TIMBRE SPACE AS SYNTHESIS SPACE: TOWARDS A NAVIGATION BASED APPROACH TO TIMBRE SPECIFICATION

A Seago London Metropolitan University
S Holland The Open University
P Mulholland KMI, The Open University

1 INTRODUCTION

Much research into timbre, its perception and classification over the last forty years has modelled timbre as an *n*-dimensional co-ordinate space or *timbre space*, whose axes are measurable acoustical quantities (variously, spectral density, simultaneity of partial onsets etc). Typically, these spaces have been constructed from data generated from similarity/dissimilarity listening tests, using multidimensional scaling (MDS) analysis techniques.

Our current research is the computer assisted synthesis of new timbres using a timbre space search strategy, in which a previously constructed simple timbre space is used as a search space by an algorithm designed to synthesize desired new timbres steered by iterative user input. The success of such an algorithm clearly depends on establishing suitable mapping between its quantifiable features and its perceptual features. We therefore present here, firstly, some of the findings of a series of listening tests aimed at establishing the perceptual topography and granularity of a simple, predefined timbre space, and secondly, the results of preliminary tests of two search strategies designed to navigate this space. The behaviour of these strategies in a circumscribed space of this kind, together with the corresponding user experience is intended to provide a baseline to applications in a more complex space.

1.1 Background

The notion of sounds occupying an *n*-dimensional co-ordinate space can be traced back to Licklider[1], and Plomp [2], and has since been explored from a number of perspectives. Plomp defines timbre space as a space 'derived from the timbre dissimilarities among a set of complex tones'. There is an important distinction to be made here, however. Individual sounds in a timbre space can be presented as points whose distances from each other either reflect and arise from similarity/dissimilarity judgments made in listening tests [3], or, alternatively, where the space is the *a priori* arbitrary choice of the analyst, where the distances between points reflect calculated differences derived from, for example, spectral analysis [4]. For the purposes of this paper, we will use the term *perceptual space* for the former, and *attribute space* for the latter. In either case, the axes will be vectors representing measurable attributes of the sounds inhabiting the space.

A much used technique for the study of timbre in general has been that of multidimensional scaling, in which estimates of similarity/dissimilarity between all pairs of a set of sounds is used to construct an *n*-dimensional perceptual space, typically of low dimensionality, each of whose axes corresponds to some measurable acoustic correlate[5-11]. While MDS provides valuable data for informing theories for the 'salient dimensions or features of classes of sounds' [7], such data, of course, is in itself insufficient as a basis for a search strategy which would aid the selection of a desired timbre from a previously generated perceptual space. This is because the scaling solutions do not necessarily define a given

sound such that it could be re-synthesised from this data alone. This has been noted in a number of studies[12, 13]; a sound in the MDS space may have perceptually important features that no other sounds in the same space have – and, by the same token, two sounds could occupy the same location in a given MDS perceptual space, and nevertheless be audibly different.

That a simple perceptual space can have predictive as well as descriptive power, however, has been demonstrated. It has been shown, for example, that exchanging the spectral envelopes of tone pairs previously part of an MDS spatial solution, and then generating a new spatial solution, results in the sounds exchanging places on the axis previously interpreted as relating to spectral shape. 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[14]. Another study compared mappings of a set of synthesized stimuli generated by a Kohonen self-organising map algorithm and a perceptual matrix derived from similarity ratings acquired from listening tests, and found significant correlation[15].

1.2 Existing work

Approaches for exploiting timbre space for synthesis vary, but typically involve the mapping of coordinates of a point in the space to synthesis parameters, often those of frequency modulation (FM). A timbre space derived from that used by Hourdin, Charbonneau and Moussa[16] has been mapped to a FM synthesizer in order to produce sounds for audio interfaces[17].

Timbre space can be used as a search space; genetic algorithms (GAs) provide a means of arriving at an optimal solution within a search space[18], by encoding (usually in binary form) a population of possible solutions, evaluating each solution using a problem-specific *fitness function*, allowing the 'best' solutions to breed new solutions, and iteratively re-evaluating them. GAs have been exploited in systems which are designed to converge on the correct parameters for a given synthesis algorithm – frequency modulation[19, 20] or granular synthesis[21] - in order to generate a desired sound.

2 LISTENING TESTS

2.1 Motivation, aims and objectives

We turn now to discussion of the empirical work. The attribute space chosen is deliberately simple and low-dimensional; its perceptual topography and the physical parameters used to generate it are expected to relate more or less linearly. The work presented here is intended to investigate the details of this relationship. With detailed knowledge of the relationship, we can use the space as a vehicle for exploring the properties of candidate frameworks for user-driven search.

The first aim of the study is to establish the extent to which the Euclidian distances between three sounds, A, B and C, disposed in a predefined simple attribute space, such that the distance AC is different from BC, are reflected in perceptual differences. We wish to determine this for the following reason. In general terms, a search algorithm (such as a GA, for example) picks one or more candidate solutions from a search space, evaluates them using a fitness function and converges on the best solution available. Where the fitness function is provided by the judgement of a user, candidate solutions are selected on the basis of perceived similarity to a pre-heard or imagined target; the extent to which perceptual distances map to objectively measurable distances in the attribute or generating space therefore makes a difference to the effectiveness of the fitness function.

2.2 The attribute space

The chosen space for this empirical work is such that all of the sounds are time invariant i.e. have static spectra and are characterized by three prominent formants. A formant is a broad frequency region which causes an increase in amplitude of any spectral component partial falling within its range. Slawson, and subsequently Plomp and Steeneken[22] 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, and the stimuli chosen 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 attribute space, loosely based on vowels, has been chosen for this study; firstly, for its simplicity, 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.

2.3 Stimuli

A set of electronically synthesized pitched waveform stimuli was generated. The spectra of the pitched stimuli contained 73 harmonics of a fundamental frequency 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 the three dimensional space illustrated below.

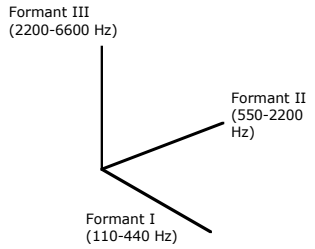
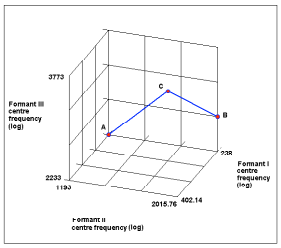


Figure 1: The three dimensional attribute space investigated in this study.

All stimuli were generated using Csound, and were exactly two seconds in duration, with attack and decay times of 0.4 seconds. Each test consisted of a triplet of pitched stimuli, **A**, **B** and **C,** disposed in the space such that ABC did not form a straight line, AC and BC had projections on all three axes, and the Euclidian distance AC was greater than that of BC (a ratio of AC:BC = 1.732 : 1) In all cases, C was the initial stimulus and A and B were the probes. Each of the forty-eight test triplets were constructed as follows: if A_x, A_y and A_z are respectively the Formant I, II and III centre frequency coordinates for A; B_x, B_y and B_z those for B and C_x, C_y and C_z those for C, positions were found for A, B and C such that the Euclidian distances AC and BC were then as shown in figure 2. The forty-eight triplets were uniformly distributed throughout the space.



$$AC = \sqrt{\left[\log\left(\frac{C_x}{A_x}\right)\right]^2 + \left[\log\left(\frac{C_y}{A_y}\right)\right]^2 + \left[\log\left(\frac{C_z}{A_z}\right)\right]^2} = 0.227797 \quad \text{Eq. 1}$$

$$AC = \sqrt{\left[\log\left(\frac{B_x}{A_x}\right)\right]^2 + \left[\log\left(\frac{B_y}{A_y}\right)\right]^2 + \left[\log\left(\frac{B_z}{A_z}\right)\right]^2} = 0.131518 \quad \text{Eq. 2}$$

Figure 2: Test stimulus triplet

2.4 Procedure

Twenty test subjects were used for this study (only nineteen responses proved to be usable, however). All were students in the Sir John Cass Department of Art, Media and Design of London Metropolitan University, studying either music technology or musical instrument building – consequently, these subjects were accustomed to listening critically to sound. Fifteen subjects used Sennheiser PX-30, the remaining five Sony MDR-V300 headphones. The forty eight tests were presented to the test subjects in the form of a series of Web pages accessed individually from a desktop computer; half the subjects received the sequence in one random order, and to the other half in another random order. The procedure was explained, and subjects encouraged to acclimatise themselves to the sounds, and to set the headphone volume at a comfortable level.

For each of the tests, each subject was asked to indicate which of the first two stimuli of the triplet sounded more like the third. (The first two stimuli of half the triplets, randomly chosen, were swapped to avoid giving clues to the subjects). In all cases, subjects were able to audition any sound as often as they wished, before making a decision.

2.5 Results

The mean number of 'correct' identifications for all 48 tests was 35.05 (73.02%) - standard deviation (s) = 6.739. The probability of this result, based on a binomial distribution B(48,0.5), is $p = 7 * 10^{-5}$, well below both the five and one per cent levels of statistical significance. We conclude from this that subjects are, in general, able to perceive relative Euclidian distances between three pitched sounds A, B and C in this particular attribute space.

3 SEARCH STRATEGIES

We turn now to consideration of two search strategies. It should be noted that while a statistically significant correlation has been found between relative Euclidian and perceptual distances in this particular attribute space, distance judgments are obviously not made with 100% accuracy; to put it another way, discrimination and distance perception in this space is errorful, and this must necessarily inform any successful search strategy.

In order to examine and compare two search strategies, the same attribute space as that described above was constructed. For our purposes here, we can define the space as a three-dimensional matrix **$S=(s_{x,y,z})$,** with axes **1..X, 1..Y** and **1..Z** being the Formant I, II and III axes respectively, where X = 16, Y=16 and Z = 10, and where x, y and z are then the co-ordinates of any sound within the space.

Two navigational strategies were tested in this pilot study. The first was one in which the subject was given direct access to the axes which describe the space - the strategy is that of **multidimensional line**

search, where the search is conducted along the axes of the space. It may be that for a space with low dimensionality, such an approach may be effective. The second is an adapted Bayes filter process, in which a three dimensional probability space is iteratively updated by user input. We discuss first the **multidimensional line search** strategy**.**

4 MULTIDIMENSIONAL LINE SEARCH

The test software presented the subject with three sliders and two buttons. Clicking on a button labelled 'Play target' played a fixed sound stimulus chosen from the attribute space, which is used as the target. The stimulus produced by the other button ('Play sound'), however, could be varied at will by moving any or all of the three sliders; this sound was used as the 'probe'. The sliders corresponded to three axes of the attribute space; moving the first slider increased/decreased x, resulting in a shift of the 'probe' sound along the Formant I axis; moving the second slider increased/decreased y, and so on. There was one unique slider configuration which produced a sound identical to the target. The software logged the changing position of the probe sound in the attribute space, and its Euclidian distance to the target.

4.1 Procedure

Three students took part in this pilot study. The subjects were asked to listen through headphones to both the 'target' sound and the 'probe' sound and then to incrementally change the 'probe' sound by moving one or more of the sliders until the two sounds were perceived by the subject to be indistinguishable.

4.2 Results and discussion

The three subjects completed the test in 43, 12 and 43 iterations, arriving at sounds whose Euclidian distances from the target were respectively 4.123, 2.236 and 3.000. The trajectories through the space are given below.

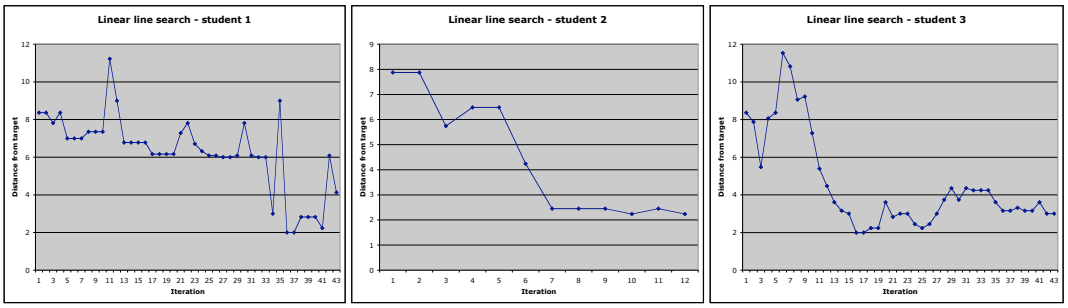


Figure 3: Multidimensional line search – test trajectories

The graphs show a slow convergence on the target sound, with occasional deviations which are quickly corrected. Note that student 3, while repeating the process forty-three times, actually achieved the minimum distance in sixteen. This seems an effective way of navigating a target sound within this attribute space of low dimensionality.

5 ADAPTED BAYES FILTER PROCESS

We turn now to the other strategy tested within this attribute space. The search strategy is an **adapted Bayes filter** process. The use of Bayesian or probabilistic networks as a means of representing and solving decision problems under uncertainty is well established in the literature. More recently, Bayesian methods have been applied to the identification and filtering of junk e-mails. The search strategy described here is an adaptation of the Bayes filter, in that a network of probabilities is iteratively updated by new input, in this case from the user. We describe here the principle, before going to the implementation.

The approach makes use of a three dimensional matrix **M=($\mathbf{m_{x,y,z}}$)** of cells with axes **1..X**, **1..Y** and **1..Z**, where X=13, Y =13 and Z =10, and **N = XYZ** is the number of cells in the matrix. The matrix **M** corresponds to the attribute space **S**, such that each cell $\mathbf{m_{x,y,z}}$ holds a numerical value representing the probability that the stimulus $\mathbf{s_{x,y,z}}$ is the target sound; at the outset this value is set to 100 for all values of x y and z.

The subject is presented with two probe stimuli A and B, and a target sound T; all three taken from the attribute space S. Each subject is then asked to judge which of A or B more closely resembles T. The subject having made a choice or A or B, each cell in the probability space M is then updated for all values of x, y and z such that, if the Euclidian distance of $\mathbf{s_{x,y,z}}$ is closer to the selected stimulus (A or B) than to the rejected stimulus, the value of $\mathbf{m_{x,y,z}}$ is multiplied by a factor of $\sqrt{2}$; otherwise it is multiplied by a factor of $1/\sqrt{2}$ (thus the space is effectively divided by a line perpendicular to a line joining A and B). After M has been updated, two new probe stimuli are randomly generated, and the process repeated. Clearly, should the user selection be 'correct' most, or all of the time, the probability values of cells in M associated with the stimuli immediately at and around the target sound T will go incrementally to a maximum.

The chosen metric for assessing the strategy is the Euclidian distance between the cell in M associated with the target sound T and the cell in M which is the **weighted centroid** – that is to say, its weighted centre of gravity - of M at any moment. The coordinates of this cell are $\bar{x}$, $\bar{y}$ and $\bar{z}$, and are given by

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}, \bar{y} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}, \bar{z} = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i} \qquad \text{Eq. 3}$$

where x, y and z are the coordinates of the *i*th cell in matrix M, and *w* is the value of the *i*th cell in M. To illustrate this, a simulated 'perfect' run (in which only 'correct' choices were made) was performed; this resulted in the graph shown in figure 4, in which the trajectory of the weighted centroid relative to the target over eleven iterations is indicated.

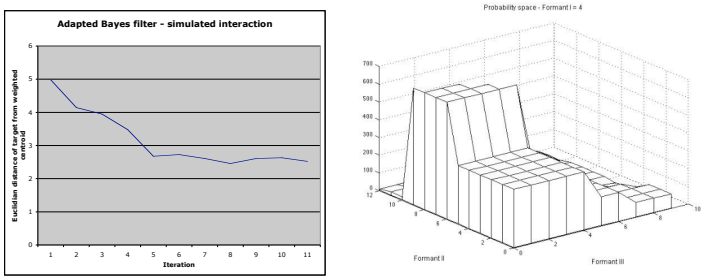


Figure 4: Adapted Bayes filter – simulated interaction trajectory and final probability values

The final probability distribution showed a peak around the cell associated with the target sound T. The second diagram shows a slice taken from the matrix M, showing the peak probability values .

5.1 Procedure

The test implementation of this strategy is now discussed. As before, the interface presents the subject with a 'Play target' button; however, in addition, two buttons which play the probe sounds A and B are presented, together with two selection buttons, allowing the subject to indicate which of A or B more closely resembles the target. Three students took part in this pilot study (not the same ones as in the previous experiment). After a brief period of familiarisation with the sounds and with the software, each subject was prompted to decide which of the probes A and B more closely resembles the target, and to respond by clicking on the appropriate selection button. Following each selection, two new probes A and B were generated. This process was repeated eleven times, and the results analysed.

5.2 Results and Discussion

The graphs shown below show the trajectory followed by the weighted centroid of the probability space M, relative to the target, for each subject. Also indicated in each graph are the projections of the trajectory along each of the three axes corresponding to the formant I II and III axes of the attribute space.

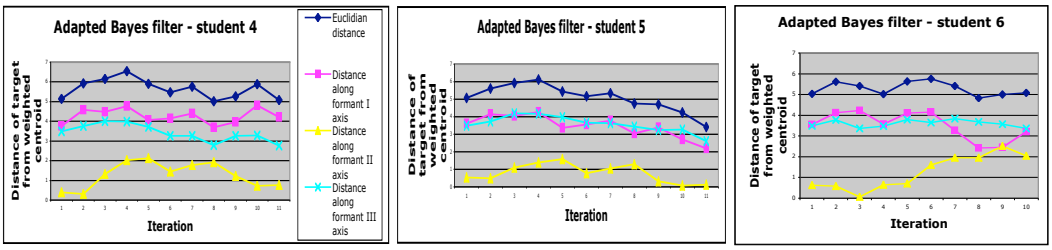


Figure 5: Weighted centroid trajectory for three test subjects.

Of the three subjects, only one (student 5) gave responses that resulted in a more-or-less smooth convergence on the target; the weighted centroids in the cases of students 4 and 6 do not seem to follow similar trajectories. While more tests clearly need to be conducted, on this evidence, it seems that a general ability to discern relative Euclidian distances in the space does not, of itself, form the basis of a robust search strategy, and that the Bayes filter strategy, in particular, does not recover well from error.

6 FURTHER WORK

The research described above is ongoing, and in the immediate future it is intended to refine the Bayes strategy, and to evaluate other search methods One important modification which is proposed is the introduction of a feedback element to the interaction, enabling the subject to identify when the navigation is adrift and to take corrective action. The more long term objective of the work is to apply a successful search strategy to other suitable attribute spaces of low dimensionality, before extending it to more complex (six or seven dimensional), but, at the same time, more musically useful spaces.

7 REFERENCES

1. J.C.R. Licklider. Basic Correlates of the Auditory Stimulus, Wiley. (1951)
2. R. Plomp. Timbre as a Multidimensional Attribute of Complex Tones, Suithoff. (1970)
3. J.C. Risset and D.L. Wessel. Exploration of Timbre by Analysis and Synthesis, Academic Press. (1999)
4. R. Plomp. Aspects of tone sensation, Academic Press. (1976)
5. L. Wedin and G. Goude. Dimension analysis of the perception of instrumental timbre, Scandinavian Journal of Psychology, 13, 228-240 (1972)
6. J.R. Miller and E.C. Carterette. Perceptual space for musical structures, Journal of the Acoustical Society of America 58(3) (1975)
7. J.M. Grey. Multidimensional perceptual scaling of musical timbres, J. Acoust. Soc. Am 61:5 (1977)
8. G. Sandell. Perception of concurrent timbres and implications for orchestration, Proceedings, International Computer Music Conference (1989)
9. G. Sandell. Effect of spectrum and attack properties on the evaluation of concurrently sounding timbres, Meeting of the Acoustical Society of America (1989)
10. R.A. Kendall and E.C. Carterette. Perceptual scaling of simultaneous wind instrument timbres, Music Perception 8 4. (1991)
11. P. Iverson and C.L. Krumhansl. Isolating the dynamic attributes of musical timbre, Journal of the Acoustical Association of America 94(5) (1993)
12. C.L. Krumhansl. Why is musical timbre so hard to understand?, Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg symposium (1989)
13. S. McAdams. Perspectives on the Contribution of Timbre to Musical Structure, Computer Music Journal 23:3 (1999)
14. D. Ehresman and D.L. Wessel.Perception of Timbral Analogies,Technical report 13,IRCAM. (1978)
15. P. Toiviainen, M. Kaipainen and J. Louhivuori. Musical timbre: similarity ratings correlate with computational feature space distances, Journal of New Music Research 24 3. (1995)
16. C. Hourdin, G. Charbonneau and T. Moussa. A Multidimensional Scaling Analysis of Musical Instruments' Time Varying Spectra, Computer Music Journal 21:2 (1997)
17. C. Nicol, S. Brewster and P. Gray. Designing Sound: Towards a system for designing audio interfaces using timbre spaces, Proceedings of ICAD 04 -Tenth Meeting of the International Conference on Auditory Display (2004)
18. J.H. Holland. Adaptation in Natural and Artificial Systems, University of Michigan Press. (1975)
19. 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 (1993)
20. J. McDermott, N.J.L. Griffith and M. O'Neill. Toward User-Directed Evolution of Sound Synthesis Parameters, Springer. (2005)
21. C.G. Johnson. Exploring the sound-space of synthesis algorithms using interactive genetic algorithms, AISB'99 Symposium on Musical Creativity (1999)
22. R. Plomp and J.M. Steeneken. Pitch versus timbre, Proceedings of the 7th International Congress of Acoustics (1971)