Sound synthesis and musical timbre: a new user interface

Allan Seago
London Metropolitan University
41, Commercial Road, London E1 1LA
a.seago@londonmet.ac.uk

Sound creation and editing in hardware and software synthesizers presents usability problems and a challenge for HCI researchers. Synthesis parameters vary considerably in their degree of usability, and musical timbre itself is a complex and multidimensional attribute of sound. This paper presents a user-driven search-based interaction style where the user engages directly with sound rather than with a mediating interface layer.

1. INTRODUCTION

Most 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 twenty years, the user interface of the typical synthesizer is, in many respects, unchanged since the 1980's 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. The difficulty that this presents has led to a situation where most synthesizer users seem to have limited their choices of sound to selections from a bank of preset sounds. (Evidence for this is largely anecdotal, but allegedly ‘nine out of ten DX7s coming into workshops for servicing still had their factory presets intact’ (ComputerMusic 2004). This paper examines the reasons for this and proposes a system which essentially removes the mediating synthesis controller layer, and recasts the problem of creating and editing sound as one of search.

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; frequency modulation, or FM synthesis, for example, is a synthesis method that is 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 of them require a significant degree of understanding of the synthesis method being employed, and therefore present to a greater or lesser extent, usability problems for the naïve user; 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 the synthesis method being employed. The design of intuitive controllers for synthesis is difficult because of the complex nature of musical timbre.

3. TIMBRE

The process of creating and editing of a sound typically involves incremental adjustments to its various sonic attributes - pitch, loudness, timbre, and the way that these three 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 implicit in Western musical theory and notational practice. More to the point for our purposes here, it is also reflected in the design of 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).

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, Hollan et al. 1986) for the purpose of specifying any but the most crudely characterized sounds.

5. SOUND SYNTHESIS USING LANGUAGE

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

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).

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. Musical sounds are complex, differing from each other in many ways; they can only be adequately represented in a timbre space of high dimensionality (fifteen dimensions or more) – i.e. the value of n needs to be quite high.
- 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 1995; Hourdin, Charbonneau et al. 1997; Nicol 2005). Other researchers have sought to bridge the gap between attribute/perceptual space and parameter space by employing techniques drawn from artificial intelligence.

8. SEARCH ALGORITHMS

The position taken in this paper 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, Hahn et al. 1993; Johnson 1999; Dahlstedt 2001; Mandelis 2001; Mandelis and Husbands 2006). Such methods typically use interactive genetic algorithms (IGAs) 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). An interesting technique for addressing this problem was proposed by McDermott, Griffith et al (2007); this system made use of interactive interpolation or ‘sweeping’, to speed up the search process.

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. WEIGHTED CENTROID LOCALISATION

The structure and function of the system is summarised here before considering it in greater detail. An n-dimensional attribute space is constructed 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 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 element s 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, correspond 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 consider each strategy in turn here.

9.1 WCL-2 strategy

Three sounds, 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 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, the space P is effectively bisected by a line which is perpendicular to the line AB (see figure 2). The probability space P having been updated, two new probes $A_{new}$ and $B_{new}$ are generated, and the process repeated.

9.2 WCL-7 strategy

The WCL-7 strategy works slightly differently. 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 A to 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.

10. THE TIMBRE SPACES

Having considered the algorithm itself, the three timbre spaces constructed for its testing will be considered. The first one, referred to here as the formant space, was inhabited by sounds of 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. 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, McAdams, Smith and Winsberg (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 seconds in eleven 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 fifteen linear steps between 3 to 8 in harmonic rank units - that is to say, between 933 and 2488 Hz. Finally, the EHA - the attenuation of even harmonics relative to odd harmonics (EHA) - ranged from 0 (no attenuation) to 10 dB, and could take eleven different values, separated by equal steps. The sounds used in the space were synthetically generated pitched tones with a fundamental of 311 Hz (E4), containing 20 harmonics.

The last of the three spaces on which the two WCL strategies were tested, called the MDS space, 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.

Both the space and the construction technique used to build it are derived in part from the work of Hourdin et al (1997); the list of fifteen instrumental timbres is broadly the same as that used in that study. 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. Using MDS, a six dimensional solution was generated. An additional seventh dimension was added, based on rise time, with the same characteristics as those of the SCG-EHA space. The disposition of the various instruments in the timbre space is shown in figure 2.

The strategy was tested on a number of subjects - fifteen in the case of the formant and SCG-EHA space tests, and twenty for the MDS space tests (which were conducted later). 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. After making the selection by clicking on the appropriate button, new sounds were generated by the software (two in the case of the WCL-2 strategy and seven in the case of WCL-7 strategy), and the process repeated.

11. RESULTS

In all cases, the trajectory of the weighted centroid relative to the position in the space of the target sound was logged. Figure 3 shows the averaged trajectory over fifteen iterations for all subject interactions. 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.

Figure 2: The 15 instrumental sounds located in a three dimensional space following MDS analysis.
12. 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.

To what extent could this technique, or others similar to it, be used for non-musical HCI problems? The WCL technique is proposed here precisely 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.

The WCL strategy described here indicates that a search-based approach can be successful, and suggests that an approach to timbre specification in which the user is able to engage directly with sounds, rather than with a mediating layer of synthesis parameters, descriptive language or graphic representation, may be fruitful. However, both WCL and IGAs suffer from the ‘bottleneck’ problems described earlier, and further work will be needed to accelerate the interaction.

13. REFERENCES


