Chapter 9
A New Interaction Strategy for Musical Timbre Design

Allan Seago

Abstract Sound creation and editing in hardware and software synthesizers presents usability problems and a challenge for HCI research. Synthesis parameters vary considerably in their degree of usability, and musical timbre itself is a complex and multidimensional attribute of sound. This chapter presents a user-driven search-based interaction style where the user engages directly with sound rather than with a mediating interface layer. Where the parameters of a given sound synthesis method do not readily map to perceptible sonic attributes, the search algorithm offers an alternative means of timbre specification and control. However, it is argued here that the method has wider relevance for interaction design in search domains which are generally well-ordered and understood, but whose parameters do not afford a useful or intuitive means of search.

9.1 Introduction

Much recent research in music HCI has concerned itself with the tools available for real time control of electronically generated or processed sound in a musical performance. However, the user interface for so-called 'fixed synthesis' – that part of the interface which allows the design and programming of sound objects from the ground up (Pressing 1992) – has not been studied to the same extent. In spite of the migration that the industry has seen from hardware to software over the last 20 years, the user interface of the typical synthesizer is, in many respects, little changed since the 1980s and presents a number of usability issues. Its informed use typically requires an in-depth understanding of the internal architecture of the instrument and of the methods used to represent and generate sound. This chapter

A. Seago (✉)
Sir John Cass Faculty of Art, Architecture and Design, London Metropolitan University, London, UK
e-mail: a.seago@londonmet.ac.uk

S. Holland et al. (eds.), *Music and Human-Computer Interaction*, Springer Series on Cultural Computing, DOI 10.1007/978-1-4471-2990-5_9,

proposes a system which essentially removes the mediating synthesis controller layer, and recasts the problem of creating and editing sound as one of search. The search strategy presented here has wider HCI applications, in that it provides a user-driven method of exploring a well-ordered search space whose axes may either be unknown, or else not connected uniquely to any one parameter. This will be considered further in the conclusion.

We begin, however, by summarising usability problems associated with sound synthesis methods and with the building of intuitive controllers for timbre. Approaches to the problem which have used visual representations of sound or sought to bridge the semantic gap between sound and language are briefly reviewed. The discussion goes on to consider timbre space and its use in sound synthesis, and then describes the operation of the weighted centroid localization (WCL) search method in three contrasting timbre spaces. The chapter concludes by considering the extent to which the WCL algorithm could be applied in other application domains, and discusses relevant issues arising from the building and testing of the system. Finally, directions for future work are proposed.

9.2 Synthesis Methods

Synthesis methods themselves present varying degrees of difficulty to the uninformed user. Some, like subtractive synthesis, offer controllers which are broadly intuitive, in that changes to the parameter values produce a proportional and predictable change in the generated sound. Other methods, however, are less easily understood. FM synthesis, for example, is a synthesis method that may be viewed as essentially an exploration of a mathematical expression, but whose parameters have little to do with real-world sound production mechanisms, or with perceived attributes of sound. However, all synthesis methods require a significant degree of understanding of the approach being employed, and therefore present usability problems for the naïve user, to a greater or lesser extent. The mapping of the task language familiar to musicians – a language which draws on a vocabulary of colour, texture, and emotion – is not easily mapped to the low-level core language of any synthesis method. In other words, the design of intuitive controllers for synthesis is difficult because of the complex nature of musical timbre.

9.3 Timbre

The process of creating and editing of a synthesized sound typically involves incremental adjustments to its various sonic attributes – pitch, loudness, timbre, and the way that these evolve and change with respect to time. Regardless of architecture or method of synthesis, the last of these three attributes – timbre – presents the most intractable usability issues.

The difficulties attached to the understanding of timbre have been summarised in a number of studies; notably by Krumhansl (1989) and Hajda et al. (1997). It has variously been defined as the 'quality' or 'character' of a musical instrument (Pratt and Doak 1976), or that which conveys the identity of the originating instrument (Butler 1992). However, most recent studies of timbre take as their starting point the ANSI standards definition in which timbre is stated as being "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" – that is to say, timbre is what is left, once the acoustical attributes relating to pitch and loudness are accounted for. This definition, of course, raises the question of how timbral differences are to be defined in isolation from loudness and pitch when these qualities are not dissimilar.

Timbre has been traditionally presented as an aspect of sound quite distinct from and orthogonal to pitch and loudness. This 'three axis' model of musical sound is reflected in the design of commercial subtractive synthesizers, where the user is provided with 'handles' to these three nominal attributes in the form of voltage-controlled oscillators (for pitch), amplifiers (for loudness) and filters (for timbre). However, it has long been understood that timbre is a perceptual phenomenon which cannot be simply located along one axis of a three dimensional continuum. Instead, it arises from a complex interplay of a wide variety of acoustic elements, and is itself multidimensional; to a great extent, it subsumes the uni-dimensional vectors of pitch and loudness (which map, more or less linearly, to frequency and amplitude respectively).

9.4 Sound Synthesis Using Visual Representations of Sound

A number of sound synthesis systems offer a user interface in which the user engages with a visual representation of sound in either the time or frequency domain. A good example of such a system is *Metasynth*, by U and I Software. However, audio-visual mapping for sound visualisation presents problems (Giannakis 2006). It is difficult to make an intuitive association between the waveform and the sound it generates – the information is simply at too low a level of abstraction. No user is able to specify finely the waveform of imagined sounds in general, either in the time or frequency domains. In other words, there is no 'semantic directness' (Hutchins et al. 1986) for the purpose of specifying any but the most crudely characterized sounds (Seago et al. 2004).

9.5 Sound Synthesis Using Language

Much research has been focussed on the design of systems whose interfaces connect language with synthesis parameters (Ashley 1986; Ethington and Punch 1994; Vertegaal and Bonis 1994; Miranda 1995, 1998; Rolland and Pachet 1996; Martins

et al. 2004). Many of these systems draw on AI techniques and encode rules and heuristics for synthesis in a knowledge base. Such systems are based on explicitly encoded rules and heuristics which relate to synthesis expertise ('bright sounds have significant energy in the upper regions of the frequency spectrum', 'a whole number modulator/carrier frequency relationship will generate a harmonic sound'), or to the mapping of specific acoustic attributes with the adjectives and adverbs used to describe sound.

While the idea of presenting the user with an interface which mediates between the parameters of synthesis and a musical and perceptual vocabulary is an attractive one, there are a number of problems. There is a complex and non-linear relationship between a timbre space and a verbal space. The mapping of the sound space formed by a sound's acoustical attributes to the verbal space formed by semantic scaling is, as has been noted (Kendall and Carterette 1991), almost certainly not linear, and many different mappings and sub-set spaces may be possible for sounds whose envelopes are impulsive (e.g., xylophone) or non-impulsive (e.g., bowed violin). There are also questions of the cross-cultural validity and common understanding of descriptors (Kendall and Carterette 1993). Most studies of this type make use of English language descriptors, and issues of cultural specificity are inevitably raised by studies of this type where the vocabulary used is in a language other than English (Faure et al. 1996; Moravec and Stepánek 2003). Similarly, it has been found that the choice of descriptors for a given sound is likely to vary according to listener constituency – whether they are keyboard players or wind players, for example (Moravec and Stepánek 2003). Apparently similar semantic scales may not actually be regarded by listeners as similar (Kendall and Carterette 1991); it is by no means self-evident, that, for example, soothing-exciting is semantically identical with calm-restless, or would be regarded as such by most subjects.

9.6 Timbre Space

One approach to timbre study has been to construct timbre spaces: coordinate spaces whose axes correspond to well-ordered, perceptually salient sonic attributes. Timbre spaces can take two forms. The sounds that inhabit them can be presented as points whose distances from each other either reflect and arise from similarity/dissimilarity judgments made in listening tests (Risset and Wessel 1999). Alternatively, the space may be the *a priori* arbitrary choice of the analyst, where the distances between points reflect calculated (as distinct from perceptual) differences derived from, for example, spectral analysis (Plomp 1976).

More recent studies have made use of multidimensional scaling to derive the axes of a timbre space empirically from data gained from listening tests. That such spaces are firstly, stable and secondly, can have predictive as well as descriptive power has been demonstrated (Krumhansl 1989), and this makes such spaces interesting for the purposes of simple synthesis. For example, hybrid sounds derived from combinations of two or more instrumental sounds were found to occupy positions

in an MDS solution which were located between those of the instruments which they comprised. Similarly, exchanging acoustical features of sounds located in an MDS spatial solution can cause those sounds to trade places in a new MDS solution (Grey and Gordon 1978). Of particular interest is the suggestion that timbre can be 'transposed' in a manner which, historically, has been a common compositional technique applied to pitch (Ehresman and Wessel 1978; McAdams and Cunible 1992).

9.7 Timbre Space in Sound Synthesis

If timbre space is a useful model for the analysis of musical timbre, to what extent is it also useful for its synthesis? Here, we summarise and propose a set of criteria for an ideal *n*-dimensional attribute space which functions usefully as a tool for synthesis.

- It should have good coverage – that is to say, it should be large enough to encompass a wide and musically useful variety of sounds.
- It should have sufficient resolution and precision.
- It should provide a description of, or a mapping to a sound sufficiently complete to facilitate its re-synthesis.
- The axes should be orthogonal – a change to one parameter should not, of itself, cause a change to any other.
- It should reflect psychoacoustic reality. The perceived timbral difference of two sounds in the space should be broadly proportional to the Euclidean distance between them.
- It should have predictive power. A sound C which is placed between two sounds A and B should be perceived as a hybrid of those sounds.

The first of these criteria – that the number of timbre space dimensions needs to be high – poses clear computational problems. Some studies have sought to address this by proposing data reduction solutions (Sandell and Martens 1995; Hourdin et al. 1997a; Nicol 2005). Other researchers have sought to bridge the gap between attribute/perceptual space and parameter space by employing techniques drawn from artificial intelligence.

9.8 Search Algorithms

The position taken in this chapter is that the problem can usefully be re-cast as one of search – in which a target sound, located in a well-ordered timbre space, is arrived at by a user-directed search algorithm.

A number of such algorithms already exist (Takala et al. 1993; Johnson 1999; Dahlstedt 2001; Mandelis 2001; Mandelis and Husbands 2006). Such methods

typically use interactive genetic algorithms (IGAs). The features of IGAs are reviewed by McDermott (2013) in this volume; for our purposes, the main drawback of IGAs is the so-called 'bottleneck'; genetic algorithms take many generations to converge on a solution, and human evaluation of each individual in the population is inevitably slower than in systems where the determination of fitness is automated (Takagi 2001).

We present here another form of search algorithm, called weighted centroid localization (WCL), based, not on the procedures of breeding, mutation and selection characteristic of GAs, but on the iterative updating of a probability table. As with interactive GAs, the process is driven by user selection of candidate sounds; the system iteratively presents a number of candidate sounds, one of which is then selected by the user. However, in this approach, a single candidate solution (rather than a population) is generated; over the course of the interaction, this series of choices drives a search algorithm which gradually converges on a solution.

9.9 Weighted Centroid Localization (WCL)

The search strategy employs an adapted *weighted centroid localisation* (WCL) algorithm, which is used to drive the convergence of a candidate search solution on to a 'best-fit' solution, based on user input. The technique has been shown to be an effective search method for locating individual sensors within wireless sensor networks (Blumenthal et al. 2007).

The structure and function of the system is summarised here before considering it in greater detail. A target sound T is identified or imagined by a user. The target T, which is assumed to be fixed, is unknown to the system. An n-dimensional attribute space is constructed which is assumed to have the ability to generate the target sound T, and in which will be created a number of iteratively generated probe sounds. In addition, the system holds an n-dimensional table P, such that, for each point s in the attribute space, there is a corresponding element p in the probability table. The value of any element p represents the probability, at any given moment, that the corresponding point s is the target sound, based on information obtained so far from the user.

At each step of the user/system dialog, the user is asked to listen to a number of system generated probes, and to judge which of the probes most closely resembles the target sound T. The user's judgement on the probes is used by the system to generate a new candidate sound C, whose coordinates correspond, at generation time, to those of the weighted centroid of the probability table. Two versions of this search strategy were tested; the first of these, referred to here as the WCL-2 strategy, presented two probes to subjects. The second, the WCL-7 strategy, presented seven. We begin by considering the WCL-2 strategy.

For the purposes of a formal test of the system, three sounds, initially chosen randomly from the space, are presented to the subject – a target sound T and two probes A and B. On each iteration of the algorithm, the subject is asked to judge

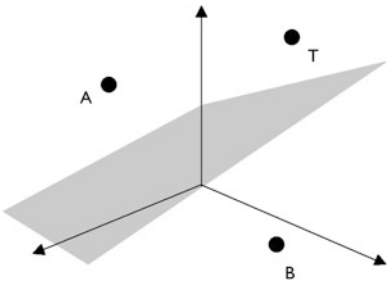

Fig. 9.1 Bisection of probability table P

which of the two probes *A* or *B* more closely resembles *T*. If *A* has been chosen, the values of all cells in *P* whose Euclidean distance from *B* is greater than their distance from *A* are multiplied by a factor of $\sqrt{2}$; the values of all other cells are multiplied by a factor of $1/\sqrt{2}$. Similarly, if *B* has been chosen, the values of all cells in *P* whose Euclidean distance from *A* is greater than their distance from *B* are multiplied by a factor of $\sqrt{2}$; the values of all other cells are multiplied by a factor of $1/\sqrt{2}$. Thus, on each iteration (in the case of a three dimensional space), the space *P* is effectively bisected by a plane which is perpendicular to the line *AB* (see Fig. 9.1). The probability space *P* having been updated, two new probes A_{new} and B_{new} are generated, and the process repeated.

As *P* is progressively updated, its weighted centroid *C* starts to shift. If all, or most, of the subject responses are correct (i.e. the subject correctly identifies which of *A* or *B* is closer to *T*), the position of *C* progressively approaches that of *T*.

The WCL-7 strategy works slightly differently. A target sound *T* and seven probes *A* . . . *G*, initially chosen randomly from the space, are presented to the subject. On each iteration of the algorithm, the subject is asked to judge which of the seven probes *A*–*G* more closely resembles *T*. For each cell in the probability table *P*, establish its Euclidean distance *d* from the cell corresponding to the selected probe, and multiply its value by 100/*d*. In effect, the value of a cell increases in inverse proportion to its distance from the selected probe. The weighted centroid *C* is recalculated, and a new set of probes *A* . . . *G* generated.

In both cases, the search strategy is user driven; thus, the subject determines when the goal has been achieved. At any point, the subject is able, by clicking on the 'Listen to candidate' button, to audition the sound in the attribute space corresponding to the weighted centroid *C*; the interaction ends when the subject judges *C* and *T* to be indistinguishable. In operational use, T might be a sample or an imagined sound.

In order to have a baseline against which to assess the success of the WCL strategy, another program was developed which provided the user with the means of manually navigating the attribute space. The user interface afforded direct access to the attribute space via individual sliders which control navigation along each axis. This is a form of multidimensional line search (MLS). It has the virtue of simplicity; indeed, for a space of low dimensionality, it may be the most effective search method. However, a successful interaction using this technique is entirely

dependent on the ability of the user to hear the individual parameters being modified and, crucially, to understand the aural effect of changing any one of them.

9.10 The Timbre Spaces

Having considered the algorithm itself, the three timbre spaces constructed for the purposes of testing will be considered.

The first timbre space, referred to here as the *formant space*, is fully described in Seago et al. (2005) and is summarized here. It was inhabited by sounds of exactly 2 s in duration, with attack and decay times of 0.4 s. Their spectra contained 73 harmonics of a fundamental frequency (*F0*) of 110 Hz, each having three prominent formants, *I*, *II* and *III*. The formant peaks were all of the same amplitude relative to the unboosted part of the spectrum (20 dB) and bandwidth ($Q = 6$). The centre frequency of the first formant, I, for a given sound stimulus, was one of a number of frequencies between 110 and 440 Hz; that of the second formant, *II*, was one of a number of frequencies between 550 and 2,200 Hz, and that of the third, *III*, was one of a number of frequencies between 2,200 and 6,600 Hz. Each sound could thus be located in a three dimensional space.

The second space, referred to here as the *SCG-EHA space*, was derived from one studied by Caclin et al. (2005). The dimensions of the space are rise time, spectral centre of gravity (SCG) and attenuation of even harmonics relative to the odd ones (EHA). The rise time ranged from 0.01 to 0.2 s in 11 logarithmic steps. In all cases, the attack envelope was linear. The spectral centre of gravity (SCG), or spectral centroid is defined as the amplitude-weighted mean frequency of the energy spectrum. The SCG varied in 15 linear steps between 3 and 8 in harmonic rank units – that is to say, between 933 and 2,488 Hz. Finally, the EHA – the attenuation of even harmonics relative to odd harmonics – ranged from 0 (no attenuation) to 10 dB, and could take 11 different values, separated by equal steps. Again, the sounds used in the space were synthetically generated pitched tones with a fundamental of 311 Hz (E4), containing 20 harmonics.

These two spaces were chosen because a mapping between perceptual and Euclidean distances in the space could be demonstrated; in the case of the formant space, this was shown in Seago (2009) and in that of the SCG-EHA space, in Caclin et al. (2005). The construction of the last of the three spaces – the MDS space – is more complex, and for this reason will be discussed here in greater detail. Based on a space constructed by Hourdin et al. (1997b), it was generated through multi-dimensional scaling analysis of a set of instrumental timbres – e.g. alto flute, bass clarinet, viola etc. Multidimensional scaling (MDS) is a set of techniques for uncovering and exploring the hidden structure of relationships between a number of objects of interest (Kruskal 1964; Kruskal and Wish 1978). The input to MDS is typically a matrix of 'proximities' between such a set of objects. These may be actual proximities (such as the distances between cities) or may represent people's similarity-dissimilarity judgments acquired through a structured survey

or exposure to a set of paired stimuli. The output is a geometric configuration of points, each representing a single object in the set, such that their disposition in the space, typically in a two or three dimensional space, approximates their proximity relationships. The axes of such a space can then be inspected to ascertain the nature of the variables underlying these judgments.

However, the technique can also be used as a means of data reduction, allowing the proximity relationships in a multidimensional set of data to be represented using fewer dimensions – notably, for our purposes, by Hourdin et al. (1997a). As already mentioned, both the space and the construction technique used to build it are derived in part from this study, and the list of 15 instrumental timbres is broadly the same. The pitch of all the instrumental sounds was, again, Eb above middle C (311 Hz). Each instrumental sample was edited to remove the onset and decay transients, leaving only the steady state portion, which was, in all cases, 0.3 s.

For the purposes of our test, a multi-dimensional timbre space was constructed. One of its dimensions was attack time, with the same characteristics as those of the SCG-EHA space described in the previous paragraph (i.e. ranging from 0.01 to 0.2 s). The remaining dimensions were derived by MDS techniques, and an initial MDS analysis was performed to determine the minimum number of dimensions required to represent the audio samples with minimum loss of information. This resulted in a seven-dimensional timbre space.

This preparatory process is described fully in Seago (2009); we will summarise it here. The audio files were spliced together and processed using heterodyne filter analysis. Heterodyne filtering resolves periodic or quasi-periodic signals into component harmonics, given an initial fundamental frequency; a fuller account of the technique is given in Beauchamp (1969) and Moorer (1973). After a process of editing, the output was a 15 row by 20 column matrix of data in which each row held the Long Time Averaged Spectrum (LTAS) for one instrumental sound, and the columns contained the amplitude values of each harmonic. This matrix was used to build a dissimilarity matrix which was in turn input to a classical multidimensional scaling function. This generated two outputs: a solution space to the input dissimilarity matrix, and the eigenvalues of each axis of the solution space. The magnitude of each eigenvalue is an indicator of the amount of information associated with that axis. Inspection revealed that 95% of the total information required to reconstruct the spectra was associated with just six axes; thus, MDS can be used to reduce the dimensionality from 20 to 6 with minimal loss of information. A new six-dimensional MDS space was thus generated from the dissimilarity matrix. The following scatter graph (Fig. 9.2) shows the 15 instrumental sounds placed in a three dimensional space (the first three columns of the reduced space dataset).

Sounds represented in the reduced space can be auditioned by means of a data recovery process. The harmonic amplitudes of a given sound can be dynamically generated from a single six-coordinate point in the space and input to an additive synthesis process for playback.

Having considered the three spaces in which the search strategies were tested, we turn now to consider the testing procedure itself.

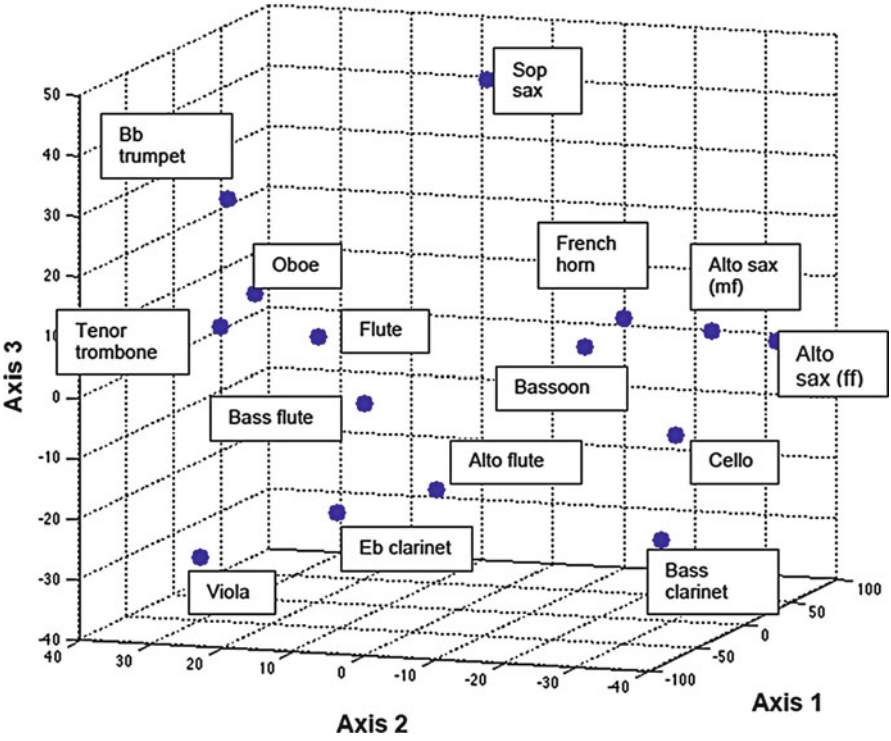

Fig. 9.2 The 15 instrumental sounds located in a three dimensional space following MDS analysis

9.11 Strategy Testing Procedure

The strategy was tested on a number of subjects – 15 in the case of the formant and SCG-EHA space tests, and 20 for the MDS space tests (which were conducted later). The purpose of the test was explained, and each subject given a few minutes to practise operating the interfaces and to become accustomed to the sounds. The order in which the tests were run varied randomly for each subject. Tests were conducted using headphones; in all cases, subjects were able to audition all sounds as many times as they wished before making a decision.

For the multidimensional line search test, each subject was asked to manipulate the three software sliders, listening to the generated sound each time until EITHER the 'Play sound' button had been clicked on 16 times OR a slider setting was found for which the generated sound was judged to be indistinguishable from the target. For the WCL-2 and WCL-7 tests, each subject was asked to listen to the target and then judge which of two sounds A or B (in the case of the WCL-2 strategy) or of seven sounds (in the case of the WCL-7 strategy) more closely resembled it. After making the selection by clicking on the appropriate button, new probe sounds were

generated by the software, and the process repeated until EITHER 16 iterations had been completed OR the sound generated by clicking on the 'Candidate' button was judged to be indistinguishable from the target.

Finally, in order to determine whether the strategy was, in fact, operating in response to user input and was not simply generating spurious results, the WCL-2 and WCL-7 strategies was run with a simulation of user input, but where the 'user response' was entirely random.

9.12 Results

Figures 9.3 and 9.4 summarise the mean WCL-2 and WCL-7 weighted centroid trajectories, averaged for all 15 interactions, in the formant and SCG-EHA attribute spaces respectively; in each case, they are compared with the trajectory in the respective spaces of the sound generated by the user on each iteration of the multidimensional line search strategy – again, averaged for all 15 interactions.

In all three strategies deployed in the two attribute spaces, there was considerable variation in individual subject performance. However, the mean trajectories of both the WCL-2 and WCL-7 strategies show a greater gradient (faster convergence on the target) than that of the MLS strategy, with the WCL-7 trajectory being, in both cases, the steepest. It is noticeable that the MLS trajectories, both individually and taken as an average, show no significant convergence in the formant space, but do converge in the SCG-EHA space, suggesting that the parameters of the SCG-EHA space are more perceptually salient (and therefore useful) than those of the formant space.

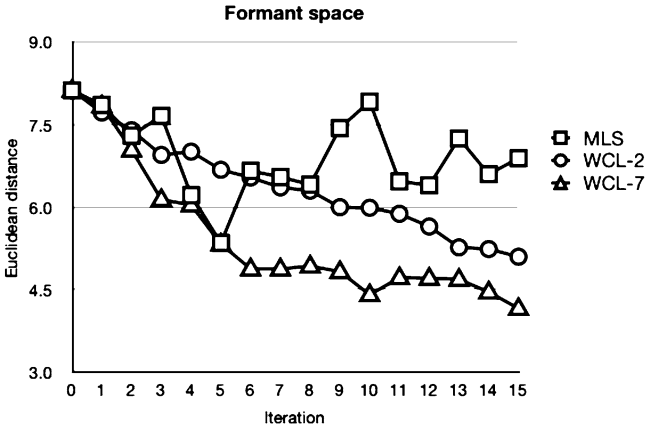

Fig. 9.3 Mean weighted centroid trajectory in formant space for MLS, WCL-2 and WCL-7 strategies

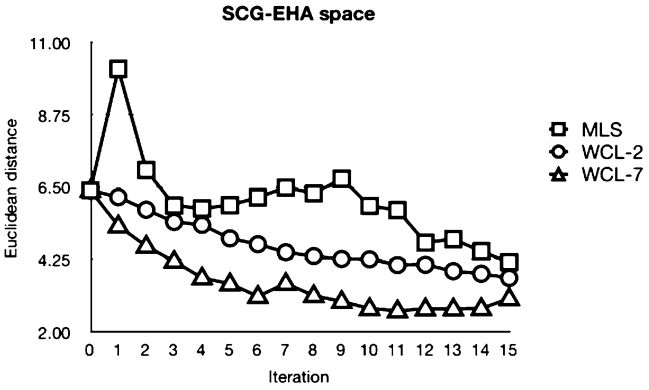

Fig. 9.4 Mean weighted centroid trajectory in SCG-EHA space for MLS, WCL-2 and WCL-7 strategies

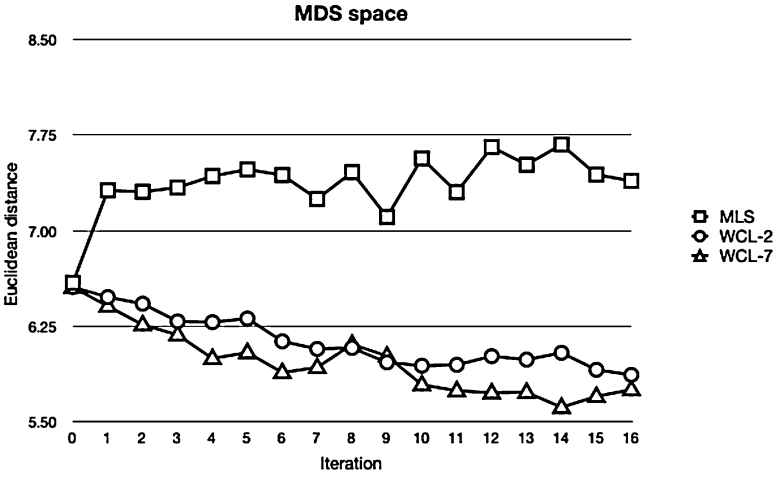

Fig. 9.5 Mean weighted centroid trajectory in MDS space for MLS, WCL-2 and WCL-7 strategies

We now consider the results in the MDS space. Figure 9.5 shows the averaged trajectories for the multidimensional line search, WCL-2 and WCL-7 search strategies.

Overall, the line search method is not a satisfactory search strategy in this particular attribute space. Inspection of the individual trajectories showed only one example of a subject who was able to use the controls to converge on the target. By contrast, the averaged data from the WCL-2 and WCL-7 tests shows a clear and steady gradient, with that from the WCL-7 strategy showing a more rapid convergence.

Finally, the average trajectories of the 'control' strategy (in which 'random' user responses were given) showed no convergence at all.

9.13 Summary of Results

A summary of the results is presented in Fig. 9.6. In order to make possible direct comparison of the results from three attribute spaces that otherwise differed, both in their sizes and in their characteristics, the vertical axis represents the percentage of the Euclidean distance between the target and the initial position of the weighted centroid.

While it should be borne in mind that, in all cases, there was considerable variation in individual subject performance, the six mean weighted centroid trajectories from the WCL-2 and WCL-7 search strategies in the three spaces all show, to a greater or lesser extent, a convergence on the target. Two observations can be made from the above results. Firstly, the gradients of the two traces representing the weighted centroid mean trajectory in the seven-dimensional MDS space are considerably shallower than those in either of the two three-dimensional spaces. One probable reason for this is the greater difficulty of the task; a seven dimensional space is clearly more complex and difficult to navigate than a three dimensional one. Secondly, in each of the three attribute spaces, the WCL-7 strategy (in which subjects were asked to choose from seven probes) produced a swifter convergence (expressed as the number of subject iterations) on the target than the WCL-2

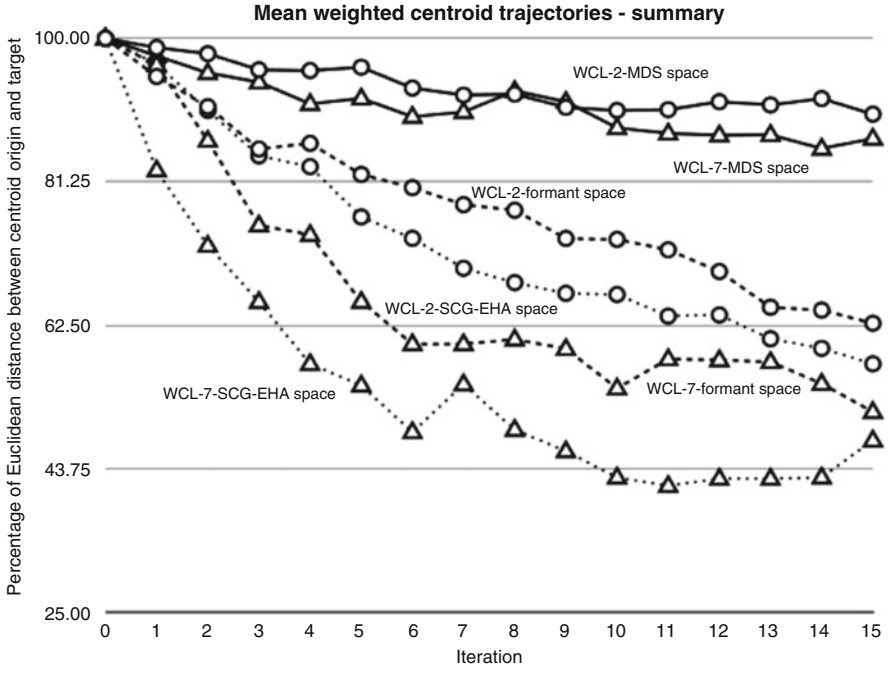

Fig. 9.6 Mean trajectories of weighted centroid for WCL-2 and WCL-7 strategies in three different timbre spaces

strategy, where only two probes were offered. This was observable in a number of individual subject performances, as well as in the overall graph, and is an interesting result. The task of critically evaluating seven, rather than two probes imposes on the subject a greater cognitive load and it had been speculated that this would result in a slower (or even zero) rate of convergence. It should be emphasised, however, that the metric used here is the number of iterations, not the elapsed time or the number of individual actions (i.e. mouse clicks) required to audition the seven probes. Several subjects reported that the WCL-7 task was more difficult than the WCL-2 task; and although this was not measured, it was noticeable that the time required by a number of subjects to complete the task was significantly greater in the case of the WCL-7 task than for either of the other two.

9.14 Conclusion

This paper has presented a discussion of usability in sound synthesis and timbre creation, and the problems inherent in current systems and approaches. Interactive genetic algorithms (IGAs) offer a promising means of exploring search spaces where there may be more than one solution. For a timbre search space which is more linear, however, and whose dimensions map more readily to acoustical attributes, it is more likely that there is (at best) only one optimum solution, and that the fitness contour of the space consists only of one peak. In this case, the WCL strategy offers a more direct method for converging on an optimum solution without the disruptive effects of mutation and crossover.

The WCL technique is proposed here in order to address the usability problems of other methods of sound synthesis. However, it can be generalised to a broader range of applications. As with IGAs, it can be applied to search problems where a fitness function (i.e. the extent to which it fulfils search criteria) is difficult to specify, and convergence on a best solution can only be achieved by iterative user-driven responses based on preference. Examples of such application domains are the visual arts and creative design. This concluding section considers issues arising from this research, potential for further work and implications for HCI work in other non-musical domains.

An important feature of the WCL interaction described here is that it affords engagement with the sound itself, rather than with some visual representation of it. Earlier trial versions of the software (not discussed in this chapter) did, in fact, include a visual representation of the probability space, in which the cell values were colour-coded. The probability space was thus represented as a contour map, from which users were able to see areas of higher probability in the search space. Interestingly, during pilot testing, subjects found the visual element to be a distraction; they stopped listening critically to the sounds and instead, selected those probes which appeared to be closer to these ‘high probability’ areas. In this way,

a dialogue was established in which the software was driving user choices, rather than the other way round, and where convergence on the target was poor. This is not to say that a visual component should or could not be used, but does suggest that visual aids may not always be helpful in interactions of this kind.

The efficacy of the algorithm for sound synthesis, or indeed in any of these domains rests on two assumptions. In order to test the strategies, a target sound was provided for the subjects, whereas the ultimate purpose, as previously noted, is to provide a user interaction which converges on a target which is imaginary. The assumption is, firstly, that the imagined sound actually exists in the space and can be reached; and secondly, that it is stable – the user's imagined sound does not change. Consideration of the efficacy of these, or any other search strategies when the goal is a moving target – that is to say, the user changes his/her mind about the target – is outside the scope of this study, but should be nevertheless noted; it is the nature of interactive search that some searches prove not to be fruitful, or the target changes during the course of the interaction (because the user has changed his/her mind). This being the case, the interface might incorporate a 'backtrack' feature, by use of which previous iterations could be revisited, and new choices made.

Future directions of research are briefly outlined here. First of all, the sounds inhabiting all three spaces are spectrally and dynamically invariant (although the SCG-EHA and MDS spaces include a variable attack time dimension). Clearly, for the strategy to be a useful tool for timbral shaping, this will need to be addressed.

Neither the WCL process itself, nor the spaces in which it was tested took account of the non-linearity of human hearing. That the sensitivity of the hearing mechanism varies with frequency is well known; it would be of interest to establish whether the WCL search strategy performed significantly better in such spaces which incorporated a perceptual model which reflected this.

The number of iterations required to achieve a significant degree of convergence with the target is unacceptably high. Essentially, this is the 'bottleneck' problem, characteristic of interactive GAs. Convergence on the target might be significantly accelerated if the user, instead of being offered two or more probes for consideration, is provided with a slider which offers sounds which are graduated interpolations between two points in the space, or alternatively a two dimensional slider which interpolate between four points. Very much the same technique is described in McDermott et al. (2007) as a means of selection; what is proposed here is an adapted version of it, in which the sliders are dynamically attached to a vector which joins two probes in the MDS space whose positions are updated on each iteration. A two dimensional slider could be similarly used for a vector which joined three probes. Another direction which could prove fruitful is to provide the user with the means of rating two or more probes for perceived similarity to the target (rather than simply selecting one). The probability space could then be given an additional weighting based on the relative rating of the probes, which in turn might result in a swifter convergence of the weighted centroid on the target.

References

Ashley, R. (1986). A knowledge-based approach to assistance in timbral design. In *Proceedings of the 1986 international computer music conference*, The Hague, Netherlands.

Beauchamp, J. (1969). A computer system for time-variant harmonic analysis and synthesis of musical tones. In H. von Foerster & J. W. Beauchamp (Eds.), *Music by computers*. New York: Wiley.

Blumenthal, J., Grossmann, R., Golatowski, F., & Timmermann, D. (2007). Weighted centroid localization in Zigbee-based sensor networks. WISP 2007. In *IEEE international symposium on intelligent signal processing*, Madrid, Spain.

Butler, D. (1992). *The musician's guide to perception and cognition*. New York: Schirmer Books.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America, 118*(1), 471–482.

Dahlstedt, P. (2001). *Creating and exploring huge parameter spaces: Interactive evolution as a tool for sound generation proceedings of the 2001 international computer music conference*. Havana: ICMA.

Ehresman, D., & Wessel, D. L. (1978). *Perception of timbral analogies*. Paris: IRCAM.

Ethington, R., & Punch, B. (1994). SeaWave: A system for musical timbre description. *Computer Music Journal, 18*(1), 30–39.

Faure, A., McAdams, S., & Nosulenko, V. (1996). Verbal correlates of perceptual dimensions of timbre. In *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC4),* McGill University, Montreal, Canada.

Giannakis, K. (2006). A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound, 11*(3), 297–307.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America, 63*(5), 1493–1500.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliège & J. Sloboda (Eds.), *The perception and cognition of music*. London: Psychology Press.

Hourdin, C., Charbonneau, G., & Moussa, T. (1997a). A multidimensional scaling analysis of musical instruments' time varying spectra. *Computer Music Journal, 21*(2), 40–55.

Hourdin, C., Charbonneau, G., & Moussa, T. (1997b). A sound synthesis technique based on multidimensional scaling of spectra. *Computer Music Journal, 21*(2), 40–55.

Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1986). Direct manipulation interfaces. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: new perspectives on human-computer interaction*. Hillsdale: Lawrence Erlbaum Associates.

Johnson, C.G. (1999). Exploring the sound-space of synthesis algorithms using interactive genetic algorithms. In *AISB'99 symposium on musical creativity*, Edinburgh.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception, 8*(4), 369–404.

Kendall, R., & Carterette, E. C. (1993). Identification and blend of timbres as basis for orchestration. *Contemporary Music Review, 9*, 51–67.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzen & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music*. Amsterdam: Elsevier (Excerpta Medica 846).

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Newbury Park: Sage Publications.

Mandelis, J. (2001). Genophone: An evolutionary approach to sound synthesis and performance. In E. Bilotta, E. R. Miranda, P. Pantano, & P. Todd (Eds.), *Proceedings of ALMMA 2002: Workshop on artificial life models for musical applications*. Cosenza: Editoriale Bios.

Mandelis, J., & Husbands, P. (2006). Genophone: Evolving sounds and integral performance parameter mappings. *International Journal on Artificial Intelligence Tools, 20*(10), 1–23.

Martins, J.M., Pereira, F.C., Miranda, E.R., & Cardoso, A. (2004) Enhancing sound design with conceptual blending of sound descriptors. In *Proceedings of the workshop on computational creativity (CC'04),* Madrid, Spain.

McAdams, S., & Cunible, J. C. (1992). Perception of timbral analogies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 336*(1278), 383–389.

McDermott, J. (2013). Evolutionary and generative music informs music HCI—and vice versa. In S. Holland, K. Wilkie, P. Mulholland, & A. Seago (Eds.), *Music and human-computer interaction* (pp. 223–240). London: Springer. ISBN 978-1-4471-2989-9.

McDermott, J., Griffith, N. J. L., & O'Neill, M. (2007). Evolutionary GUIs for sound synthesis. In *Applications of evolutionary computing*. Berlin/Heidelberg: Springer.

Miranda, E. R. (1995). An artificial intelligence approach to sound design. *Computer Music Journal, 19*(2), 59–75.

Miranda, E. R. (1998). Striking the right note with ARTIST: An AI-based synthesiser. In M. Chemillier & F. Pachet (Eds.), *Recherches et applications en informatique musicale*. Paris: Editions Hermes.

Moorer, J. A. (1973). *The heterodyne filter as a tool for analysis of transient waveforms*. Stanford: Stanford Artificial Intelligence Laboratory.

Moravec, O., & Stepánek, J. (2003). Verbal description of musical sound timbre in Czech language. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC'03),* Stockholm.

Nicol, C. A. (2005). *Development and exploration of a timbre space representation of audi*o. PhD thesis, Department of Computing Science. Glasgow: University of Glasgow.

Plomp, R. (1976). *Aspects of tone sensation*. New York: Academic.

Pratt, R. L., & Doak, P. E. (1976). A subjective rating scale for timbre. *Journal of Sound and Vibration, 45*(3), 317–328.

Pressing, J. (1992). *Synthesiser performance and real-time techniques*. Madison: A-R Editions.

Risset, J. C., & Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In D. Deutsch (Ed.), *The psychology of music*. San Diego: Academic.

Rolland, P.-Y., & Pachet, F. (1996). A framework for representing knowledge about synthesizer programming. *Computer Music Journal, 20*(3), 47–58.

Sandell, G., & Martens, W. (1995). Perceptual evaluation of principal components-based synthesis of musical timbres. *Journal of the Audio Engineering Society, 43*(12), 1013–1028.

Seago, A. (2009). A new user interface for musical timbre design. Ph.D thesis, Faculty of Mathematics, Computing and Technology, The Open University.

Seago, A., Holland, S., & Mulholland, P. (2004). *A critical analysis of synthesizer user interfaces for timbre*. HCI 2004: Design for Life, Leeds, British HCI Group.

Seago, A., Holland, S., & Mulholland, P. (2005). Towards a mapping of timbral space. In *Conference on Interdisciplinary Musicology (CIM05),* Montreal, Canada.

Takagi, H. (2001). Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE, 89*(9), 1275–1296.

Takala, T., Hahn, J., Gritz, L., Geigel, J., & Lee, J.W. (1993). Using physically-based models and genetic algorithms for functional composition of sound signals, synchronized to animated motion. In *Proceedings of the International Computer Conference (ICMC'93),* Tokyo, Japan.

Vertegaal, R., & Bonis, E. (1994). ISEE: An intuitive sound editing environment. *Computer Music Journal, 18*(2), 21–29.