A single event sequence is dispatched to the server that begins with events that contain signal messages to be reported at the time of dispatch, but empty synthesis requests. Each time an event is dispatched, the number of the event is conveyed back to the Capture program which draws the ball at the corresponding point along its trajectory. Seven events at the midpoint of the trajectory (where the ball approaches the ground, contacts it and rebounds) have synthesis instructions. The first initiates a tone having a timbre similar to the sound of a gasoline powered lawn mower. The next five synthesis instructions cause that tone to grow louder as the ball approaches the floor, and then fade as it moves away. The final event terminates the tone. When the sequence is played at a relatively fast pace, the ball appears to follow a parabolic trajectory, making a bouncing noise as it contacts the floor.

When several balls are active, each one is controlled by a separate event sequence registered with the sound server. The sequence controlling every ball except the first contains no synthesis instructions since only one ball generates a bouncing noise. The sequences are initialized from a pair of sequence objects, (one containing tone synthesis instructions) which in turn are initialized from a file created interactively by the sound modeller. The sound server is therefore extremely active when even a modest number of balls are active, even though it is generating sound for only one of them. An obscure bug in the sound server occasionally results in a misdirected signal message, causing the display to simply freeze. The sound server still responds properly to other applications, so terminating and restarting the Capture application will rectify the problem. The bug arises infrequently enough that a successful demonstration can almost always be delivered.

3.4 System Requirements

The tone generation requirements of the demonstration programs are generally quite simple. Sinusoidal tones with gradual amplitude attack and decay are sufficient more most of the effects. The Shepard's Tones and Tritone demonstrations require greater control. If one foregoes the ability to adjust the tone parameters, digital samples can be precomputed for simple play back. A responsive interactive demonstration requires a synthesis system that can generate tones with precisely controlled harmonic content. In this case, only octave harmonics are required. It may also be possible to generate the tones by playing seven simple sinusoidal tones simultaneously. However, imprecision in synchronizing either the onset or termination of the tones will weaken the illusion.

Accurate timing presents a much greater challenge, especially on Unix or other time sharing platforms. We have occasionally had difficulty achieving a desired effect during a presentation, only to discover that the machine was being used remotely for a seemingly innocuous activity such as text editing. CPU intensive tasks such as compilation or rendering destroy the effects completely. The Music Kit exploits features of Mach such as light weight threads and fixed priority scheduling to minimize operating system interference in the timing of musical presentations, but the impact of concurrent activity has not been removed entirely. Our sound scheduler also uses a small number of high priority threads in an attempt to process events in a timely fashion, but small inaccuracies are inevitable since neither Unix nor Mach can make real time guarantees. The best platform for timing purposes would be one whose operating system kernel is either very light weight and non-intrusive, or which is capable of accurate real time performance.

Perhaps the most significant challenge is in the area of coordination of sound and image. On the NeXT platform, for example, the DSP is capable of extremely high precision in the control of large and complex musical performances. Drawing and image presentation, on the other hand, is accomplished by sending suitable PostScript instructions to a window server process that manages the display and evaluates or renders the PostScript it is sent. Drawing is therefore delayed both by the overhead of communication with the window server and by the time taken to render the PostScript instructions. These delays can easily amount to several milliseconds or more, which may produce unacceptable synchronization (as is the case in the rhythmic streaming demonstration). A more highly integrated system that combines the presentation of all media types is therefore advisable.

These observations suggest that to create a complex auditory environment of the kind we have described, one will require a combination of hardware and software that provides relatively high quality digital sample play back and sustained tone synthesis, and an integrated media presentation system that, together with the underlying operating system, can guarantee the presentation of both auditory and visual material with delays not exceeding a few milliseconds.

4 Summary and Conclusions

The work discussed in this report represents the first phase of a project whose long term goal is to produce a rich acoustic environment in which the behaviour of multiple independent activities is communicated through perceptually distinguishable auditory streams. Although much is known about auditory phenomena that are experienced in isolation, there is currently neither an empirical nor a theoretical basis for choosing auditory elements for a complex and heterogeneous display in an uncontrolled acoustic environment.

A set of demonstration programs have been designed and implemented in an effort to gain preliminary experience with many of the issues involved. They include two demonstrations of circularities in judgements of relative pitch, three variations of auditory streaming effects in which independent streams can be fused and separated by modifying a small number of parameters, and a program that allows one to explore the limits of asynchrony in visual capture. These programs have been used in demonstrations given in our laboratory and a number of other local institutions. Audiences have ranged in number from one or two individuals to groups of twenty or more and have included interested colleagues, visiting researchers and speakers, benefactors, high school and undergraduate students, and members of the general public.

We have found the relative pitch discrimination illusions to be easily produced and effective under a wide range of ambient acoustic conditions. Auditory stream effects based on variation of pitch, loudness and presentation rate are more difficult to create and control. It is likely that timbre will provide a more effective basis for stream discrimination. These two observations together point to a potential pitfall. In the kind of acoustic environment we envision, some sounds will be brief but other streams will be continuous and of arbitrary duration. This is more easily accomplished with real time synthesis than with the playback of precomputed samples. A system that separates a large number of streams on the basis of timbre may need to use relatively simple algorithms for computing additional harmonic content in order to meet real time constraints. The relative pitch discrimination illusions demonstrate the need for care in this situation.

The demonstration of visual capture is quite simple, but has proven very interesting and suggests several possibilities for immediate investigation. A variety of simple experiments based on the idea of identifying the graphical object associated with an auditory stimulus are possible.

Perhaps of more immediate concern is the need to synthesize independent auditory streams that are naturally perceived as distinct. Although our informal investigations suggest that streams are poorly differentiated by pitch and loudness, a more formal confirmation of this would be valuable. Similarly, the ability of timbre to differentiate streams should be assessed. It should not be difficult to design experiments in which one or more streams are presented simultaneously and subjects are asked to indicate the number they perceive.

Visual capture may also be important in the context of auditory stream separation. Two streams that in isolation are interpreted as a single more complex one may separate perceptually if each is naturally associated with a distinct visual object. Experiments of this type may prove more challenging because of the need to introduce additional factors without confounding the results.

Finally, it is our hope that many of the sounds in the environment we have described will be capable of *subconscious acoustic priming*. That is, a user whose attention has not been focused on a particular auditory stream may nonetheless assimilate the information it carries at a subconscious level. We have begun the design of an experiment to test this hypothesis. It generates a continuous auditory stream that encodes the state of a simple object while the subject performs a continuously engaging distractor task (placement of pieces in a jigsaw puzzle). Both the puzzle display and the auditory stimulus are periodically interrupted to ask a question regarding the state of the auditory object.

References

[1] J. A. Ballas and J. H. Howard, Jr. Interpreting the language of environmental sounds. *Environment and Behaviour*, 19:91–114, 1987.

[2] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1):11–44, Spring 1989.

[3] S. Bly, S. P. Frysinger, D. Lunney, D. L. Mansur, J. J. Mezrich, and R. C. Morrison. Communicating with Sound. In R. Baecker and W. A. S. Buxton, editors, *Readings in Human Computer Interaction: A Multidisciplinary Approach*, pages 420–424. Morgan Kaufmann, Los Altos, California, 1987.

[4] Sara Bly. Presenting information in sound. In *Proceedings of the CHI'82 Conference on Human Factors in Computer Systems*, pages 371–375, New York, 1982. ACM.

[5] Sara Bly. *Sound and Computer Information Presentation*. PhD thesis, University of California, Davis, Computing Science Group, University of California, Davis, March 1982. with Lawrence Livermore National Laboratory.

[6] Albert S. Bregman. *Auditory scene analysis: the perceptual organization of sound*. MIT Press, Cambridge, Mass., 1990.

[7] William Buxton, William Gaver, and Sara Bly. The Use of Non-Speech Audio at the Interface. CHI-89 tutorial 10, April 1989.

[8] Diana Deutsch. The tritone paradox: Effects of spectral variables. *Perception & Psychophysics*, 41(6):563–575, 1987.

[9] Diana Deutsch. Paradoxes of Musical Pitch. *Scientific American*, pages 88–95, August 1992.

[10] Christopher J. DiGiano and Ronald M. Baecker. Program Auralization: Sound Enhancements to the Programming Environment. In Norman Jaffe, editor, *Proceedings of Graphics Interface '92*, pages 44–52, Vancouver, B.C., May 11-15 1992. Canadian Information Proceesing Society.

[11] David C. Earle and Stephen J. Maskell. Fraser cords and reversal of the café wall illusion. *Perception*, 22(4):383–390, 1993.

[12] Brian Evans. Enhancing Scientific Animations with Sonic Maps. In *Proceedings of the International Computer Music Conference*, November 1989.

[13] Brian Evans. Correlating Sonic and Graphic Materials in Scientific Visualization. In Edward J. Farrell, editor, *Extracting Meaning from Complex Data: Processing, Display, Interaction*, pages 154–162, P.O. Box 10, Bellingham, Washington 98227-0010 USA, February 1990. SPIE – The Internation Society for Optical Engineering.

[14] Nicholas Falletta. *The Paradoxicon*, pages 24–34. John Wiley & Sons, Inc., New York, 1990.

[15] Tecumseh Fitch. Sonifying the Body Electric. Imager Laboratory Invited Seminar, January 1995.

[16] Scott Flinn. Coordinating Heterogeneous Time-Based Media Between Independent Applications. Technical Report 95-16, Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, B.C., Canada, V6T 1Z4, July 1995. Available electronically as http://www.cs.ubc.ca/tr/1995/TR-95-16.

[17] Steven P. Frysinger. Applied Research in Auditory Data Representation. In Edward J. Farrell, editor, *Extracting Meaning from Complex Data: Processing, Display, Interaction*, pages 130–139, P.O. Box 10, Bellingham, Washington 98227-0010 USA, February 1990. SPIE – The International Society for Optical Engineering.

[18] William W. Gaver. Auditory Icons: Using Sound in Computer Interfaces. *Human-Computer Interaction*, 2:167–177, 1986.

[19] William W. Gaver. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction*, 4(1):67–94, Spring 1989.

[20] William W. Gaver and Randall B. Smith. Auditory Icons in Large-Scale Collaborative Environments. In *Proceedings of IFIP INTERACT'90: Human-Computer Interaction*, pages 735–740. IFIP, 1990.

[21] William W. Gaver, Randall B. Smith, and Tim O'Shea. Effective Sounds in Complex Systems: the ARKola Simulation. In *Proceedings of ACM SIGCHI '91*, pages 85–90, 1991.

[22] Richard L. Gregory and Priscilla Heard. Border locking and the Café Wall illusion. *Perception*, 8:365–380, 1979.

[23] Fred Karlin and Rayburn Wright. *On the Track: A Guide to Contemporary Film Scoring*. Schirmer Books, A Division of Macmillan, Inc., New York, NY, 1990.

[24] Lester F. Ludwig, Natalio Pincever, and Michael Cohen. Extending the Notion of a Window System to Audio. *IEEE Computer*, 23(8):66–72, 1990.

[25] David Lunney and Robert C. Morrison. Auditory Presentation of Experimental Data. In Edward J. Farrell, editor, *Extracting Meaning from Complex Data: Processing, Display,*

Interaction, pages 140–146, P.O. Box 10, Bellingham, Washington 98227-0010 USA, February 1990. SPIE – The International Society for Optical Engineering.

[26] Mark E. McCourt. Brightness induction and the Café Wall illusion. *Perception*, 12(2):131–142, 1983.

[27] Christian Metz. Aural Objects. In Elisabeth Weis and John Belton, editors, *Film Sound: Theory and Practice*, pages 154–161. Columbia University Press, New York, NY, 1985.

[28] R. Timothy Mullins. Causal Uncertainty and Contextual Cues in the Recognition of Environmental Sounds. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, volume 1, pages 247–251. Human Factors Society, 1988.

[29] Lawrence M. Paul. When Lips and Voice Disagree: Determining the Practical Limits and Consequences of Visual-Auditory Asynchrony. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, volume 1, pages 229–231. Human Factors Society, 1988.

[30] Irwin Pollack and Lawrence Ficks. Information of Elementary Multidimensional Auditory Displays. *Journal of the Acoustical Society of America*, 26(2):155–158, March 1954.

[31] David A. Rabenhorst, Edward J. Farrell, David H. Jameson, Thomas D. Linton, and Jack A. Mandelman. Complementary Visualization and Sonification of Multi-Dimensional Data. In Edward J. Farrell, editor, *Extracting Meaning from Complex Data: Processing, Display, Interaction*, pages 147–153, P.O. Box 10, Bellingham, Washington 98227-0010 USA, February 1990. SPIE – The International Society for Optical Engineering.

[32] Linda A. Roberts and Joel Angiolillo-Bent. The Relative Pleasantness and Distinctiveness of a Variety of Auditory Patterns. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, volume 1, pages 554–557. Human Factors Society, 1988.

[33] Andrew B. Rostron. Some Observations on the Auditory Staircase Illusion. *Perceptual and Motor Skills*, 39:212–214, 1974.

[34] Roger N. Shepard. Circularity in Judgements of Relative Pitch. *The Journal of the Acoustical Society of America*, 36(12):2346–2353, 1964.

[35] Roger N. Shepard. Structural Representations of Musical Pitch. In Diana Deutsch, editor, *The Psychology of Music*, chapter 11, pages 343–390. Academic Press, New York, NY, 1982.

[36] Frankie K. Sun, William B. Cowan, and Kellogg S. Booth. Understanding Visual Effects in a Windowed Environment. In Stephen MacKay and Evelyn M. Kidd, editors, *Proceedings of Graphics Interface '90*, pages 100–107, Halifax, N.S., May 14-18 1990. Canadian Information Processing Society.

[37] N. J. Vanderveer. Ecological acoustics: Human perception of environmental sounds. Dissertation Abstracts International, 40/09B, 4543, 1979. University Microfilms No. 8004002.

[38] W. H. Warren and R. R. Verbrugge. Auditory perception of breaking and bouncing events. *Journal of Experimental Psychology: Human Perception and Performance*, 10:704–712, 1984.

[39] Edward S. Yeung. Pattern Recognition by Audio Representation of Multivariate Analytic Data. *Analytical Chemistry*, 52(7):1120–1123, 1980.