A Comparison of Input Devices in an ISEE Direct Timbre Manipulation Task

Roel Vertegaal
Department of Ergonomics
University of Twente
P.O. Box 217
7500 AE Enschede, The Netherlands

E-mail: R.Vertegaal@wmw.utwente.nl

Barry Eaglestone
Department of Computing
University of Bradford
Bradford BD7 1DP, United Kingdom

E-mail: B.Eaglestone@comp.brad.ac.uk

Abstract

The representation and manipulation of sound within multimedia systems is an important, currently under-researched area. This paper gives an overview of our work on the direct manipulation of audio information, and describes a solution based upon the navigation of four-dimensional scaled timbre spaces. We experimentally evaluated three hardware input devices for use in a timbre space navigation task: the Apple Standard Mouse, Gravis Advanced Mousestick II joystick (absolute and relative) and the Nintendo Power Glove. Results show that the usability of these devices significantly affected the efficacy of the system, and that conventional low-cost, low-dimensional devices provided better performance than the low-cost, multidimensional dataglove.

Keywords: Empirical Research, Direct Manipulation, Human-Synthesizer Interaction, Auditory Perception, Timbre Space, Input Devices, Information Retrieval.

Introduction

The audio dimension of multimedia technology has attracted relatively little attention from HCI researchers [11]. This is in spite of its importance as a medium for human communication and a commodity in a number of specialist applications. Audio-intensive applications exist in the arts and the music, television, cine and video sound industries, and often require large databases of sounds [7, 16]. This paper addresses a particular aspect of this area, the design of interfaces for the manipulation of sound. The first part presents our research into direct manipulation of the timbre of sound. After reviewing related work, mainly in computer music, we describe our development of a specific candidate solution involving the use of a presentation model based upon scaled timbre spaces, with direct manipulation by relocation of sounds within the spaces. Timbre space is a multidimensional space of sounds where each dimension models the variability of sounds with respect to some perceived characteristic. This solution has been implemented for musical application in the Intuitive Sound Editing Environment (ISEE) system [27].

ISEE allows musicians to manipulate the timbre of synthesized sounds by navigating a hierarchy of four-dimensional scaled timbre spaces, effectively integrating expert synthesis design methodology with a perceptual organization of sounds. Two important factors which affect the usability of this type of interface are the visual representation of the timbre spaces, and the hardware interface by which users interact. The second part of the paper describes our research into the second of these factors. Results from an experimental evaluation of input devices for this form of timbre manipulation are presented. Conclusions on the efficacy of the ISEE interface and of each of the devices are drawn from an analysis of the integration of movement, accuracy and movement times.

Direct Manipulation of Timbre

Timbre is best defined as the quality of sound that enables one to discriminate two static tones with equal pitch and loudness. Each natural sound consists of a combination of sine waves with different frequencies, amplitudes and phases. It is the pattern of these so called overtones or harmonics, relative to the perceived pitch and loudness of the sound, that constitutes the timbre of a sound. The interactive control of timbre has always played an important role in our everyday lives. Our speech communication depends on it as a means of conveying information. In western classical music, however, timbre control has traditionally played a somewhat less important role, since much of the *musical* information is conveyed by pitch using complex harmony. For musicians, the control of complex harmony typically requires a different modality than the refined control of timbre. Some musical instruments (e.g., our voice, the violin) provide the musician with considerable timbre control but can only generate a limited number of notes simultaneously. Other instruments, most notably the piano, offer great polyphonic capability at a loss of timbre control. It is for this reason that the piano keyboard has become the most prominent input device for compositional purposes in the classical tradition.

In computer synthesized music, timbre control plays a much more important role than in classical music. Digital sound synthesis technology has provided musicians with an almost unlimited number of timbral possibilities. Nonetheless, the piano keyboard remained the most prominent input device in sound synthesizers. Synthesizer user interface standards for timbre modification have not kept up with recent advances in HCI. With the advent of graphical user interfaces in sound synthesis systems, one would expect the notion of direct manipulation of timbre to have gained ground. An important aspect of direct manipulation is the principle of transparency, where attention shifts from issuing commands to observing results conveyed by feedback [22]. This requires feedback to be consistent with the user's expectations of the task's results. Shneiderman [24] argues that with direct manipulation systems, there may be substantial task-related semantic knowledge, but users need to acquire only a modest amount of computer-related semantic and syntactic knowledge. Task-related semantics should dominate the users' concerns, reducing the distraction of dealing with the computer semantics and syntax. Current synthesis user interfaces, however, are based on the direct use of synthesis model (i.e., sound generating process) parameters, which need not necessarily behave in a perceptually linear or consistent fashion. For example, to change the brightness of a tone digitally synthesized using the Frequency Modulation (FM) synthesis model [6], one would change the *modulation index*. Though most of the time this seems to affect the brightness of the sound, when *modulation feedback* is active the sound can suddenly turn into noise, resulting in a loss of correspondence between the task-related semantics and synthesizer-related semantics. This has led many novice users to the impression that creating sounds on an FM synthesizer is in fact a stochastic process. A more direct mapping between task-related semantics (I want to make a sound brighter) and synthesizer-related semantics (then I need to change the *modulation index* or the *modulation feedback level* or both) could easily be achieved if control would operate at a higher level of abstraction, using a perceptually-based, hardware-independent interface.

Perceptually-based Timbre Control

Buxton, Patel et al. [3] recognized early-on that the manipulation of on-screen sliders representing synthesis hardware parameters is no more than a substitute for the direct manipulation of timbre, and that timbre should ideally be controlled according to perceptual rather than acoustical attributes. They also emphasized the importance of minimizing non-musical problems of the sound synthesis task and permitting the composer to understand the perceptual consequences of their actions. The following paragraphs will review recent advances towards achieving that goal.

Timbre Space

Wessel [29], Grey [12] and Plomp [21] proved it possible to explain differences in musical timbre with far fewer degrees of freedom than are needed by most synthesis algorithms. Grey [12] investigated the perception of musical instrument timbres using multidimensional scaling techniques [23]. Wessel addressed the timbre control problem by implementing a simple control structure based on a perceptual mapping produced with the same technique [30]. In this approach, a *timbre space* is derived from a matrix of timbre dissimilarity judgements made by humans comparing all pairs of a set of timbres. In such a space timbres that are close sound similar, and timbres that are far apart sound different. To use a timbre space as a synthesis control structure one specifies a coordinate in the space using an input device. Synthesis parameters are then generated for that particular point in space. This involves interpolation between the different originally judged timbres. Lee and Wessel [18, 19] have demonstrated how a Mattel Power Glove was used in combination with a neural network to produce this real-time interpolation during performances. This approach elegantly features all constraints for achieving a direct manipulation of timbre, including a well-based formalism for the real-time mapping of low-dimensional perceptual parameters to high-dimensional synthesis model parameters. However, Plomp [21] indicated that when constructing timbre spaces, the dimensions proliferate with the variance in the assessed timbres. This makes it difficult to derive a generalized synthesis model from this strategy. When trying to construct two low-dimensional timbre spaces for two different sets of instruments the dimensions defining these two spaces might vary considerably, which could cause usability problems in generic user interface applications. Generic use of timbre space is also inhibited by the need to use existing sound examples judged by a human panel. How could a musician construct his own timbre spaces? What if he wants to generate totally new sounds?

Grey [12] theorized about the semantics of the dimensions of the 3-D timbre space he derived from an experiment in which 16 closely related instrumental stimulus tones (varying from wind instruments to strings) were compared on timbral similarity. He indicated that one dimension could express instrument family partitioning, another dimension could relate to the brightness of the tones, and a third dimension could relate to the temporal pattern of timbral development at the start of the tones (e.g., noise patterns). Though these conclusions cannot simply be generalized, they do give us an indication of the nature of appropriate parameters to be used when generalizing timbre space as a synthesis model.

The Intuitive Sound Editing Environment

The practicality of using timbre space as a basis for a sound design system is demonstrated by the Intuitive Sound Editing Environment (ISEE) [27]. ISEE is a synthesizer and synthesis model independent user interface designed for sound synthesis applications in both composition and performance[1]. The following description of ISEE omits technical details—those interested should refer to [27].

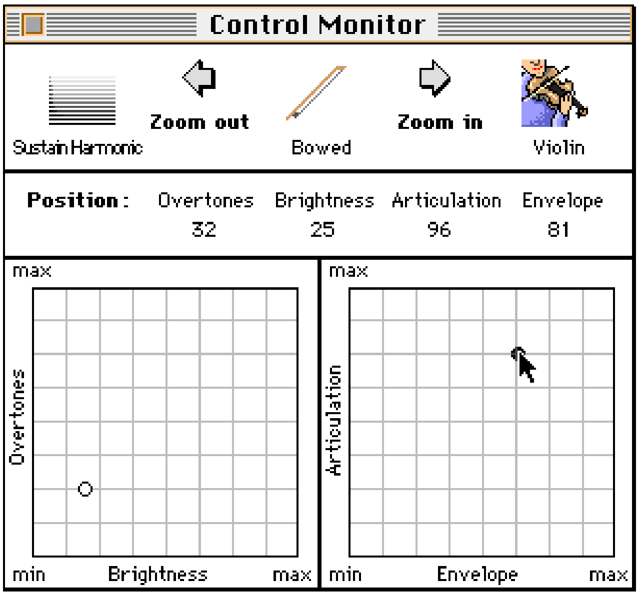

Figure 1: The ISEE controller front-end.

ISEE attempts to generalize the timbre space paradigm for generic user interface purposes by concentrating on the defining dimensions of timbre space. Assuming these parameters have orthogonal properties, every point in space can be defined by combining synthesis data associated with the projections of its coordinates. Four timbre parameters are presented to the user by means of two two-dimensional projections of the four-dimensional timbre space they constitute (see figure 1). The first two of these parameters relate to the frequency contents (spectral envelope) of the sound and the last two to its development in time (temporal envelope): the Overtones parameter controls the frequency pattern of the overtones; the Brightness parameter controls the energy balance between the lower and higher frequencies; the Articulation parameter controls

[1]A demo version of the software can be downloaded from URL `http://reddwarf.wmw.utwente.nl/isee/welcome.html`

the development of the overtones at the start of the sound and the behaviour of noise throughout the sound; and the Envelope parameter controls the speed of the temporal development of the sound. The first three parameters find their origins in the semantics of timbre space as identified by Grey [12], with the Envelope parameter adding generic control of envelope models other the sustained sound used in Grey's research.

Our application of high-level semantics derived from timbre space research can prevent Plomp's proliferation of timbre space dimensions [21] and allows the number of parameters presented to the user to be kept to a consistent minimum. The problem is one of defining functions which map from timbre space dimensions to synthesis parameters, and is simplified by decomposition of the timbre space into a hierarchy of spaces, each of which defines sub-classes of related timbres (see figure 2). Separate mapping functions are then defined for each sub-class (see the section below and appendix A for a description of these mappings). The actual implementation of the ISEE timbre parameters thus depends on the required refinement of synthesis control. The high level of abstraction of the ISEE timbre parameters allows them to be applied to different sub-classes in ways relevant to the behaviour of specific instruments within that sub-class. For example, the Envelope parameter controls the duration of the rise of the sound (attack speed) in the Sustaining instruments sub-classes, but the duration of the decay of the sound (decay speed) in the Decaying instruments sub-classes.

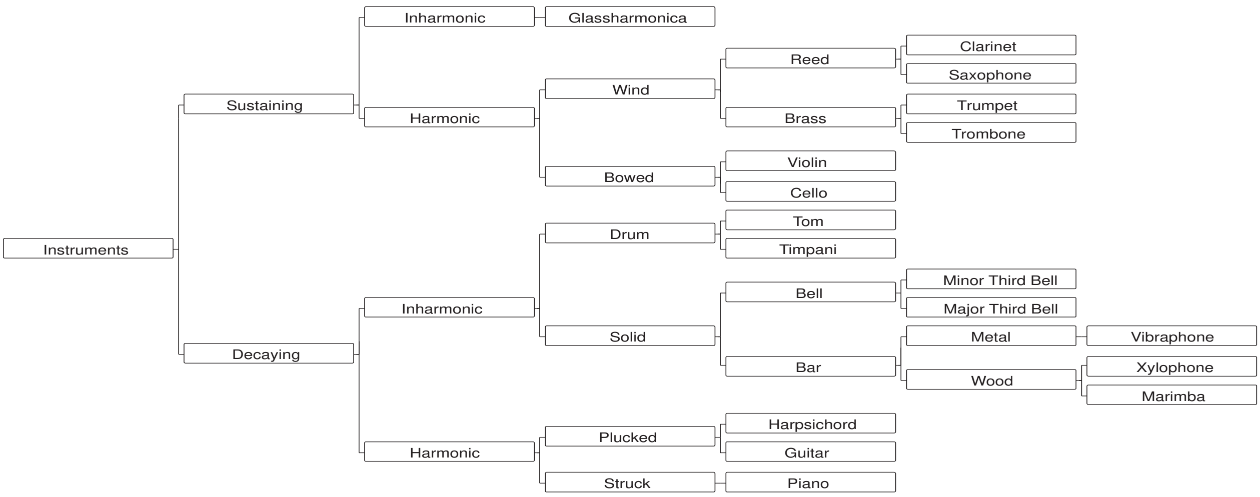

Figure 2: The instrument space hierarchy, based upon a perceptual taxonomy of instruments.

ISEE classes are called *instrument spaces*, because as well as allowing control of the timbre, each class also defines the range and type of pitch and loudness behaviour of the instrument(s) it contains. The instrument space hierarchy is based upon a categorisation scheme derived from expert analysis of existing instruments using think-aloud protocols, card sorting and interview techniques [26]. The first criterion used to structure a search is the envelope model (i.e., whether the sound sustains or decays). The next criterion is harmonicity of spectrum, after which transient behaviour (the timbral development at the start of a sound) becomes important. Further classification roughly follows the traditional Sachs-Hornbostel model, where instruments are classified into families according to characteristics of the vibrating source [13]. Each instrument space in the hierarchy is associated with a position of the four timbre parameters in its parent. By pressing the *Zoom In* button, the user can select the sub-space nearest to the current location of the four timbre parameters. This way, the properties of the four timbre parameters are used to structure a hierarchical search. The user can zoom in on refined instrument spaces from grosser higher level spaces, selecting constraints for each of the four timbre parameters in the process. Alternatively, when interested in a broader perspective of instruments, the user can jump to the parent instrument space by pressing the *Zoom Out* button. Users can also make use of a more declarative representation in the form of a traditional hierarchy browser, for example, when constructing new instrument spaces.

Instrument Space Organization

At an abstract level, the parameters of each component instrument space are consistently mapped according to the following heuristics: Overtones from harmonic (string-like sounds where individual overtones cannot be distinguished) to inharmonic (bell-like sounds where individual overtones can be heard), Brightness from dull to bright, Articulation from mellow to harsh, and Envelope from fast to slow. The underlying implementation of synthesis parameters, however, depends on the synthesis method used and on the level of refinement. In principle, any synthesis method can be used to define any instrument space. The only constraints are the capabilities of the synthesis hardware and the appropriateness of a synthesis method given a particular class of instruments. The specific parameter mappings for different synthesis models are beyond the scope of this paper (see Appendix A for an example). Instead, we will give an overview of the lower-level semantics of instrument space definitions at different levels in the hierarchy. Only timbre parameters of which the definition differs from that of the superclass are described.

At the root of the hierarchy, the *Instruments* space contains a crude characteristic of the sub-spaces it encloses. This space is defined by a mapping of the Overtones from harmonic to inharmonic, Brightness from dull to bright and Envelope from short decay to sustain. The mapping of the Articulation depends on the setting of the Envelope parameter. When it is set to decay, the Articulation controls the type of material of which the mallet is made with which the instrument is hit (the "knock" at the start of the sound), from mellow to hard. When it is set to sustain, the Articulation maps from a mellow brass-like transient to a harsh bowed-like transient. In order to limit the depth of the hierarchy and reduce redundancy in timbral manipulation, the next two levels in the hierarchy can be combined. This way, Envelope position and Overtones position in the *Instruments* space can be used to decide which of the four aggregate spaces (*Sustaining Inharmonic, Sustaining Harmonic, Decaying Inharmonic* and *Decaying Harmonic*) in the next layer will be selected when the user zooms in. In the *Sustaining Harmonic* space, the definition of the Overtones parameter is similar to that of the harmonic part of the *Instruments* space with enhanced resolution, and the Envelope parameter controls the duration of the rise of the sound. In this space, Articulation position maps from a mellow brass-like transient to a harsh bowed-like transient, and can be used to decide which of the two spaces (*Wind* or *Bowed*) in the next layer will be selected when the user zooms in.

From this level of classification instrument spaces are organized according to the Sachs-Hornbostel model. In order to be able to differentiate between the instruments in one family, however, pitch becomes an important criterion, which does not fit well with our model. As a compromise, Overtones is used in the *Bowed* space to control the range of the pitch, with a resolution of octaves. At the final level, instrument control properties become predominant, and timbre control is now very refined. In the *Violin* space, the Overtones parameter describes the position of the bow on the violin, from the fingerboard (sul tasto) to the bridge (sul ponticello). The Brightness parameter relates to the bow pressure on the string, the Articulation parameter relates to the force with which the bow is dropped on the strings while the Envelope parameter still controls the duration of the rise of the sound. In the decaying branch of the hierarchy, mapping strategies similar to those in the sustaining branch are found. As a brief example of the mappings used in this branch, the *Decaying Inharmonic* space uses a definition of the Overtones parameter similar to that of the inharmonic part of the *Instruments* space with enhanced resolution. The Articulation relates to the material of which the mallet is made, while the Envelope controls the decay time only. As with the *Sustaining Harmonic* space, the Overtones is used to select which subspace to zoom in to.

Computer Music Controllers

Given the Computer Music origins of the work reviewed in the previous sections, one would expect to see practical implementations of timbre space interfaces in the form of generic timbre control devices. However, this is not the case. Studies into real-time computer music controllers have traditionally focused on skilled performance rather than generic sound synthesis. To illustrate some of the problems that arise from this approach, a selection of typical articles on real-time control of digital sound synthesis from recent years is treated below.

Cadoz, Luciani et al. [4, 5] describe a musical virtual reality system that is based on two forms of instrumental models for digital sound synthesis:

- Input devices that capture physical gestures and react to these gestures with programmable feedback;
- Sound synthesis techniques based on the simulation of physical sound producing mechanisms (physical modelling).

At the time this was a revolutionary idea, integrating the development of physical modelling as a synthesis model with the idea of reactive input devices. However, the input devices that were developed for this system were designed to physically emulate traditional musical instrument behaviour. Traditional instruments typically provide enormous control potential at a considerable cost of training time. With their performance, the idiosyncrasy of traditional input devices is modelled as well. This means different input devices are needed to play different virtual instruments. Though it is claimed that this approach is viable for use in real-time sound synthesis control, it is typically designed for skilled performance, rather than generic user interface utilization. Gibet, Florens and Marteau [9, 10] base their gestural control system on motor system theory. Their approach too follows the physical modelling paradigm. With this approach, they intend to achieve direct manipulation of sound by restoring the causal link as the natural law for sound synthesis. This relies on the theory that the objects of the perception emerge out of the representation of gestures that produce the sound. Though it is clear that a direct correlation between gesture and sound reduces cognitive processing load and enhances performance [17], the expectations of a performer are related to real world objects. This impairs utilization of the system as a generic sound synthesis control paradigm, since a generalized mapping between gesture and timbre is not provided. Another computer music control system is *GAMS* [1]. This system uses ultrasonic sound to determine the position of up to four *wands* in 3-D space. The system requires the definition of a substantial amount of relations between on-stage positions and music, lighting and imaging control. A formalism for a meaningful mapping of control information to the various media is however not discussed. Not surprisingly, the audience could not understand what was happening during trial performances. Consequently, the idiosyncrasies of the system, rather than the contents of the performance, became the point of discussion.

The above survey indicates that problems of human-synthesizer interfacing in the field of Computer Music have been tackled primarily through the development of innovative hardware controllers. However, the use of these as generic controllers is limited, because researchers often fail to develop accompanying formalisms for mapping low-dimensional controller data to the high-dimensional parameter space of the generative sound synthesis algorithms. Also, many of these systems are intended to be idiosyncratic for artistic reasons and their usability is hardly ever empirically evaluated. In fact, the tradition in music is for the user to adapt to meet the requirements of the interface, rather than the other way round, as is the tradition in HCI. Musicians achieve direct manipulation of sound through musical instrument interfaces only after years of practice, possibly involving the development of physical deformities appropriate to the interface! Adaptation of music controllers for generic use, rather than skilled use, is therefore problematic. Generic interfaces for sound synthesis require widely available generic input devices, rather than specific input devices (i.e., musical instruments), and should have greater ease of use, by virtue of a limited number of degrees of freedom and lower training requirements. Though largely ignored in Computer Music, generic input devices have been extensively studied in HCI. The following sections therefore build upon existing work in the field of HCI to address the problem of selection and use of low cost input devices for generic utilization in sound synthesis.

Materials and Methods

We selected three input devices to empirically establish their impact on performance in a four degrees of freedom (DOF) instrument space navigation task: the Apple Standard Mouse (a relative input device); the Gravis Advanced MouseStick II—an optical joystick (absolute or relative); the Nintendo Power Glove (absolute and relative). Our sample population consisted of music students from the Department of Music of the University of Huddersfield, England, with experience in the use of electronic instruments and synthesized sounds, but with marginal experience in sound synthesis. A repeated measures design was used with a group of 15 paid subjects who were asked to reach for target positions in the Sustaining instrument space using the various device types. The *Sustaining* space contained a broad selection of sustaining musical instruments generated in real-time by simple FM synthesis on a synthesizer (see appendix A for a detailed description).

The Overtones parameter was used to control the harmonicity of the timbre, the Brightness parameter was used to control the amount of high-frequency energy of the timbre, The Articulation parameter controlled the ratio between the rise time of the higher harmonics and that of the lower harmonics and the Envelope parameter controlled overall rise duration. An Apple Macintosh SE was used to filter the erratic Power Glove information and record all movement during the experiments. An Apple Macintosh LC was used to run the ISEE system. The Mac LC was placed on a 70 cm high desk with the screen elevated 20 cm from the desk. A standard office chair without arm rests was placed 80 cm away from the screen for use during mouse and joystick experimentation. The seat of the chair was 51 cm above floor level. The back was set at a 90 degrees angle from the seat. During Power Glove experimentation, the chair was removed. All systems were interconnected by MIDI (a hard- and software protocol which constitutes a musical local area network).

Four interfaces were constructed. In the first, the mouse was used to change the coordinate indicators in the Control Monitor (see figure 1) by clicking and dragging the indicator dots (each 4 mm^2 in size, moving along 4.5 cm long axes). The control to display (C:D) ratio of the mouse was equivalent to position 4 in the Apple System 7 Mouse Control Panel, which is not linear. The joystick was used in the second and third interfaces. In the second interface the joystick provided absolute control—the position of the stick corresponded directly to the position of the indicators in the Control Monitor. When measured on top of the stick, the absolute joystick had a C:D ratio of 2:1. In the third interface, the joystick provided relative control—the position of the stick controlled the speed and direction of the Control Monitor indicators. With the stick almost upright, the indicators would move 1 parameter step at a time, while with the stick pushed towards a 60 degree angle the indicators would move up to 14 parameter steps at a time (1 parameter step corresponded to one pixel on the screen, each parameter having 128 possible positions). In both joystick interfaces, the two buttons on the top of the stick were used to select the coordinate system to be controlled with the stick. The fourth interface used the Power Glove for four-dimensional positioning in the Control Monitor. Motion on the y axis controlled Overtones, the x axis controlled Brightness, the z axis controlled Articulation, and roll information (a axis) was used to control the Envelope parameter in a relative fashion. Holding the wrist level would produce no change, rolling the wrist anti-clockwise would decrease the Envelope parameter and rolling clockwise would increase the Envelope parameter. The glove was engaged by clutching and inactive when not clutching. The glove had a control space of approximately 18 m^3. The C:D ratio of the glove ranged from 60:1 for the x and z axes to 40:1 for the y axis. Although we are aware that the large difference in C:D ratio between the mouse and the glove might confound the experiment, we feel this is an inherent property of the selected glove. The Power Glove is simply not capable of performing at the same resolution as a mouse.

The subjects were given five minutes to get used to each device, except for the Power Glove, with which they were allowed to practice for 15 minutes because of the special technique involved. Each subject also performed four preparatory trials to make sure they felt comfortable with each device in order to prevent learning effects. Each subject was given 10 test blocks of four experiments, one for each of the four device types. To prevent order effects, the order of the 4 types of input devices in each test block as well as the order of the test blocks was randomised. A questionnaire was answered by each subject after the experiments.

In each experiment the subject was required to listen to a timbre and locate it in the instrument space, using one of the four interfaces. At the start of each experiment, the location of the target could be seen in the Control Monitor window while the target sound was played five times. After this, the Control Monitor indicators would centre, with the sound changing accordingly, giving the subjects an audio-visual cue to start manipulating the indicators with the input device. The target position remained visible throughout the experiment in a separate window similar to that of Control Monitor. The stimulus tones (with a 1.5 sec. duration and a C3 pitch) were repeated throughout the experiment to give the subjects sufficient auditory feedback on the position of the Control Monitor indicators. When the match was considered good enough, the subjects released the input device. All movement during the experiments was recorded at millisecond accuracy using a MIDI sequencer. This enabled us to simulate retroactively an experiment where the subject would have been required to reach a certain accuracy criterion, which would then automatically terminate the trial.

Control Experiment

A control experiment was carried out to establish the effect of the graphical representation on movement with the glove. In this experiment, the glove was used to perform the outlined task in two conditions: with the screen on and with the screen off. In the first condition, the users had audio-visual feedback, in the latter, the subjects were required to concentrate on auditory feedback only. This experiment was also used to establish whether there was a learning effect for the glove.

Analysis of Movement Time and Accuracy

The efficacy of each device was established by measuring the movement time needed to reach the four-dimensional target position within a certain accuracy (where accuracy is overall Euclidean distance to target in four-dimensional space). This combines speed and accuracy into a single measure and removes the effect of individual subjects' subjective accuracy criteria for terminating trials. Since a subject might briefly, inadvertently pass through a point that lies within the required accuracy, retroactive analysis allows us to correct this by measuring the time until the subject passed the criterion for the last time during the trial [14]. The accuracy criterion was set to 1.13 cm in 4-D Euclidean distance to target, which was the 75th percentile of the final accuracies achieved over all trials in this experiment by the least accurate device, the Power Glove. The choice of the 75th percentile is not critical; analysis with other criteria gave similar results.

Though it is usual to present error rate as a measure for the accuracy of a device in a certain task, our retroactive analysis method allowed a clearer measure: the mean accuracy. The mean accuracy of a device is the smallest 4-D Euclidean distance to target reached on average during the trials with that device.

Trajectory Analysis

Garner [8] showed that certain parameters of a task are perceived as being integrally related to one another (the user sees these as a unified whole), while others are separably related (the user sees these as a collection of separate entities). Consequently, users manipulate certain parameters simultaneously (such as the x and y

position of a graphical object) while others are manipulated separably (e.g., the colour and size of a rectangle). These relationships between attributes constitute the perceptual structure of a task. The control structure of input devices shows similar characteristics, depending on whether it is natural to move diagonally across the different degrees of freedom of the device [2]. For optimal performance, it is important that the control structure of the device correlates with the perceptual structure of the task [14]. When we use an input device which provides input of a number of integral values (e.g., the 3-space positioning system of a dataglove or Polhemus) to a certain task, we can predict the perceptual structure of that task by analysing the trajectories of movement of the device [15]. Diagonal line patterns in the data will reveal integral movement since such movement will cut across dimensions in Euclidean space, while staircase patterns will reveal separable movement since such movement will be parallel to the axes of space.

We analysed the movement patterns of each device by calculating the control integration (*CI*) of movement: a ratio scale indicating the amount of diagonal movement between two positions in a 2-D space, in degrees. Maximum integration of movement in any direction results in a maximum *CI* of 45 degrees, minimum integration in any direction results in a minimum *CI* of 0 degrees. The control integration of movement on any pair of axes (*x, y*) at any given moment in time is a function of the first order differential of that movement:

$$CI = \left|\left|\left|\tan^{-1}\left(\frac{dy}{dx}\right)\right| - \frac{1}{4}\pi\right| - \frac{1}{4}\pi\right| \qquad (1)$$

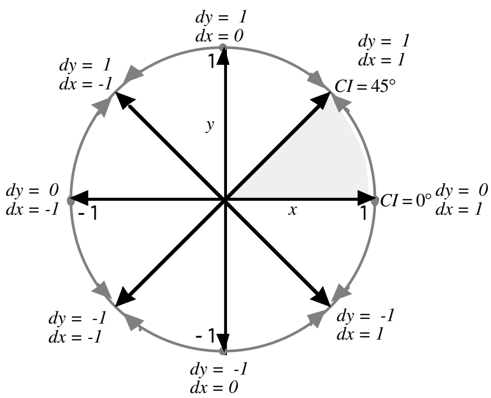

Figure 3: A graphic representation of equation (1)

Essentially, this function maps the slope of a line connecting two 2-D positions (shown in fig. 3 as vectors in a unit circle) to an angle between 0° and 45° (shown in fig. 3 as the shaded area). The function does this by folding (i.e. taking the absolute value) and shifting (i.e. subtracting 1/4 π) the slices of the pie shown in figure 3 so that all grey arrows align with the grey arrow in the shaded slice of the pie. In order to investigate the perceptual structure of our four-dimensional timbre manipulation task, the control integration was calculated for the two axes of each of the six possible projections at every sampling interval. An average $\overline{CI}$ was then calculated for each pair by adding the individual samples (CI_k) and dividing that by the number of samples *n*:

$$\overline{CI} = \frac{\sum_{k=1}^{n} CI_k}{n} \qquad (2)$$

Since $\overline{CI}$ also depends on the direction of the vector from the starting point to the target position in a trial, trials with vectors in all directions should be performed in order to get an accurate absolute indication of the control integration. However, we only used the control integration analysis data in a relative fashion. This reduced the need for such a great diversity of target vectors considerably.

Results

Analysis of variance showed that the choice of input device had a highly significant effect on performance ($F(3, 483) = 68.99$, $p < 0.001$). This indicated that differences in performance were related to the choice of input device and not just due to differences between subjects. Table 1 shows the mean movement time and mean overall accuracy for each device.

Device:	Mouse	Absolute Joystick	Relative Joystick	Power Glove
MT:	4,917	7,139	10,308	24,950
Accuracy:	1.35	1.61	2.04	8.97

Table 1: Mean movement time (msec) and accuracy (mm)

All differences were highly significant. The mouse was 1.5 times faster than the absolute joystick (paired two-tailed t-test; $p < 0.0001$), 2.1 times faster than the relative joystick ($p < 0.0001$) and 5.1 times faster than the Power Glove ($p < 0.0001$). The absolute joystick was 1.4 times faster than the relative joystick ($p < 0.0001$) and 3.5 times faster than the Power Glove ($p < 0.0001$). The relative joystick was 2.4 times faster than the Power Glove ($p < 0.0001$).

The difference in accuracy between the mouse and the absolute joystick was not significant (paired two-tailed t-test; $p > 0.15$). The mouse was 1.5 times more accurate than the relative joystick ($p < 0.005$) and 6.7 times more accurate than the Power Glove ($p < 0.0001$). The absolute joystick was 1.3 times more accurate than the relative joystick ($p < 0.05$) and 5.6 times more accurate than the Power Glove ($p < 0.0001$). The relative joystick was 4.4 times more accurate than the Power Glove ($p < 0.0001$).

Trajectory Analysis Results

Table 2 shows the mean $\overline{CI}$ per input device for each cross section of four-dimensional space.

Mean $\overline{CI}$ of axes:	Mouse	Absolute Joystick	Relative Joystick	Power Glove
$x \bullet y$	14	14.1	9.5	21.6
$a \bullet z$	12.1	13.7	8.4	3.7
$z \bullet y$	0	0.5	0.8	15.9
$z \bullet x$	0	0.5	0.7	16.4
$a \bullet y$	0	0.5	0.8	2.7
$a \bullet x$	0	0.6	0.7	2.9

(x=Brightness; y=Overtones; a=Envelope; z=Articulation)

Table 2: Mean average control integration (degrees)

Generally, movement with the Power Glove was better integrated than with the low-dimensional devices. Movement within the $x \cdot y$ and $a \cdot z$ planes was better integrated with the absolute joystick than with the relative joystick. All differences in control integration between input devices were highly significant (paired two-tailed t-test; $p < 0.01$), except for the difference in $x \cdot y$ control integration between the mouse and the absolute joystick ($p > 0.4$) and the difference in $a \cdot x$ control integration between the absolute joystick and the relative joystick ($p > 0.1$).

The Power Glove showed better integration in the $x \cdot y$ plane than in any of the other planes. The differences in control integration between the Power Glove cross sections $x \cdot y$ and $z \cdot y$ and between its $x \cdot y$ and $z \cdot x$ cross sections were highly significant ($p < 0.0001$). The difference between the $z \cdot y$ and $z \cdot x$ cross sections of the Power Glove was not significant ($p > 0.1$). The differences in control integration between the Power Glove cross sections $z \cdot x$, $z \cdot y$ and the $x \cdot y$ cross sections of the absolute joystick and mouse were highly significant ($p < 0.01$).

Control Experiment

The differences in movement time and control integration between the auditory-only condition and audio-visual feedback condition were not significant (*MT:* $p > 0.07$; *CI:* $p > 0.12$). The difference in accuracy between the two conditions was highly significant ($p < 0.0001$) (see table 3).

Screen:	Off	On
Accuracy:	15.66	9.22

Table 3: Mean accuracy (mm) in the control experiment.

Given the same conditions, the movement time of the glove in the control experiment was not significantly different from that in the first experiment ($p > 0.24$).

Qualitative Observations

Of the subjects, 73% used the mouse on a weekly or a daily basis, while 67% had little or no experience with the joystick and 93% had never used a dataglove before. Subjects found the Power Glove physically tiring and very hard to control. Initial experience with ISEE in composition and performance (compositions were made by the Hungarian composer Tamas Ungvary and a concert was given in Vienna) suggests that the timbre space approach to direct manipulation of timbre has validity. This view was reinforced by the subjects, who found ISEE a useful tool which liberated them from technicalities without restricting their artistic freedom. They regarded auditory feedback, particularly Overtones and Brightness, to be very useful as a navigational aid, but this cannot simply be generalized since they were trained listeners. The use of perceptual parameters in the interface was appreciated.

Discussion

The choice of input device significantly affected the movement time, accuracy and control integration with which the task was performed. The low-cost 2-D devices clearly outperformed the low-cost multidimensional Power Glove, most likely because of their superior technical refinement. The inferior performance of the Power Glove was most probably due to the severe impact cost-cutting measures have had on its signal/noise ratio. This affects both the resolution of the device and the amount of lag due to filtering. The observed movement time and accuracy deficiencies of the glove are best explained by the combinatory effect of lag and poor resolution on the index of difficulty of this task [20].

During the control experiment, subjects performed less accurate without visual feedback. Though this might well have been caused by the fact that the auditory feedback had a somewhat lower resolution than the visual feedback (particularly on the Overtones dimension), it might also be explained by the fact that the subjects needed to memorize their auditory target when the screen was off. However, the most probable cause for the observed accuracy deficiencies is that the subjects were simply less able to determine their exact position in the auditory-only feedback condition. Movement time did not improve between the first trials and the control experiment, which suggests there was no significant learning effect for the glove during experimentation. It therefore seems unlikely that the relative unfamiliarity of the subjects with this device confounded the results.

Control Integration

The low integration of Envelope with the other parameters during Power Glove experimentation was caused by the difference in control mode between roll and the other degrees of freedom of the glove. The current low resolution of roll impairs the glove's usability for any refined absolute utilization beyond 3 degrees of freedom. However, the glove's high amount of integration in both 2-D *and* 3-D space demonstrates the potential of such devices. The relatively low control integration of the relative joystick on the related axis pairs ($x \bullet y$, $a \bullet z$) indicates it is more separable than its absolute counterpart. This is probably due to the self-centring system of the device, which makes it strenuous to push the stick diagonally. The non-zero control integration of the joystick on the non-related axis pairs ($z \bullet y$, $z \bullet x$, $a \bullet y$, $a \bullet x$) indicates it facilitated switching between parameter sets. Since the relative joystick had no nulling problem [2], switching was slightly more frequent than with its absolute counterpart. The separability of the 2 x 2-D visual representation of the 4-D control space did not affect multidimensional control integration significantly, and corresponded well with the control structure of the 2-D devices. The high amount of integration of 3-space movement of the Power Glove suggests at least 3 of the 4 timbre parameters to be well integrated. Control integration shows the Overtones and Brightness parameters to be better integrated than other combinations of parameters such as Envelope & Articulation, Overtones & Articulation and Brightness & Articulation. This suggests Overtones and Brightness to be perceptually closer related.

Conclusions

Consistency of auditory feedback has long been disregarded in audio-related applications of direct manipulation graphical user interfaces. It is, however, also important to scrutinize the appropriateness of new input device technologies, since performance in direct manipulation systems depends heavily on that of the input device used. In Computer Music, new input device technologies are largely developed for refined simultaneous control of a large number of parameters, typically more than 6. An additional requirement for these devices seems to be that they need to look good on stage. For generic applications, however, such considerations are less important. The Power Glove, for instance, clearly impaired the usability of the ISEE system, most likely because of the lag and resolution deficiencies caused by its erratic positioning system. During regular sound synthesis applications the mouse remains the most efficient input device. When a synthesizer keyboard is used, however, the joystick is a good option. The relative joystick is best suited for gradually changing one parameter at a time and for rapidly switching parameter sets. The absolute joystick is better when speed, accuracy and navigational confidence are important. The Overtones and Brightness parameters were considered to be the most intuitive and useful parameters. Subjects were able to integrate the use of these two parameters effectively. Our visual representation of 4-D timbre space matches the control structure of the best performing devices.

The ideal timbre manipulation device?

What one considers to be ideal largely depends on whether the device is used for skilled performance or generic utilization. And although it is possible to create an inventory of important properties in the case of skilled musical performance, these properties can not be seen separate from their musical function [28]. However, as far as generic utilization is concerned, our research shows that generic 2-D devices perform well in a 4-D timbre manipulation task. We must stress that the poor performance of the Power Glove does not generalize to other, more refined tracking devices. Indeed, the high overall control integration of the glove shows the potential of such devices. However, there is evidence to suggest that in skilled musical performance, force feedback plays an important role [28]. This might reduce the usability of freely moving high-definition tracking devices in such applications. As for generic utilization, such devices are currently not an option, if only because of their high expense. Although our findings are not conclusive in this respect, it must also be noted that the separation of 4-D timbre space into 2 2-D projections seems a natural one, matching the perceived integration of the parameters. Therefore, we feel that any 2-D device which conforms well with Fitt’s Law is sufficient for generic timbre manipulation applications of ISEE. For example, if ISEE is to be integrated in the design of synthesizer user interfaces, two keyboard-mounted trackpads or touch screens might provide interesting alternatives to the joystick. We must always keep in mind, however, that we cannot expect the musician to use these as a musical instrument.

Acknowledgements

We would like to thank Apple Computer Inc. and S. Joy Mountford for supporting the above research. We would also like to thank Ernst Bonis for his early conceptual work on ISEE and Michael Clarke, Kurt Schmucker, Tom Wesley, Tamas Ungvary and Gerrit van der Veer for their support.

References

1. Bauer, W. and Foss, B. GAMS: An Integrated Media Controller System. *Computer Music Journal* 16(1), 1992, pp. 19-24.
2. Buxton, W. There's More to Interaction than Meets the Eye: Some Issues in Manual Input. In Norman, D.A. and Draper, S.W. (Ed.), *User Centered System Design: New Perspectives on HCI.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1986, pp. 319-337.
3. Buxton, W., Patel, S., Reeves, W., and Baecker, R. Objed and the Design of Timbral Resources. *Computer Music Journal* 6(2), 1982.
4. Cadoz, C., Luciani, A., and Florence, J. Responsive Input Devices and Sound Synthesis by Simulation of Instrumental Mechanisms: The Cordis System. *Computer Music Journal* 8(3), 1984.
5. Cadoz, C., Luciani, A., and Florens, J.L. CORDIS-ANIMA: A Modeling and Simulation System for Sound and Image Synthesis—The General Formalism. *Computer Music Journal* 17(1), 1993, pp. 19-29.
6. Chowning, J. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society* 21(7), 1973, pp. 526-534.
7. Eaglestone, B. and Verschoor, A. Dichtslaande deuren en mens-machine interfaces. *Kennissystemen* 5(5), 1991, pp. 17-21.
8. Garner, W.R. *The Processing of Information and Structure*. Lawrence Erlbaum, Potomac, Md., 1974.
9. Gibet, S. and Florens, J.-L. Instrumental gesture modeling by identification with time-varying mechanical models. In *Proceedings of the 1988 ICMC.* Cologne, Germany: International Computer Music Association, 1988, pp. 28-40.
10. Gibet, S. and Marteau, P.-F. Gestural Control of Sound Synthesis. In *Proceedings of the 1990 ICMC.* Glasgow: International Computer Music Association, 1990, pp. 387-391.
11. Goble, C. Multimedia Databases. Report CSTR-93-04. University of Southampton, 1993.
12. Grey, J.M. An Exploration of Musical Timbre. Ph.D. Dissertation. Dept. of Psychology, Stanford University, 1975.
13. Hornbostel, E.M. von and Sachs, C. Systematik der Musikinstrumente: ein Versuch. *Zeitschrift fuer Ethnologie* H. 4-5. Berlin, Germany, 1914.
14. Jacob, R.J.K. and Sibert, L.E. The Perceptual Structure of Multidimensional Input Device Selection. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, 1992, pp. 211-218.
15. Jacob, R.J.K., Sibert, L.E., McFarlane, D.C., and M.P. Mullen, J. Integrality and Separability of Input Devices. *ACM Transactions on Computer-Human Interaction* 1(1), 1994.
16. Jaslowitz, M., D'Silva, T., and Zwaneveld, E. Sound Genie - An Automated Digital Sound Effects Library System. *SMTE Journal* , May 1990, pp. 386-391.
17. Keele, S.W. *Attention and Human Performance*. Pacific Pallisades: Goodyear Publishing Company, 1973.
18. Lee, M., Freed, A., and Wessel, D. Real-Time Neural Network Processing of Gestural and Acoustical Signals. In *Proceedings of the 1991 ICMC.* Montreal: International Computer Music Association, 1991, pp. 277-280.
19. Lee, M. and Wessel, D. Connectionist Models for Real-Time Control of Synthesis and Compositional Algorithms. In *Proceedings of the 1992 ICMC.* San Jose: International Computer Music Association, 1992, pp. 277-280.
20. MacKenzie, I.S. and Ware, C. Lag as a Determinant of Human Performance in Interactive Systems. In *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems.* Amsterdam, Holland, 1993, pp. 488-493.
21. Plomp, R. *Aspects of Tone Sensation*. London: Academic Press, 1976.
22. Rutkowski, C. An introduction to the Human Applications Standard Computer Interface, Part 1: Theory and principles. *BYTE* 7(11), 1982, pp. 291-310.
23. Shepard, R. Representations of Structure in Similar Data: Problems and Prospects. *Psychometrica* 39, 1974, pp. 373-421.

24. Shneiderman, B. *Designing the User-Interface: Strategies for Effective Human-Computer Interaction*. Reading, MA: Addison Wesley, 1987.

25. Truax, B. Organizational Techniques for c:m Ratios in Frequency Modulation. *Computer Music Journal* 1(4), 1977, pp. 39-45.

26. Vertegaal, R. *Music Technology Dissertation*. Utrecht School of the Arts, The Netherlands, 1992.

27. Vertegaal, R. and Bonis, E. ISEE: An Intuitive Sound Editing Environment. *Computer Music Journal* 18(2), 1994, pp. 21-29.

28. Vertegaal, R. and Ungvary, T. The Sentograph: Input Devices and the Communication of Bodily Expression. In *Proceedings of the 1995 ICMC*. Banff, Canada: International Computer Music Association, 1995.

29. Wessel, D. *Report to C.M.E.* University of California, San Diego, 1974.

30. Wessel, D. Timbre Space as a Musical Control Structure. In Roads, C. and Strawn, J. (Ed.), *Foundations of Computer Music*. Cambridge, MA: MIT Press, 1985, pp. 640-657.

Appendix A: Instrument Space Definition

The synthesis model used to generate the stimulus tones during the experiments was simple FM synthesis, where one sine-wave oscillator (the carrier) is modulated in frequency by another sine-wave oscillator (the modulator). The synthesis platform was a YAMAHA SY99 synthesizer. The timbre produced by this synthesizer was controlled via MIDI system exclusive messages, while stimulus tones were generated using MIDI note-on messages. The *Sustaining* instrument space used in the experiments only contained non-decaying sounds. The four parameters of this instrument space were defined as follows: the Overtones parameter controlled the harmonicity of the produced timbre by setting the ratio between the frequency of the carrier (c) and the modulator (m). The c:m ratios were defined as follows: 1:1 (harmonic timbre), 2:1, 3:1, 4:1, 5:1 (nasal timbre), 1:2 (hollow timbre), 1:4, 1:3, 1:5, 4:5, 6:5, 1:9,1:11, 1:14, 2:3, 3:4, 2:5, 2:7, 2:9 (inharmonic timbre) (see [25] for a more detailed explanation). The Brightness parameter was used to control the amount of high frequency energy in the timbre by modifying the cutoff frequency of the low-pass filter. A low cutoff frequency would thus produce a dull timbre, while a high cutoff frequency would produce a bright timbre. The Articulation parameter controlled the ratio of the higher harmonics' attack rate to the lower harmonics' attack rate. A low Articulation would produce a brass-like attack, where the lower harmonics would rise first, while a high Articulation would produce a string-like attack, where all harmonics would rise simultaneously. The Envelope parameter controlled the rise duration of the sound. A low setting would make the sound rise quickly, while a high setting would make the sound rise slowly. These mappings were designed by an expert to approach as consistent a perceptual mapping as possible with simple FM.