Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

A Novel User Interface for Musical Timbre Design

ALLAN SEAGO[1], SIMON HOLLAND[2], and PAUL MULHOLLAND[3]

[1] London Metropolitan University, Sir John Cass Department of Art, Media and Design, E1 1LA, London, UK
a.seago@londonmet.ac.uk

[2] Open University, Centre for Computer-Human Interaction Research, Department of Computing, MK7 6AA, UK
s.holland@open.ac.uk

[3] Open University, Knowledge Media Institute, MK7 6AA, UK
p.mulholland@open.ac.uk

ABSTRACT

The complex and multidimensional nature of musical timbre is a problem for the design of intuitive interfaces for sound synthesis. A useful approach to the manipulation of timbre involves the creation and subsequent navigation or search of *n*-dimensional coordinate spaces or *timbre spaces*. A novel timbre space search strategy is proposed, based on weighted centroid localization (WCL). The methodology and results of user testing of two versions of this strategy in three distinctly different timbre spaces are presented and discussed. The paper concludes that this search strategy offers a useful means of locating a desired sound within a suitably configured timbre space.

1. INTRODUCTION

Current user interfaces for sound synthesis in both hardware and software synthesizers typically employ controllers – rotary dials, sliders and buttons - which map to the parameters of the sound synthesis method. The informed use of such an interface typically requires an in-depth understanding, both of the synthesis method used and of the internal architecture of the instrument. Devising a means of specifying, editing and controlling sound in which the directives are expressed in terms more familiar to a musician is difficult, largely because of the elusive and complex nature of musical timbre.

Most recent studies of timbre - that quality of sound which makes, for example, an oboe being played at the same pitch and degree of loudness as a clarinet

nevertheless sound different - take as their starting point the ANSI standards definition [1]. This states that timbre is "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 very subtractive definition leaves most of what we hear as timbre and timbral change unexplained.

Timbre is multidimensional – that is to say, it can be described on a wide range of semantic scales (*bright-dark*, *rough-smooth* etc), few of which, however, map in a clear and generally accepted way to any single acoustical physical attribute of sound or sound synthesis parameter. This being the case, the musician wishing to create or edit a sound is obliged to do so using the technical terminology of the particular synthesis method afforded by the electronic musical instrument being used. Thus, the design of a usable interface for the specification of timbre relies on a well-founded understanding of the relationship between timbre perception and its acoustic correlates.

One fruitful approach to the study of timbre has been to construct *timbre spaces*: coordinate spaces whose axes correspond to well-ordered, perceptually salient sonic attributes, and in which a sound can be located as a single point. This paper considers the usefulness of the timbre space model as a vehicle for sound synthesis, proposes and describes a novel timbre space search strategy for timbral shaping, in which iterative user input drives a software process where a candidate solution converges on the desired sound, and finally presents the results of user testing of this strategy in three contrasting timbre spaces.

2. BACKGROUND

A timbre space may be constructed from pre-determined acoustical attributes in such a way that distances between points in the space (i.e. sounds) reflect calculated differences derived, for example, from spectral analysis [2]. Alternatively, distances between points may reflect and arise from similarity/dissimilarity judgments made in listening tests [3]. This second category of timbre space is typically built from multidimensional scaling (MDS) analysis of data generated by such listening tests.

2.1. Multidimensional scaling (MDS)

Multidimensional scaling, or MDS, is a set of techniques used in (for example) psychology, sociology and anthropology, for uncovering and exploring the hidden structure of relationships between a number of objects of interest [4, 5]. The input to MDS is typically a matrix of 'proximities' between such a set of objects. These may be actual proximities (such as the geographical distances) or may represent subjects' 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. In timbre research, it has been used to identify those acoustic attributes which are salient to similarity-dissimilarity judgments of sound stimuli [6-13].

2.2. Timbre space as synthesis space

If a timbre space can be successfully used in the study of timbre and timbral relationships, can it also be used for sound synthesis? In order for a timbre space to function effectively as a synthesis space, it should be large enough to encompass a wide and musically useful variety of sounds, and be of sufficient resolution and precision; secondly, it should provide a description of, or a mapping to a sound which is complete enough to facilitate its re-synthesis. Thirdly, Euclidean distances between sounds in the space should be broadly proportional to the perceived timbral difference between them, such that a sound C which is placed between two sounds A and B should be perceived as a hybrid of A and B.

A number of studies have shown that simple timbre spaces can have predictive power of this sort. Exchanging the acoustical features of sounds located in an MDS spatial solution, for example, can cause those sounds to trade places in a new MDS solution [14]. Similarly, 'hybrid' timbres, constructed by combining the acoustical features of two traditional instruments (a 'trumpar' from a trumpet and a guitar, for example) will be located between the constituent instruments in an MDS space [9]. 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 [15, 16].

2.3. Timbre space as search space

The high number of dimensions necessary to fully represent timbral attributes however presents obvious computational problems. Some studies have addressed this by proposing data reduction solutions, using MDS or principal component analysis (PCA) techniques [17-19]. Other researchers have sought to bridge the gap between timbre perception and synthesis parameters by employing search techniques drawn from artificial intelligence, using knowledge based systems [20-24] or evolutionary search algorithms [25-27].

One particular type of evolutionary search is the genetic algorithm. This technique, in which an encoded population of possible search solutions is subjected to processes of mutation, crossover and selection, has been shown to be effective for synthesis methods such as frequency modulation, where the link between a synthesis parameter (e.g. modulation index) and an audible attribute of sound is complex and unpredictable [26, 28].

However, where the mapping of synthesis parameters to audible sonic attributes is fairly straightforward (that is to say, where relative distances between sounds in the synthesis space broadly reflect those in the perceptual space), a more direct method which converges onto an optimum solution without the disruptive effects of mutation and crossover is likely to be more successful. This is the rationale for the *weighted centroid localization* timbre space search strategy presented in this paper.

3. WEIGHTED CENTROID LOCALIZATION

We give a brief overview of the strategy here, deferring more detailed discussion of its operation until later.

Similarity-dissimilarity forced-choice listening tests typically present subjects with a pair of stimuli, and ask them to rate their degree of similarity on a numerical scale of values. It is argued here that, where such tests demonstrate a correspondence between Euclidean and perceptual distances in a given timbre space, a similar and complementary process can be used as a user-driven method for the localisation of a sound chosen from that space. In essence, the subject is presented at each stage of the interaction with a number of probes and asked to make a judgment as to which one of the probes most resembles an unchanging target sound. The subject's response updates a probability table, which, in turn, is used to generate a candidate solution. Over the course of the interaction, the candidate solution converges on the target.

The candidate solution is based on the *weighted centroid* of the probability table. The centroid of a surface or body is its centre of mass, assuming uniform density; the centroid of a set of points in a two-dimensional space is the arithmetic mean of all the coordinates of the space. However, if we include in the calculation of the centroid the values (or weights) of the matrix elements, we derive the weighted centroid. The coordinates i_C and j_C of the weighted centroid are then

$$i_C = \frac{\sum_{x=1}^{N} w_x i_x}{\sum_{x=1}^{N} w_x}, j_C = \frac{\sum_{x=1}^{N} w_x j_x}{\sum_{x=1}^{N} w_x} \quad (1)$$

where N is the number of points, i_x and j_x are the coordinates of the x th point in the space, and w_x is the weight of the x th point in the space.

The notion of a weighted centroid has a number of applications and has been shown to be effective in locating individual sensors within wireless sensor networks [29]. Such networks consist of a number of nodes, some of which can determine their own positions (beacons) and others (sensors) which cannot, and which calculate their own positions by a centroid determination from the positions of the beacons in range. The weighting depends on the distance and on the characteristics of the sensor node's receivers.

The search strategy described in this paper works in an analogous way. The weighting, however, works differently; it arises from a table of cells, each corresponding to one candidate timbre in the timbre space, and whose value reflects the probability that the corresponding timbre is the target.

Two versions of the strategy were tested: one in which two probes were used (referred to in this paper as WCL-2), and another in which seven probes were presented (WCL-7). The reason for the use of two versions of the strategy was to observe the effect on the interaction (if any) of varying the number of choices offered. Clearly, two probes is the minimum number of choices that can be offered. The other end of the scale was taken to be seven; the assumption was made in the work presented here that meaningful comparison by subjects of the timbral qualities of more than seven probes would be prohibitively difficult. Even with seven probes, there is a significant increase in the cognitive load on the subject. It is clear that making comparisons of two probes and a target is more straightforward and easier to accomplish than comparisons of seven probes and a target. The question which was addressed in this experiment was whether this would be a significant factor in the rate of convergence with the target.

Before considering the WCL strategies in greater detail, we consider first the timbre spaces in which they operated.

4. TIMBRE SPACES

Two contrasting three-dimensional timbre spaces were used as vehicles for the initial testing of the WCL strategy.

4.1. Formant space

The axes of this space were formant centre frequencies; the stimuli drawn from this space sound subjectively like a collection of more or less open and closed vowel sounds. Although we are not primarily concerned with vowels as such, a simple timbre space, loosely based on vowels, has been chosen as the first timbre space; firstly, because it is simple and easily synthesizable, and secondly, because the use of such a space will allow a relatively wide range of timbral variation in the set of sounds to be generated within an otherwise very circumscribed space. The simplicity of this space will allow us to conduct the first test of the search strategy in a relatively well-understood context.

To understand the motivation for the choice of a simple, low dimensional timbre space for the empirical work presented here, it is useful briefly to review the role of formants in timbre. A formant is a broad frequency region which causes an increase in amplitude of any spectral component partial falling within its range [30]. Slawson, and subsequently Plomp and Steeneken [31] demonstrated that perceived timbral similarities were more readily attributed to invariances in formant frequencies than to invariances in the overall spectral envelope. Formant terminology is more usually applied to the description of vocal systems; however, the frequency spectrum of a given instrumental sound will also have characteristic formants, which do not shift in frequency with changes in the frequency of the fundamental.

The sounds inhabiting this space were all exactly two seconds in duration, with attack and decay times of 0.4 seconds. 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 2200 Hz, and that of the third, III, was one of a number of frequencies between 2200 and 6600 Hz.

4.2. SCG-EHA space

The second of the two 3D spaces is derived from the work of Caclin *et al* [32]. The paper takes as a baseline the work of a number of researchers who have concluded from MDS studies that spectral centre of gravity (SCG) and attack time are salient acoustical correlates of timbre perception [8] [33] [34]. Other correlates that have been identified include spectral flux (variance of the spectrum over time) [33] and spectral irregularity [35]. Caclin *et al* performed a set of experiments to confirm these earlier findings, by constructing spaces of synthetic sounds that varied only by these parameters, and conducting dissimilarity rating listening tests on those sounds. The hypothesis was that, if these correlates were correct, there should be a good match between the physical space and the perceptual space.

Of the three different spaces generated in the study, the third, whose parameters were attack time, spectral centre of gravity (SCG) and attenuation of even

harmonics (EHA) relative to odd harmonics, provided a good match.

For our purposes, this space is a suitable vehicle for testing because firstly, a good mapping between physical and perceptual dimensions has been found in this particular space, and secondly, sounds in the space vary only by their attack time, SCG and EHA - this means that they are easily synthesizable and the search will not be disrupted by timbral variations (specificities) which are not accounted for by the three axes.

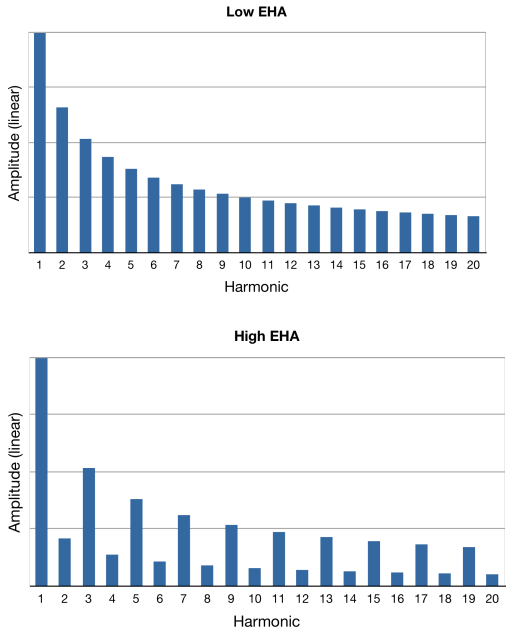

Figure 1: Spectra with low and high even harmonic attenuation.

4.2.1. Construction of the timbre space

A timbre space based on that used in Caclin *et al* was constructed. The sounds inhabiting this space were pitched with a fundamental of 311 Hz, and contained 20 harmonics. The dimensions were as follows:

- Rise time, ranging from 0.01 to 0.2 seconds in 11 logarithmic steps. (It has been noted that the logarithm of attack time appears to explain the corresponding timbre dimension better than the attack time itself [33, 35]). In all cases, the attack envelope was linear.
- EHA - attenuation of even harmonics relative to the odd ones in the range 0 dB to 10 dB – again in 11 steps. Figure 1 shows spectra with low and high EHA values respectively.
- SCG - spectral centroid. The SCG is defined here as the amplitude-weighted mean frequency of the energy spectrum. For all stimuli, the amplitude A_n of any harmonic n was calculated by

$$A_n = k \times 1/n^{\alpha} \tag{2}$$

where k is an arbitrary value and α a value determined by the SCG. This, in turn, is given by

$$SCG = \frac{\sum_n n \times A_n}{\sum_n A_n} \tag{3}$$

and is the harmonic rank number where the SCG is located. The range used here was from 3.000 to 8.000, in 15 linear steps; this corresponds to a spectral centroid range of 933 Hz to 2488 Hz.

All sounds were generated using Csound, and normalised to –3 dB relative to full amplitude using an audio editor.

While the space is broadly similar to that in Caclin *et al*, there are two small, but important differences, Firstly, the range covered on the SCG axis is wider than that in the Caclin *et al* study. This was in order to provide a greater degree of timbral variation than was apparent in that space. The range of EHA was also expanded for the same reason. The number of steps on these axes was also chosen to ensure detectable timbral difference between adjacent discrete steps on the axes.

Secondly: it was noted, when constructing the space, that a change in EHA, brought about by attenuation of even harmonics, also resulted in a small change in SCG. It could be argued that the axes of the space are not, for this reason, entirely orthogonal. It is not clear in Caclin *et al* how the authors dealt with this. In the present study, the amplitude reduction of even harmonics is accompanied by a compensating increase in the amplitudes of odd harmonics, thus preserving SCG while maintaining variation in EHA. We do not believe that these small modifications to

the space make significant alterations to its psychoacoustical properties.

5. SEARCH STRATEGIES

We return now to the discussion of the WCL-2, WCL-7 and MLS search strategies, beginning with MLS.

5.1. Multidimensional line search (MLS)

As noted above, this method provides the subject with sliders giving direct access to the axes of the space. Thus, adjusting the position of the first slider, for example, would have the effect of changing the first formant centre frequency (in the case of the formant space), and of changing the SCG of the sound (in the case of SCG-EHA space).

This strategy has the virtue of simplicity; indeed, for a space of low dimensionality, it may be the most effective. However, a successful interaction using this method is entirely dependent on the ability of the user to a) hear the individual parameters being modified and, crucially, to understand the aural effect of changing any one of them.

5.2. Weighted centroid localisation (WCL)

The WCL method is an iterated user/system dialog designed to steer a system-generated candidate sound *C* towards a goal, or target sound *T*, within the two timbre spaces described earlier, with the aim of minimising the Euclidean distance *CT*.

In general, an *n*-dimensional timbre space *S* is constructed (such as those described earlier in section 3 of this paper - thus *n*=3), which contains, at any time, a fixed target sound T and a number of iteratively generated probe sounds. In addition, we construct an *n*-dimensional table *P*, such that for each element $s_{i,j,k}$ in the timbre space (where *i*, *j* and *k* are the axis coordinates), there is a corresponding element $p_{i,j,k}$ in the probability table. The value of the element $p_{i,j,k}$ represents the probability, at any given moment, that the corresponding element $s_{i,j,k}$ is the target sound, based on information from the user.

On each step of the user/system dialog, the user is presented with the target sound *T* and a number of probes, and asked to judge which of the probes most closely resembles *T*. This information is used by the system to generate a new candidate sound *C,* whose coordinates are, at any time, those of the weighted centroid of the probability table.

As stated above, two versions of the WCL strategy were implemented and tested.

5.2.1. WCL-2 – two-alternative forced choice

In this version, three sounds, chosen randomly from the space, are presented to the subject - a target sound *T* and two probes *A* and *B*; their coordinates in the timbre space and probability table are as shown in figure 2.

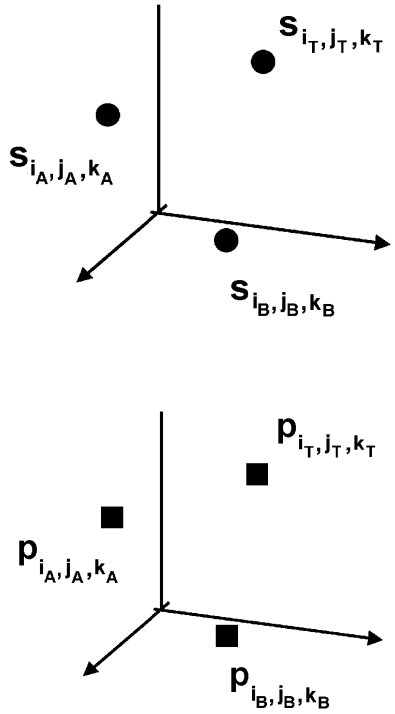

Figure 2: Timbre space S and probability table P.

On each iteration of the algorithm, the subject is asked to judge which of the two probes *A* or *B* more closely resembles *T*. The subject having made a choice, the following steps are executed:

- 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 √2; the values of all other cells are multiplied by a factor of 1/√2.

- 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 √2; the values of all other cells are multiplied by a factor of 1/√2. Thus, on each iteration, *P* is effectively bisected by a line which is perpendicular to the line *AB*.

- Recalculate the weighted centroid C.

The probability space *P* having been updated, two new probes A_{new} and B_{new} are generated, and the process repeated. The generation of coordinates for A_{new} and B_{new}, although pseudo-random, is nevertheless performed under the following constraints.

Firstly, A_{new} and B_{new} should be sufficiently far apart in the timbre space for there to be significant difference in their timbres. (Earlier pilot tests failed because subjects could not, in some cases, hear any real difference between the two, and therefore could not make a judgement about their degrees of respective similarity to *T*.) Secondly, the line connecting A_{new} and B_{new} should be more or less orthogonal to a line connecting *A* and *B*. This is to ensure that the information accumulating in the probability table *P* builds up along more than one dimension.

As *P* is progressively updated, its weighted centroid *C* starts to shift (at the outset, because all the cells in *P* have the same value, the centroid is located exactly at the centre of the space). 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*.

As already stated, the search strategy is user driven; thus the subject determines when the goal has been achieved. At any point, the subject was able to audition the sound in the timbre space corresponding to the weighted centroid *C*; the interaction ended when the subject judged *C* and *T* to be indistinguishable.

Data pertaining to the iteration was logged by the software – in particular, the successive positions of *A, B* and *C*, the degree to which the interaction has been successful is measured by the gradient of the approach of *C* to *T*, and the number of iterations required.

5.2.2. WCL-7 – seven-alternative forced choice

A target sound T and seven probes A..G, 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 more closely resembles T. The subject having made a choice, the following steps are executed:

- 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.

- Recalculate the weighted centroid C,

- Generate a new set of probes A .. G. As before, this is not entirely random, as the Euclidean distance between the probes needs to be of a sufficient magnitude to allow audible timbral differences to be perceived by the subject.

- As before, data relating to the interaction is logged by the software.

6. PROCEDURE

Six versions (two timbre spaces x three search strategies) of the software **I .. VI** were prepared and loaded onto six Apple Mac eMac computers. The characteristics of the target sounds were as shown in table 1.

	Target sound parameters	**Initial Euclidean distance between weighted centroid and**
Formant space	Formant I centre frequency = 123.2 Hz Formant II centre frequency = 616 Hz Formant III centre frequency = 5447.119 Hz	8.124
SCG-EHA space	Attack time = 0.013 seconds EHA = 1 dB SCG = 6.938	6.403

Table 1: Target sound parameters in the formant and SCG-EHA spaces.

These parameters place the target sounds near the edge of their respective timbre spaces. As the weighted centroid of the probability table will, at the outset, will be at its centre (because all the probability values are the same), this will facilitate the tracking of its movement.

6.1. Testing

Fifteen subjects were used for this test, who were paid for their time. 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. Each subject was then asked to run each test I to VI; 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.

6.1.1. MLS (tests I and II)

Each subject was asked to manipulate the three software sliders, listening to the generated sound each time until EITHER sixteen iterations had been completed OR a slider setting was found for which the generated sound was judged to be indistinguishable from the target. The choice of sixteen was pragmatically arrived at in the course of a number of pilot tests; it provided a sufficient search of the space for assessing the efficacy of the approach while minimising the risk of fatigue in the task. It was also noted that there was little or no further convergence on the target after about the sixteenth iteration.

6.1.2. WCL-2 : (tests III and IV)

Each subject was asked to listen to the target and then judge which of two sounds A or B more closely resembled it. After making the selection by clicking on the appropriate button, two new sounds A and B were generated by the software, and the process repeated until EITHER sixteen iterations had been completed OR the sound generated by the software was judged to be indistinguishable from the target.

6.1.3. WCL-7: (tests V and VI)

The same procedure was adopted here, except that subjects were asked to listen to the target and then judge which one of seven sounds presented by the software more closely resembled it.

6.1.4. Control

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 strategy was run with a simulation of user input, where the 'user response' was entirely random.

7. RESULTS

Figure 3 shows the three trajectories, MLS, WCL-2 and WCL-7, averaged out for all fifteen subjects in the formant space. We consider first the mean trajectory followed by the sound generated by subjects using the MLS strategy. There was some degree of variety in individual subject trajectories – however, overall, the first five iterations show a convergence on the target. Trajectories from iteration five onwards, however, become increasingly erratic, as subjects attempted to 'fine tune' the sound arrived at. Many subjects began the search by adjusting all three sliders to their minimum value and then incrementally adjusting the value of a single slider for a few iterations before turning their attention to another one.

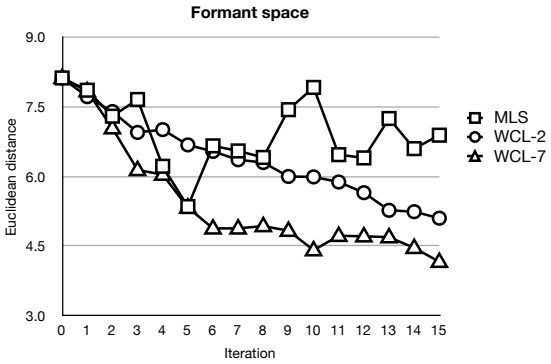

Figure 3: Mean weighted centroid trajectories in the formant space.

By contrast, the WCL-2 mean trajectory shows a steady reduction in the distance; from 7.72 at the first iteration to 5.1 at the fifteenth. In fact, nearly all the individual subject traces showed a similar trajectory. An even steeper gradient is shown in the WCL-7 mean trajectory, from 7.84 at the first iteration to 4.98 at the fifteenth.

Turning now to the operation of the three strategies in the SCG-EHA space, we see in figure 4 a much clearer overall convergence in all cases. The MLS strategy trace shows again that many subjects began the search by moving the sliders to their minimum value, thus causing the overall jump in value between iterations 0 and 1. After iteration 1, the individual trajectories show, in most cases, the probe sound approaching the target. Again, the trajectories of the WCL-2 and WCL-7 traces both show a steady reduction in the distance.

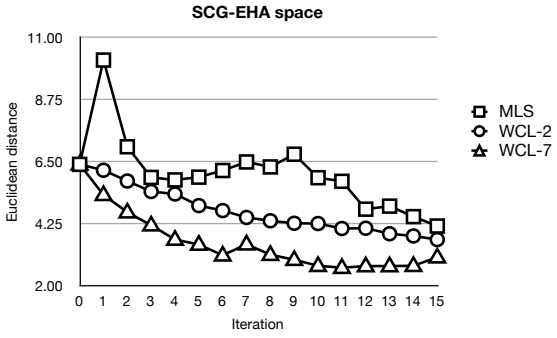

Figure 4: Mean weighted centroid trajectories in the SCG-EHA space.

7.1. 'Control' results

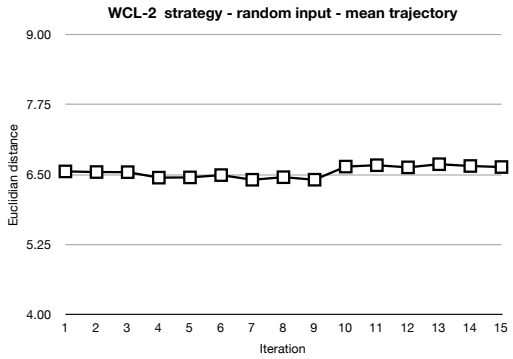

Figure 5: Mean weighted centroid trajectory using random user input for the WCL-2 strategy.

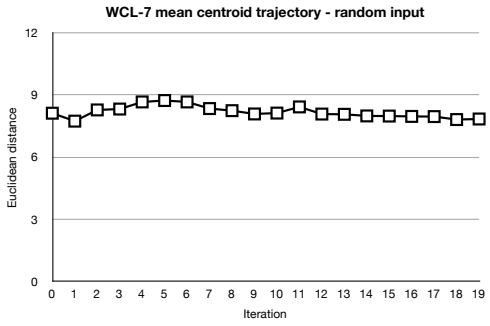

Figure 6: Mean weighted centroid trajectory using random user input for the WCL-7 strategy.

Figures 5 and 6 show the weighted centroid trajectories when random 'user' responses were given in the WCL-2 and WCL-7 versions of the search strategy. Individual subject trajectories showed the characteristics of Brownian motion, or the 'drunkard's walk' – the random path taken by, for example, a particle suspended in a fluid. However, the mean trajectory in both cases was more or less a straight line, suggesting that random input produced random output.

To summarise: in all three strategies deployed in the two timbre spaces, there was considerable variation in individual subject performance. However, the mean trajectories of 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.

8. SEARCHING A MULTIDIMENSIONAL MDS SPACE

The two timbre spaces examined in the previous section are very simple and limited in their coverage. In order to determine how effectively the WCL method might operate in a more realistic and 'real world' timbre space, we now need to test it against a wider and more musically useful range of timbres. To build such a space, we can begin by assembling a palette of sounds drawn from a list of orchestral musical instruments whose timbres are very diverse. Such a timbre space would necessarily be highly multidimensional; however, data reduction techniques such as MDS and PCA can be used to represent these sounds in a space of reduced dimensionality while, at the same time, preserving most of the variance between them. Such a space can be used as a vehicle for synthesis.

8.1. Multidimensional scaling (MDS) for synthesis - background

As MDS has been shown to be effective both as a means of determining salient features of timbre and for representing similarity/dissimilarity relationships between timbres in an timbre space of reduced dimensionality, a number of studies have investigated its suitability as a vehicle for sound synthesis. In particular, Hourdin, Charbonneau and Moussa [19, 36] demonstrated that an MDS space of reduced dimensionality, created, not from psychoacoustic listening tests, but from physical descriptions of the sound had potential use as a synthesis space. In this study, a set of forty orchestral instrument tones, partly based on that used by Grey [8], but also including a number of others, such as a marimba, tubular bells, and a harp, was analysed using heterodyne filtering [37-41]. The eighty-column matrix generated by this process represented time-varying frequency and amplitude data for each harmonic of the input tones was, in turn, reduced in an MDS analysis to one of seven factors – i.e. a seven dimensional solution was generated in which the relative distance relationships between the objects (each corresponding to a row or spectral snapshot in the original heterodyne data) reflected their distances in the eighty dimensional space; this reduced dimensionality space could nevertheless be used to resynthesize the original tones with a high degree of accuracy.

Secondly, particular trajectory curves in the derived space seemed to be associated with particular timbres. While no listening tests were performed to verify this, the authors suggested that this may be 'an interesting tool for composers' – certainly it implies a degree of congruence between the physical and perceptual spaces. Thirdly, the space seemed to be stable. This was demonstrated by removing twelve sounds from the list of forty tones and rebuilding the MDS space; neither the shape of the space, nor the distance relationships between the tones inhabiting it differed significantly from the space originally generated. Finally, distances in the space seemed to reflect perceptual distances - when intermediate curves were plotted which were interpolations between two existing curves (in this case, tenor trombone and cello played *martelé*), the resultant tone sounded plausibly like a hybrid of these two.

McAdams, Beauchamp *et al* [13] noted that MDS does not necessarily generate components which map in a one-to-one fashion to a clearly identified acoustical quantity that could be varied in sound synthesis. This, in turn, implies that simply providing a user with a set of sliders each of which corresponds to a principal component (which is the essence of the MLS method) might not be a successful strategy. The tests described here investigate this.

8.2. Construction of the MDS space

The timbre space which was constructed was seven dimensional. Six dimensions were derived through MDS. The seventh was attack time, with the same characteristics as those of the SCG-EHA space described previously. The choice of a six dimensional space will be justified in the section describing the MDS process in detail.

Both the space and the construction technique used to build it were derived in part from the work of Hourdin *et al* [19, 36]; the list of fifteen instrumental timbres was broadly the same as that used in that study, and were : alto saxophone (no vibrato, *mezzo forte*), bass clarinet, bass flute, bassoon, Bb trumpet, cello (*sul* A), Eb clarinet, flute, French horn, oboe, soprano saxophone, tenor trombone and viola (*sul* G). The samples were taken from the sample library of the University of Iowa Electronic Music Studios [42]; all samples were recorded anechoically in mono, 16 bit, 44.1 kHz AIFF format. The pitch of all the instrumental sounds was Eb above middle C (311

Hz); and all were played *mezzo forte*, except where otherwise indicated. Each instrumental sample was then edited to remove the onset and decay transients, leaving only the steady state portion, which was, in all cases, 0.3 seconds.

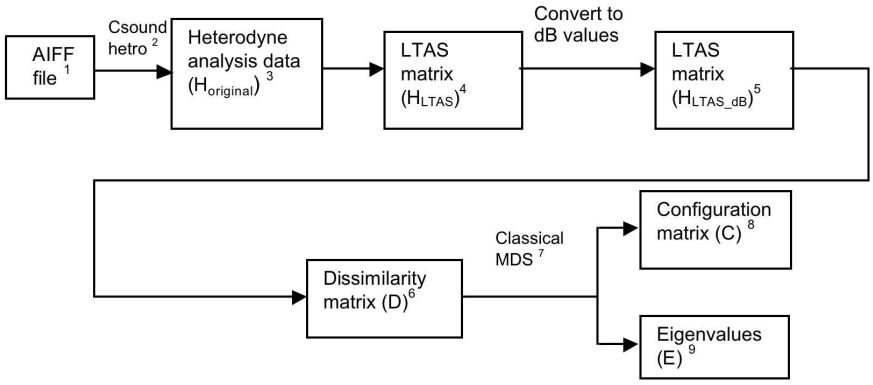

Figure 7: Multidimensional scaling of instrument samples.

The process is summarised in figure 7 and discussed here.

1. All the edited samples were normalized to –3 dB relative to full amplitude, using sound editing software, and spliced together to form one single AIFF file.

2. The audio file was then processed using *heterodyne filter analysis*. Heterodyne filtering resolves periodic or quasi-periodic signals into component harmonics, given an initial fundamental frequency: the multiplication of the input waveform by a sine and cosine function at harmonic frequencies and the summing of the results over a short time period yields amplitude and phase data for each harmonic. The implementation used here was that provided as a utility (*hetro*) as part of the *Csound* audio programming environment [43].

3. The output of *hetro* is a matrix of data in which the columns contain the time-varying amplitude and frequency values of each harmonic, and each row is a breakpoint snapshot of the instantaneous spectrum. However, because we are concerned with steady state spectra, the columns representing harmonic frequency fluctuations (F) were not included in the analysis and were removed – thus the N x 40 matrix becomes an N x 20 one.

4. A new 15 x 20 matrix H_{LTAS} was generated from the heterodyne data matrix $H_{original}$, such that each row held the Long Time Averaged Spectrum (LTAS) for one instrumental sound.

5. The heterodyne data contained linear harmonic amplitudes. These were converted to decibels, as shown in equation 7; firstly, to be consistent with the space and secondly, because logarithmic rather than linear axes more closely align with amplitude perception.

$$H_{LTAS\_dB} = 20\log\left(H_{LTAS}\right) \tag{4}$$

6. A dissimilarity matrix D was built from H_{LTAS_dB} using the *pdist()* function in MATLAB. This is a 15 x 15 matrix whose (ij)th element is equal to the Euclidean distance between the (i)th and (j)th points in H_{LTAS}.

7. The dissimilarity matrix D was then used as input to a classical multidimensional scaling function *cmdscale()*. This has two outputs.

8. The first output is a 15 x p configuration matrix C, where $p < 15$, which is a solution space to the input dissimilarity matrix D (and which may or may not be identical to H_{LTAS_dB}).

9. The second output is a vector *E* holding the eigenvalues of *C*C'*.

Each eigenvalue corresponds to one axis of the p dimensional configuration matrix C; its magnitude indicates the relative contribution of the corresponding axis to the building of the dissimilarity matrix D (in other words, the amount of information associated with that axis). We can express these eigenvalues as percentages of the total amount of information – see figure 8.

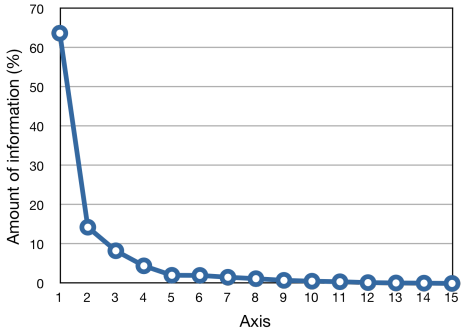

Figure 8: Eigenvalues of C*C'

Note that the first six eigenvalues are considerably greater in magnitude than the remaining nine. In fact, 95 % of the total information required to reconstruct the spectra is associated with just six axes; thus, MDS can be used to reduce the dimensionality from twenty to six with minimal loss of information.

1. Having established this, a new 6-dimensional space Y was generated using MDS from the dissimilarity matrix D (see figure 9) This was done in MATLAB using the *mdscale()* function.

2. The alignment of points in the reduced space is such that translation (centering the data), rotation, reflection and scaling are needed in order to recover, with minimum error, the original heterodyne data for synthesis. The data to do this was obtained using the *procrustes*() function in MATLAB; this is a function which determines a linear transformation of the points in one matrix which best conforms them to those in another. In this case, the two matrices are the reduced space just generated and the original matrix holding the Long Time Averaged Spectrum (LTAS) for each instrumental sound (H_{LTAS_dB}).

Figure 10 shows the fifteen instrumental sounds placed in a three dimensional space (the first three columns of the reduced space dataset).

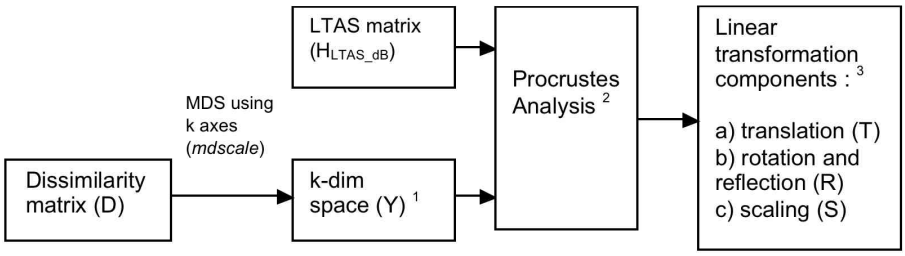

Figure 9: Process of building the reduced dimensionality space.

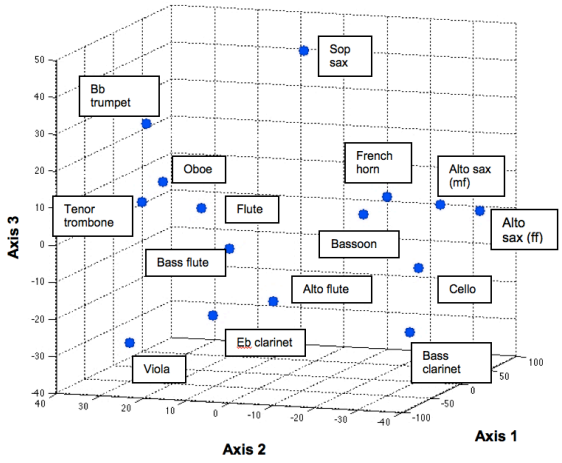

Figure 10: The 15 instrumental sounds located in a three dimensional space following MDS analysis.

The six dimensions of the reduced space describe sounds which are dynamically invariant. In order to maintain a degree of methodological continuity with the other two spaces, the seventh axis describes the attack envelope. The attributes are the same as that of the rise time axis of the SCG-EHA space – i.e. ranging from 0.01 to 0.2 seconds.

Sounds represented in the reduced space Y can be auditioned by means of a data recovery process, in which a given sound is dynamically generated from a single six-coordinate point in the space.

The single row, 6 column matrix P containing the coordinates of a point in Y is transformed using the data obtained from the *procrustes*() function, in order to recover the heterodyne data and to best align it with the original matrix HLTAS.

$$H_{LTAS\_dB\_reconstructed} = (P * R) + T \qquad (5)$$

The resultant single row, 20 column matrix HLTAS_dB_reconstructed contains the long time averaged amplitudes of the harmonics of the desired sound. The elements of this matrix are converted to linear form, as follows:

$$H_{LTAS\_reconstructed} = 10^{\frac{H_{LTAS\_db\_reconstructed}}{20}} \qquad (6)$$

This data can be input to an additive synthesis process for playback. The overall envelope (variable rise time and fixed decay envelope is imposed at this point).

9. PROCEDURE

Three versions of the software (as before, the MLS, WCL2 and WCL7 versions) were prepared and loaded onto three Apple eMac computers. The testing procedure was as described previously, except that this time, twenty subjects were used.

10. RESULTS

Figure 11 shows the averaged trajectory for all twenty interactions in the space over nineteen iterations, and seems to indicate that, overall, the MLS method is not a satisfactory search strategy in this particular timbre space. Inspection of the individual trajectories shows only one example of a subject who was able to use the controls to converge on the target.

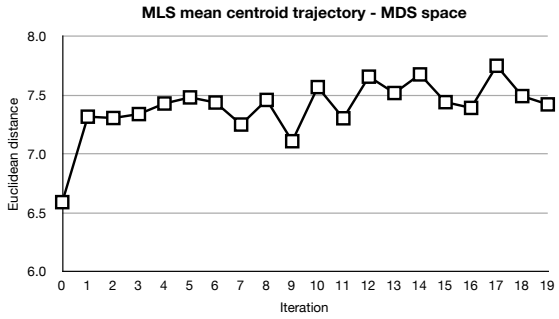

Figure 11: Mean weighted centroid trajectory in MDS space using multidimensional line search.

By contrast, the mean trajectory of the weighted centroid in the WCL-2 strategy shows a small but steady convergence on the target, although there was considerable variation in the individual subject trajectories.

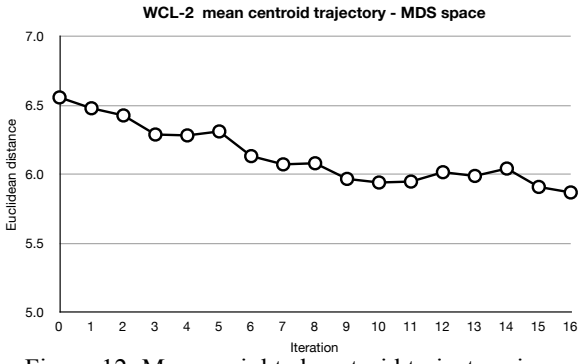

Figure 12: Mean weighted centroid trajectory in MDS space using WCL-2 strategy.

The WCL-7 trajectory, shown in figure 13 shows a slightly steeper convergence on the target; again, there was considerable variation in the individual trajectories.

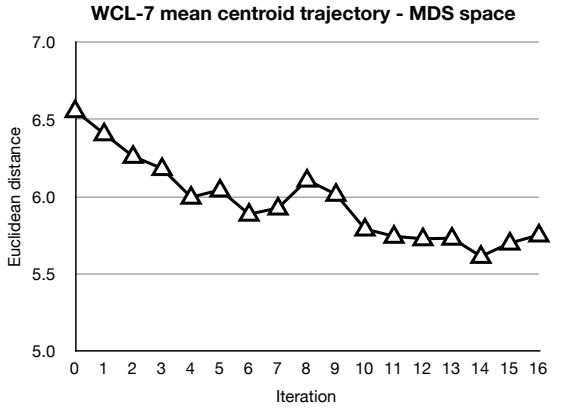

Figure 13: Mean weighted centroid trajectory in MDS space using WCL-7 strategy.

10.1. Summary of results

A summary of the results from this section, combined with those from the formant and SCG-EHA spaces is presented in figure 13. 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 difficult to navigate than a three dimensional one. Another possible reason is the far greater inertia of the MDS space probability table (consisting of 77 = 823543 cells) relative to that of the formant and SCG-EHA probability tables (1690 and 1815 cells, respectively), which would cause a slower shift of the weighted centroid across the space. This might be addressed by increasing the factor by which probability values are updated on each iteration corresponding to a sound which is 'closer' to a chosen sound.

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 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. (It is important, however, to note that subjects invariably required more time to complete the WCL-7 interaction, reflecting the greater difficulty of the task.) 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. Again, the gradient is likely to be highly sensitive to the value of the multiplication factor.

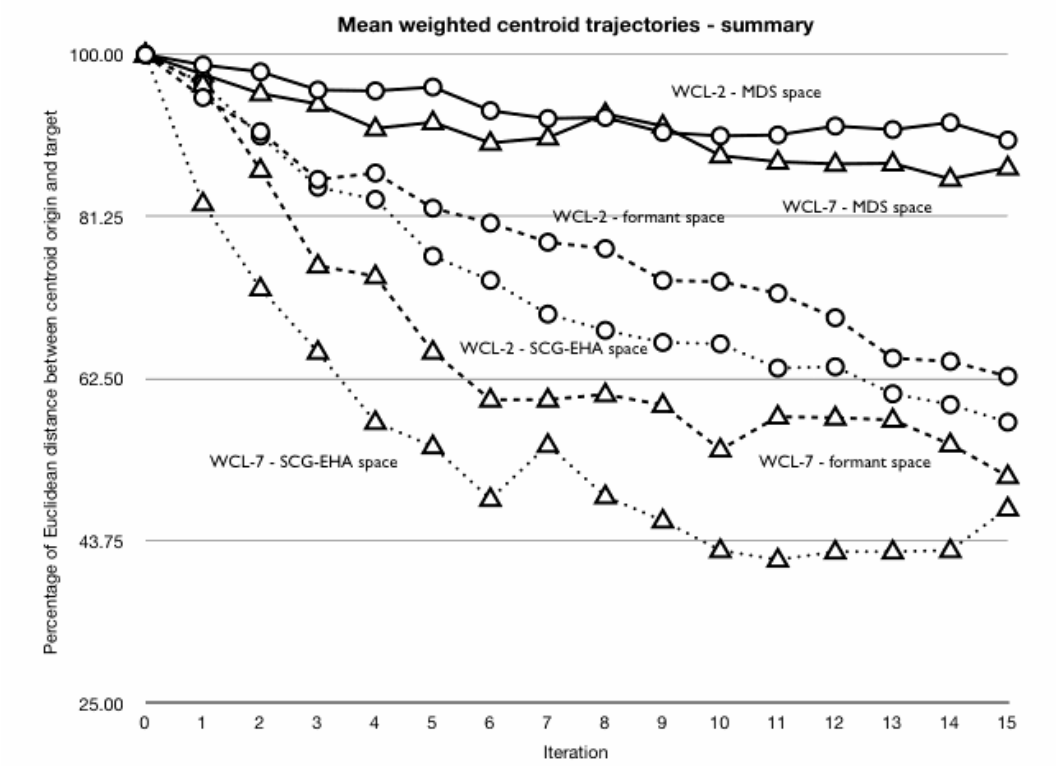

Figure 13: Mean trajectories of weighted centroid for WCL-2 and WCL-7 strategies in three different timbre spaces.

11. CONCLUSION AND DISCUSSION

We have presented a new user-driven search strategy based on weighted centroid localization (WCL), for searching suitably configured timbre spaces. The empirical work suggests that the WCL method performs significantly better in relatively simple three dimensional spaces (in this case the formant space and the SCG-EHA spaces) than in spaces where the dimensionality is greater (the MDS space) – a result which is, in itself, not surprising; but in all three spaces, the two versions of the WCL performed better than the multidimensional line search strategy (where subjects had direct access to the dimensions of the space).

While, in principle, the work presented here offers the user a novel means of searching a suitable configured timbre space, the system is both too limited in scope and too slow in operation for 'real world' use. Possible directions for new work which would address these and other problems are discussed here, together with more general proposals for further work.

A 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.

The multiplication factor value ($\sqrt{2}$) used to update the probability table in the WCL-2 strategy is not necessarily optimal and further studies could be undertaken to determine a better one. In the WCL-7 strategy, the gradient of cell values in the probability table is linear (related to distance from chosen probe); it would be of interest to ascertain whether better results might be obtained if it was (for example) exponential.

All three spaces investigated in this thesis have been constructed such that distances between the sounds in the space are Euclidean distances, rather than being based on any other metric. This is justifiable, as we are primarily interested in relative distances, both real and perceived, rather than in the perceptual properties of the spaces themselves. However, there are other metrics that can be used, which were explored by McDermott, Griffith *et al* [44]; it would be of interest to see how well the WCL search strategy performed in such spaces.

Finally, as stated above, a useful interface for timbre needs to provide the means of specifying the time-variant aspects of sound. To implement this in the interface presented here would require a more complex mapping of the search space to the probability space, and this, too, is a line of research to be pursued in the future.

12. REFERENCES

[1] American National Standards Institute, "USA standard psychoacoustical terminology S3.20," New York: American National Standards Institute, (1973).

[2] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in *Frequency Analysis and Periodicity Detection in Hearing*, R. Plomp and G. F. Smoorenberg eds. Suithoff (1970).

[3] J. C. Risset and D. L. Wessel, "Exploration of Timbre by Analysis and Synthesis," in *The Psychology of Music*, D. Deutsch eds. Academic Press (1999).

[4] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1-27 (1964 March).

[5] J. B. Kruskal and M. Wish, "Multidimensional Scaling," SAGE Publications (1978).

[6] L. Wedin and G. Goude, "Dimension analysis of the perception of instrumental timbre," *Scandinavian Journal of Psychology, 13, 228-240*, vol. pp. 228-240 (1972 Sept).

[7] J. R. Miller and E. C. Carterette, "Perceptual space for musical structures," *J. Acoust. Soc. Am.*, vol. 58(3), pp. 711-720 (1975).

[8] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.*, vol. 61:5, pp. 1270-1277 (1977 May).

[9] C. L. Krumhansl, "Why is musical timbre so hard to understand?," presented at the Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg symposium,Lund, Sweden, 1989.

[10] G. Sandell, "Effect of spectrum and attack properties on the evaluation of concurrently sounding timbres," presented at the Program of the 118th meeting of the Acoustical Society of America,1989.

[11] R. A. Kendall and E. C. Carterette, "Perceptual scaling of simultaneous wind instrument timbres," *Music Perception*, vol. 8, pp. 369-404 (1991).

[12] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Am.*, vol. 94, pp. 2595-2603 (1993 Nov).

[13] P. Toiviainen, M. Kaipainen and J. Louhivuori, "Musical timbre: similarity ratings correlate with computational feature space distances," *Journal of New Music Research*, vol. 24, pp. 282-298 (1995).

[14] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *J. Acoust. Soc. Am.*, vol. 63, pp. 1493-1500 (1978 May).

[15] D. Ehresman and D. L. Wessel, "Perception of Timbral Analogies," Technical report 13, IRCAM, Paris, (1978).

[16] S. McAdams and J. C. Cunible, "Perception of Timbral Analogies," *Philosophical Transactions of the Royal Society of London - Series B - Biological Sciences*, vol. pp. 383-389 (1992 June).

[17] G. Sandell and W. Martens, "Perceptual evaluation of principal components-based synthesis of musical timbres," *J. Audio Eng. Soc.*, vol. 43, pp. 1013-1028 (1995).

[18] C. Nicol, "Development and Exploration of a Timbre Space Representation of Audio," PhD thesis, University of Glasgow (2005).

[19] C. Hourdin, G. Charbonneau and T. Moussa, "A Sound Synthesis Technique Based on Multidimensional Scaling of Spectra," *Computer Music Journal*, vol. 21, pp. 56-58 (1997 Summer).

[20] R. Ashley, "A Knowledge-Based Approach to Assistance in Timbral Design," presented at the Proceedings of the 1986 International Computer Music Conference,The Hague, Netherlands, 1986.

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

[22] E. R. Miranda, "An Artificial Intelligence Approach to Sound Design," *Computer Music Journal*, vol. 19:2, pp. 59-75 (1995).

[23] P.-Y. Rolland and F. Pachet, "A Framework for Representing Knowledge about Synthesizer Programming," *Computer Music Journal*, vol. 20, pp. 47-58 (1996).

[24] R. Ethington and B. Punch, "SeaWave: A System for Musical Timbre Description," *Computer Music Journal*, vol. 18:1, pp. 30-39 (1994).

[25] P. Dahlstedt, "Evolution in creative sound design," in *Evolutionary Computer music* E. R. Miranda and J. A. Biles eds. Springer (2007).

[26] A. Horner, J. Beauchamp and L. Haken, "Machine Tongues XVI: Genetic Algorithms and Their Application to FM Matching Synthesis," *Computer Music Journal 17:4 pp 17-29*, vol. pp. 17-29 (1993).

[27] C. G. Johnson, "Exploring the sound-space of synthesis algorithms using interactive genetic algorithms," presented at the AISB'99 Symposium on Musical Creativity,Edinburgh, 1999

[28] T. J. Mitchell and A. G. Pipe, "Convergence Synthesis of Dynamic Frequency Modulation Tones Using an Evolution Strategy," in *Applications on Evolutionary Computing*, eds. Springer (2005).

[29] J. Blumenthal, R. Grossmann, F. Golatowski and D. Timmermann, "Weighted Centroid Localization in Zigbee-based Sensor Networks," presented at the WISP 2007. IEEE International Symposium on Intelligent Signal Processing, 2007.

[30] S. Handel, "Listening," MIT Press (1989).

[31] R. Plomp and J. M. Steeneken, "Pitch versus timbre," presented at the Proceedings of the 7th International Congress of Acoustics,Budapest, 1971.

[32] A. Caclin, S. McAdams, B. K. Smith and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *J. Acoust. Soc. Am.*, vol. 118, pp. 471-482 (2005 July).

[33] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions specificities and latent subject classes," *Psychological Research*, vol. 58, pp. 177-192 (1995).

[34] J. Marozeau, A. de Cheveigne, S. McAdams and S. Winsberg, "The dependency of timbre on fundamental frequency," *J. Acoust. Soc. Am.*, vol. 114, pp. 2946 – 2957 (2003 Nov).

[35] J. Krimphoff, S. McAdams and S. Winsberg, "Caractérisation du timbre des sons complexes. II. Analyses acoustiques at quantification psychophysique.," *Journal de Physique IV. Colloque C5, supplément au Journal de Physique III*, vol. 4, pp. 625-628 (1994).

[36] C. Hourdin, G. Charbonneau and T. Moussa, "A Multidimensional Scaling Analysis of Musical Instruments' Time Varying Spectra," *Computer Music Journal*, vol. 21:2, pp. 40-55 (1997 Summer).

[37] M. D. Freedman, "A technique for analysis of musical instrument tones," PhD thesis, University of Illinois (1965).

[38] M. D. Freedman, "Analysis of musical instrument tones," *J. Acoust. Soc. Am.*, vol. 41, pp. 793-806 (1967 April).

[39] J. Beauchamp, "A computer system for time-variant harmonic analysis and synthesis of musical tones," in *Music by Computers*, H. von Foerster and J. W. Beauchamp eds. Wiley (1969).

[40] J. A. Moorer, "The Heterodyne Filter as a Tool for Analysis of Transient Waveforms," Memo 208, Stanford Artifical Intelligence Laboratory, Stanford, California, (1973).

[41] J. A. Moorer, "On the loudness of complex, time variant tones," *Report no STAN-M-4, Center for Computer Research in Music and Acoustics, Department of Music, Stanford University*, vol. pp. (1975).

[42] L. Fritts, "The University of Iowa Electronic Music Studios - Musical Instrument Samples," http://theremin.music.uiowa.edu/MIS.html, 2008 9th June.

[43] M. Klapper, "Working with Csound's ADSYN, LPREAD, and LPRESON Opcodes," in *The Csound Book*, R. Boulanger eds. MIT Press (2000).

[44] J. McDermott, N. J. L. Griffith and M. O'Neill, "Toward User-Directed Evolution of Sound Synthesis Parameters," in *Applications on Evolutionary Computing*, eds. Springer (2005).