Towards a Mapping of Timbral Space

Allan Seago, Sir John Cass Department of Art, Media and Design, London Metropolitan University, UK
a.seago@londonmet.ac.uk www.londonmet.ac.uk

Simon Holland, Department of Computing, Open University, UK
s.holland@open.ac.uk computing.open.ac.uk

Paul Mulholland, KMI, Open University, UK
p.mulholland@open.ac.uk kmi.open.ac.uk

Proceedings of the Conference on Interdisciplinary Musicology (CIM05)
Actes du Colloque interdisciplinaire de musicologie (CIM05)
Montréal (Québec) Canada, 10-12/03/2005

Abstract

This paper presents the interim findings of a series of listening tests designed i) to test the extent to which the relative positions of sounds inhabiting a simple three-dimensional co-ordinate space, whose axes are formant centre frequencies, are reflected in perceptual distances between those sounds, and how this varies through the space, ii) to establish the perceptual granularity of the space - how the maximum distance between two sounds in the space where no difference can be heard varies throughout the co-ordinate space, and iii) the extent to which the results of i) and ii) correlate. The paper concludes that, firstly, subjects are, in general, able to perceive relative distances in this co-ordinate space, and that ability varies throughout the space, and that, secondly, the ability to discriminate timbral shifts varies quite considerably in the space. In both cases, the shift in the second formant centre frequency seems to be salient; some errors in test data, however, prevent any firm conclusions on the extent to which the results of i) and ii) correlate. The purpose of this study is to provide empirical data to inform the development of more intuitive user interfaces for timbre specification.

Background

The range of tools and techniques available to the musician for the design and editing of sound is considerable – however, the poor usability of hardware and software synthesizers, in this respect, has been noted by a number of researchers (Charbonneau 1981, 10-19; Ethington and Punch 1994, 30-39; Gaver 1993, 228-235; Miranda 1995, 59-75; Mynatt 1994, 269 - 270; Seago et al. 2004). The full use of commercial electronic synthesizers has been limited by awkward user interfaces (UI) for controlling timbre, largely because of the problems posed by its complex and multidimensional nature.

This multidimensional aspect has driven much research into timbre over the last thirty years. In many studies, this has involved identifying perceptually salient properties of timbre by deriving, using multidimensional scaling techniques, a 'best fit' *n*-dimensional co-ordinate space whose axes are not known at the outset (Grey 1977, 1270-1277; Hourdin et al. 1997, 40-55; Kaminskyj 1999, 36-9; McAdams 1999, 85-102; Misdariis et al. 1998, 3005-3006; Toiviainen et al. 1995, 282-298). The concept of a 'timbral space' has similarly underpinned studies aimed at bridging the gap between task related semantics and synthesizer related semantics by, for example, employing genetic algorithms to navigate a search space of encoded FM parameters (Horner et al. 1993, 17-29; Johnson 1999); such approaches tend to focus on a search space whose parameters are those of a particular synthesis method, but which do not map easily onto perceptual space. By contrast, we propose an approach where the space to be navigated is one which maps to perceptual space, but at the same time whose parameters lend themselves easily to synthesis.

Whether the co-ordinate space maps to perceptual space can be demonstrated by showing a link between Euclidian distances in the co-ordinate space and perceptual distances. A number of studies (Grey 1977, 1270-1277; Plomp 1970, 397-414; Toiviainen et al. 1995, 282-298) have identified classes of *n*-dimensional timbral space where there is correlation between perceptual and Euclidian distances separating sounds in the space – other studies suggest that listeners are able to perceive abstract relations and analogies between such sounds (Ehresman and Wessel 1978, ; McAdams and Cunible 1992, 383-389). It is, however, probable that a correlation may not be consistent throughout a co-ordinate space, and that the perception of timbral differences may vary.

The research presented here presents the findings of a perceptual exploration of a simple three dimensional timbral space, whose axes are based on formant centre frequencies. While formant amplitudes and frequencies do not, of course, exclusively explain timbre perception, it has been proposed (Balzano 1986, 297-314; Slawson 1968, 97-101) that they provide a better way of describing the steady state behaviour of dynamic systems (such as musical instruments) than is the case for spectrum alone. Furthermore, formant based steady state sounds are easy to synthesize with a minimum of parameters, while at the same time offering maximum timbral variation.

A number of studies on vowel formant frequency discrimination (Flanagan 1955, 613-617; Hermansky 1987, 533-534; Mermelstein 1978, 572-580; Nakagawa et al. 1982, ; Nord and Sventelius 1979) have been carried out over the years, and have reported varying results. This variation has been attributed to a variety of test subjects, methods and degree of subject training (Kewley-Port and Watson 1994, 485-496). The parameters of these and other studies were, variously, formant bandwidth (Flanagan 1957, ; Gagne and Zurek 1988, 2293-2299), formant bandwidth and fundamental frequency (Dissard and Darwin 2001, 409-415), fundamental frequency, slope, and phase relations between spectral components (Lyzenga and Horst 1997, 1755-1767). The aim of all of these studies has been primarily to identify those features which support vowel identification; the study described in this paper, while drawing on these studies, is not concerned with vowel identification as such, but with the perceptual features of the space itself.

Aims and objectives

This paper reports the interim findings of listening tests aimed at establishing the perceptual topography and granularity of a simple vowel-based three dimensional timbral space whose axes are formant centre frequencies. In particular, the degree of correlation that exists between the Euclidian distances in the co-ordinate space used to generate the sounds, and perceptual differences is reported.

1. The first aim of the study is to establish the extent to which the Euclidian distances of three sounds A, B and C, disposed in the space such that the distance AC is different from BC, is reflected in perceptual differences. The hypothesis is that there is a correlation, but that the degree of correlation will vary in different parts of the space.
2. The second aim is to examine the perceptual granularity of the space – specifically, the maximum distance between two sounds for which there there is no difference in perceived timbre. This has been studied in the literature from the point of view of the difference limen – the smallest timbral shift that can be detected. We take the converse approach by looking at the largest timbral shift that cannot be detected. The hypothesis is that this distance, averaged out for all subjects used in the listening tests, will vary in different parts of the space.
3. It is also hypothesized that there will be a correlation between the findings in 1 and 2.

This particular formant-based space is chosen, firstly because of its simplicity, and secondly, because of its similarity to the kind of co-ordinate space occupied by vowel sounds. It is argued that the use of such a space will allow maximum timbral variation in the set of sounds to be generated within an otherwise very circumscribed space.

This study is part of a wider research program into the development of more intuitive synthesizer UIs. Our contention is that a UI which embodies a perceptual model of a simple three-dimensional timbral space and which can be demonstrated successfully to enable a user to navigate successfully that timbre space would serve as a valuable reference model to inform the application of the same UI principles to more complex and realistic *n*-dimensional spaces.

Test material for relative Euclidian distance perception

The material for the tests consisted of electronically synthesized pitched and non-pitched waveform samples. For this first phase, the material comprised 168 pitched samples, each having a spectrum containing 73 harmonics of a fundamental frequency (F0) of 110 Hz, and 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 sample, 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 3210 and 6320 Hz. Each sample could thus be located in a three dimensional space whose axes were the formant centre frequencies, as indicated in figure 1.

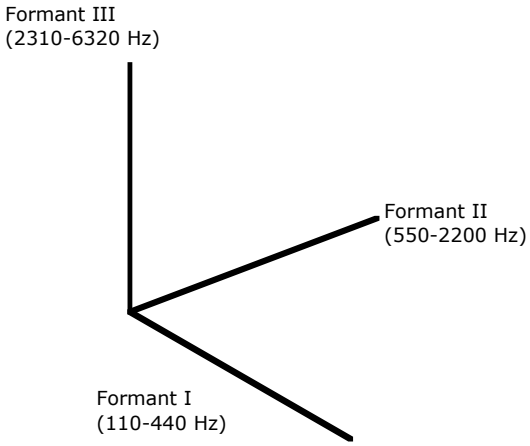


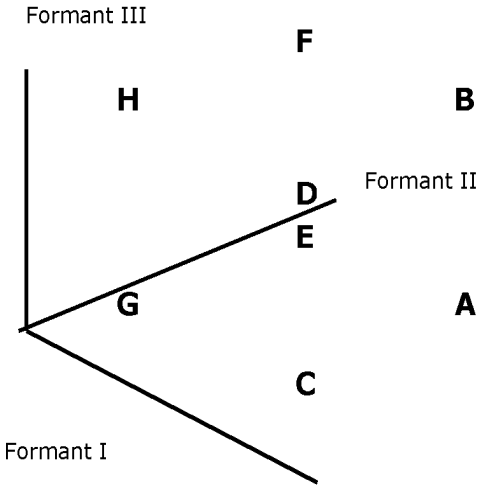
Figure 1. The three dimensional timbral space investigated in this study

56 tests were compiled from the pool of samples. Each test consisted of an equally spaced triplet of samples, whose alignment in the timbral space took one of the following trajectories:

- along the formant I axis
- along the formant II axis
- along the formant III axis
- along both the formant I and II axes
- along both the formant I and III axes
- along both the formant II and III axes
- along all three axes

The three samples making up each triplet were separated from other by a frequency ratio of 1.41 – so, for example, the samples making up a triplet aligned along the formant II axis had identical formant I and III centre frequencies, but their formant II centre frequencies were f, 1.41f and $(1.41)^2$f.

The timbral space was subdivided into eight smaller three-dimensional areas (areas A-H) as shown in figure 2.



	Formant I centre frequency range	Formant II centre frequency range	Formant III centre frequency range
Area A	220-440 Hz	1100-2200 Hz	2310-3650 Hz
Area B	220-440 Hz	1100-2200 Hz	4000-6320 Hz
Area C	220-440 Hz	550-1100 Hz	2310-3650 Hz
Area D	220-440 Hz	550-1100 Hz	4000-6320 Hz
Area E	110-220 Hz	1100-2200 Hz	2310-3650 Hz
Area F	110-220 Hz	1100-2200 Hz	4000-6320 Hz
Area G	110-220 Hz	550-1100 Hz	2310-3650 Hz
Area H	110-220 Hz	550-1100 Hz	4000-6320 Hz

Figure 2. Subdivision into areas A-H of the three dimensional space, whose boundaries are given in the table.

The triplets were disposed in the space, such that each area A-H contained seven triplets, each aligned along one of the trajectories described above.

Test material for perceptual granularity

For the second phase of the study, a set of noise-based sounds, with formant structures as described above, was compiled as previously. Each formant had identical bandwidths (Q=10) and boost (20 dB). Each of these tests consisted of a pair of sounds drawn from this pool, whose alignment in the space took one of the following forms:

- co-incident (i.e. the sounds were identical)
- separated along the formant I axis by a formant centre frequency difference of Δf_1. This is expressed as a Weber ratio, equating to

  $$\frac{\Delta f_1}{f_1} * 100 = 5.95$$

  where f_1 is the lower frequency of the pair. (This figure was arrived as a result of a pilot study, and corresponds to shifting the formant peak by about a semitone.)
- separated along the formant I axis by a formant centre frequency difference of Δf_2, expressed as a Weber ratio of

  $$\frac{\Delta f_2}{f_2} * 100 = 12.25$$

  where f_2 is the lower frequency of the pair.
- separated along the formant II axis by a Weber ratio of 5.95
- separated along the formant II axis by a Weber ratio of 12.25
- separated along the formant III axis by a Weber ratio of 5.95
- separated along the formant III axis by a Weber ratio of 12.25

Sound pairs were placed in areas A- H, as described in the previous section, so that in each area, there was a pair located along each axis separated by both Δf_1 and Δf_2, six pairs in all; so, for example, in area C, the pairs were as follows:

	Lower frequency of pair	Δf_1	Δf_2
Formant I shift	310	328.43	-
Formant I shift	310	-	347.96
Formant II shift	776	822.14	-
Formant II shift	776	-	871.02
Formant III shift	2903	3075.61	-
Formant III shift	2903	-	3258.49

Figure 3. Sound sample pairs located in area C. A corresponding pattern of sound sample pairs were located in the other areas of the space.

Areas A, B, D, E, F, G and H were correspondingly populated. A number of co-incident pairs were also generated and included in the test as a control.

All sounds for both phases of the test, pitched and non-pitched, were generated using Csound, and were exactly 2 seconds in duration, with attack and decay times of 0.4 seconds.

Procedure

Twenty test subjects were used for the study, who were paid for their participation. They were mostly (but not all) students in the Sir John Cass Department of Art, Media and Design of London Metropolitan University, studying either music technology or musical instrument building, restoration and repair – consequently, these subjects were accustomed to listening critically to sound. The tests were presented through Sony MDR-V300 headphones to the test subjects in the form of a series of Web pages (*www.city.londonmet.ac.uk/~seago/Experiment1.html*) accessed individually from a desktop computer. The procedure was explained, and test subjects encouraged to acclimatise themselves to the sounds, and to set the headphone volume at a comfortable level.

In the first phase of the experiment, intended to test Euclidian distance perception, each subject was asked to listen to the 56 tests, and for each of the tests, to indicate which of the first two samples of the triplet sounded more like the third. (The first two samples of half the triplets, randomly chosen, were swapped to avoid giving clues to the subjects). The tests were presented in random order. There is a risk in listening tests of this kind that perceptions can be distorted at the beginning of the test by unfamiliarity with the material, and at the end by fatigue, with the data from the tests in the middle being, as it were, most 'reliable'. In order to diminish this effect, the subjects were divided into two groups of ten; one group received the sequence in one random order, the other group received it in another order.

Before starting on the next phase, subjects were given a ten to twenty minute break.

In the second phase, intended to test perceptual granularity, subjects were asked to listen to each pair, and rate them for the degree of perceived difference – choices were 'no difference', 'slight difference' and 'clear difference'. As before, the tests were presented to the two groups of subjects in different random orders.

In all cases, subjects were able to audition any sound as often as they wished, before making a decision.

Results of perceptual granularity tests

We first consider the results from the perceptual granularity tests. Data from the second phase of the listening tests were averaged out for all 20 subjects and broken down by formant – that is to say, by the axis along which the sounds in each pair were aligned - as shown in figure 4.

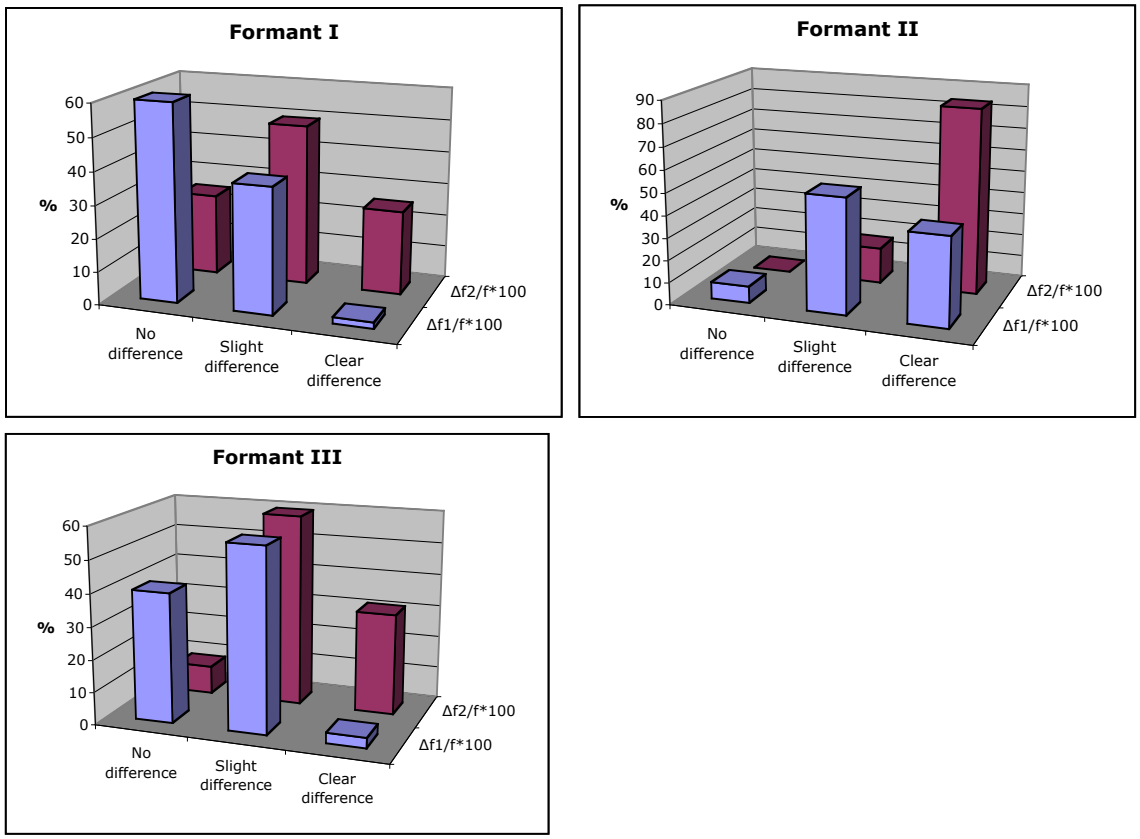


Figure 4. Breakdown of perceptual granularity listening tests by formant, showing ratings for 'No difference', 'Slight difference' and 'Clear difference' for pairs separated by Δf_2 and Δf_2.

The high mean percentage of 'no difference' ratings for sound sample pairs separated by Δf_1 (Weber ratio 5.95) along the formant I axis is quite striking, when compared to the ratings for pairs separated by the same ratio along the other two axes. In contrast to this, we see a very low mean percentage of 'no difference' ratings for sound sample pairs separated by Δf_1 along the formant II axis, suggesting that shifts in formant frequency in this frequency range are salient in the perception of timbral change in this particular space. It was noted, however that there was a far greater sensitivity to formant I shifts in areas A, B, C and D than in areas E, F, G and H. The data was then further broken down by these two area blocks; the results are presented in figure 5.

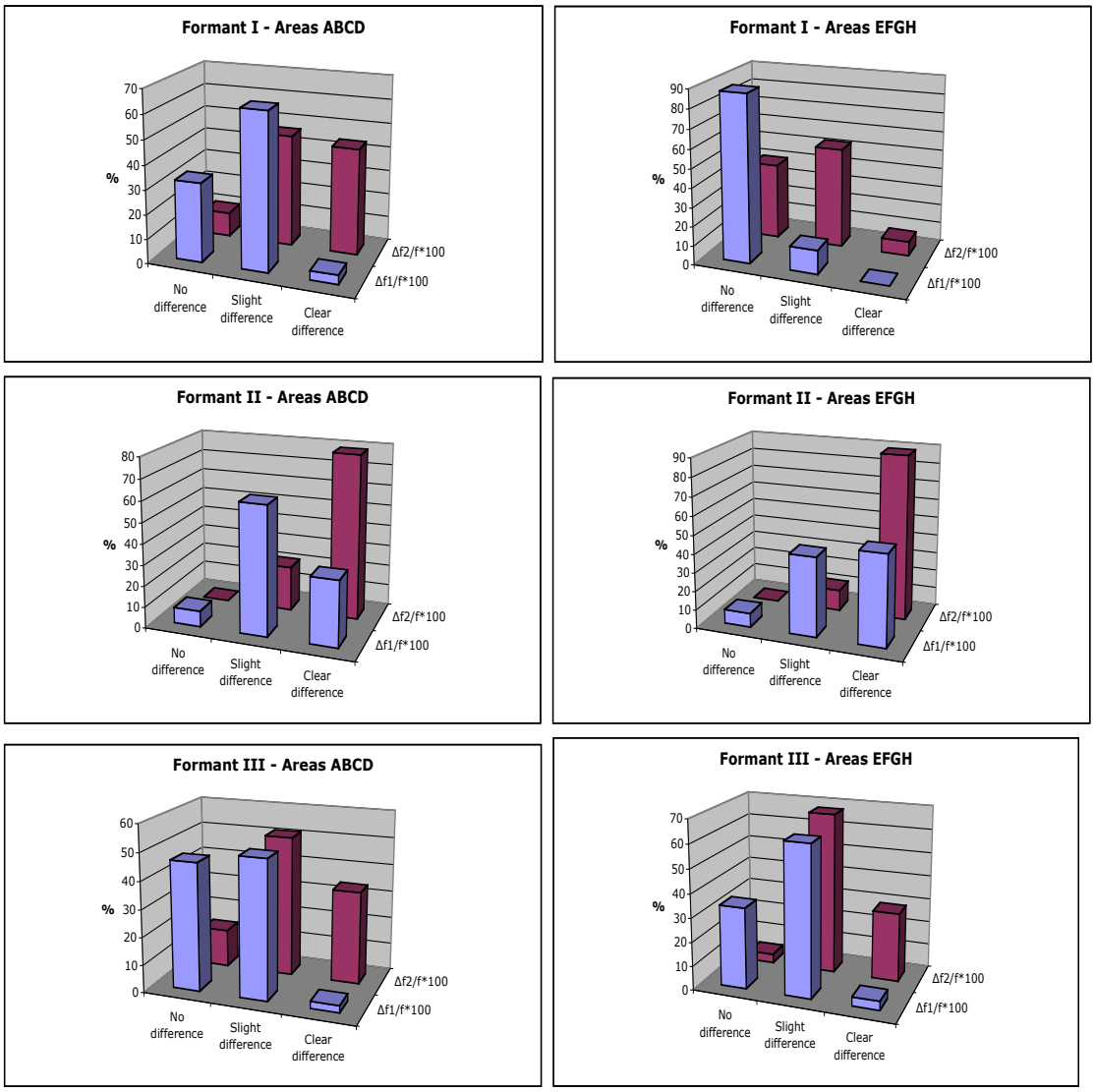


Figure 5. Breakdown of perceptual granularity listening tests by formant and area block (ABCD or EFGH), showing ratings for 'No difference', 'Slight difference' and 'Clear difference' for pairs separated by Δf_2 and Δf_2.

The following table summarises these results in order of perceptual saliency (where the lowest percentage indicates maximum perceptual saliency).

	Percentage of subjects hearing no difference
Formant II shift of Δf_1 in areas ABCD	7.5%
Formant II shift of Δf_1 in areas EFGH	7.5%
Formant I shift of Δf_1 in areas ABCD	32.5%
Formant III shift of Δf_1 in areas EFGH	33.75%
Formant III shift of Δf_1 in areas ABCD	46.25%
Formant I shift of Δf_1 in areas EFGH	87.5%

Figure 6. Percentage of subjects hearing no difference between pairs of sounds separated by Δf_1 , ranked in ascending order

We conclude from this data that the ability to discriminate timbral shifts varies quite considerably in the space; a shift of the second formant of Δf_1 in the 550-2200 Hz range is almost universally detected, whereas the same degree of shift of the first formant in the 110-220 Hz range is clearly under the perceptual threshold for most subjects.

Results of relative Euclidian distance perception tests

We turn now to consider the results of the tests where subjects were asked to consider which of two sounds more closely resembled a third.

The number of subjects giving 'correct' identifications, averaged out for all tests, was 15.7 out of 20 (79.38%). The probability of this result, based on a binomial distribution, is 0.0148; this is well below the five per cent level of statistical significance, and strongly suggests that subjects are, in general, able to perceive relative Euclidian distances in this particular timbral space. (However, see section on limitations at the end of the paper). The data was then broken down in the same way as described for the peceptual granularity tests i.e. by formant and area block (ABCD or EFGH). It was observed that the highest percentage of 'correct' identifications were made for those triplets which were aligned along the formant II axis, and that the lowest percentage occurred where the triplets were aligned along the formant I axis in areas E, F, G and H. This is summarised in figure 7.

Formant axis along which triplet aligned	Area block	Percentage of 'correct' identifications
II	EFGH	87.5%
II	ABCD	83.75%
III	ABCD	80%
I	ABCD	72.5%
III	EFGH	67.5%
I	EFGH	65%

Figure 7. Percentage of subjects making 'correct' estimations of relative distance within triplets, where the triplets are aligned along one axis, broken down by formant and area block (ABCD or EFGH) and ranked in descending order

The above pattern is reflected in the ratings for those triplets where the alignment was along two axes (figure 10). Here, the overall average success rate was higher, suggesting that the shift along two axes rather than one provided subjects with extra cues. There are some small discrepancies between the ranking order in figures 9 and 10, and it is suggested that there are some perceptual trade-offs where the alignment is along two axes. Nevertheless, overall, the highest percentage of 'correct' identifications are for those where, as before, the triplet is aligned along the formant II axis in area block EFGH, and the lowest where the triplet is aligned along the formant I axis, again in area block EFGH. However, please see section on limitations at the end of this paper.

Formant axes along which triplet aligned	Area block	Percentage of 'correct' identifications
I & II	EFGH	95%
II & III	ABCD	85%
II & III	EFGH	82.5%
I & III	ABCD	77.5%
I & II	ABCD	77.5%
I & III	EFGH	75%

Figure 8. Percentage of subjects making 'correct' estimations of relative distance within triplets, where the triplets are aligned along two axes, broken down by formant and area block (ABCD or EFGH) and ranked in descending order.

Comparison of results and discussion

Comparison of the two tables in figures 7 and 8 show suggests a strong correlation, and indicates that the ability successfully to perceive relative Euclidian distances in this timbral space is inversely related to the perceptual granularity of the space i.e. the size of the smallest timbral shift that can be perceived. However, please see section on limitations at the end of this paper.

The fact that shifts along the formant II (550-2200) axis, and, to a lesser extent, formant III (2310-6320 Hz) appear to be salient within this space can be explained by the particular sensitivity of the ear to frequencies in these regions. It should be emphasized that the particular perceptual topography revealed in this study apply only to this particular timbral space; no general claims are being made about the topography of any other timbral space, other than a speculative one that a link between Euclidian distance perception and the perceptual granularity may apply to all timbral spaces.

Further work is planned, firstly on analysis of the data, and secondly directed at using this data to inform the development of a search algorithm designed to navigate the user through the co-ordinate space described in this paper.

Finally, the listening tests have not, at the time of writing, been completed. The two tests conducted so far have been on two different sounds (pitched and noise-based), and there is a need to carry out further tests of Euclidian distance perception on noise based sounds located within the same co-ordinate space.

Limitations

Very late in the data analysis stage, an error was noted in the test data for the perception of Euclidian distances. The frequency ratio separating the sounds along the formant III axis was 1.26, whereas that for formants I and II was 1.41 . While this does not, in our view, invalidate hypotheses 1 and 2, it is likely to have distorted the ranking order given in figures 7 and 8, and must inevitably make these findings unreliable. In turn, this must put a question mark over hypothesis 3, and further work will need to be done to test this.

In addition, an error in compiling the test data resulted in ten out of the twenty responses for two out of the fifty six tests being invalid. The average number of 'correct' identifications for these two tests is out of the remaining ten responses, rather than the full twenty. However, we do not believe that, in this case, this materially affects the result.

Acknowledgments. The authors would like to express their thanks to the staff and students of London Metropolitan University who took part in this study.

References

Balzano, G.J., 1986. What are musical pitch and timbre? Music Perception, 3(3): 297-314.

Charbonneau, G., 1981. Timbre and the perceptual effect of three types of data reduction. Computer Music Journal, 5(2): 10-19.

Dissard, P. and Darwin, C.J., 2001. Formant frequency matching between sounds with different bandwidths and on different fundamental frequencies. J. Acoust. Soc of America, 110: 409-415.

Ehresman, D. and Wessel, D.L., 1978. Perception of Timbral Analogies, IRCAM, Paris.

Ethington, R. and Punch, B., 1994. SeaWave: A System for Musical Timbre Description. Computer Music Journal, 18:1: 30-39.

Flanagan, J.L., 1955. A Difference Limen for Vowel Formant Frequency. Journal of the Acoustical Society of America, 27: 613-617.

Flanagan, J.L., 1957. Estimates of the maximum precision necessary in quantizing certain dimensions of vowel sounds. J. Acoust. Soc. Am 29, 533-534.

Gagne, J.P. and Zurek, P.M., 1988. Resonance frequency discrimination. J. Acoust. Soc Am., 83: 2293-2299.

Gaver, W.W., 1993. Synthesizing auditory icons, ACM Interchi '93. ACM, pp. 228-235.

Grey, J.M., 1977. Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am, 61:5: 1270-1277.
Hermansky, H., 1987. Why is the formant-frequency difference limen asymmetric? J. Acoust. Soc. Am, 81: 533-534.
Horner, A., Beauchamp, J. and Haken, L., 1993. Machine Tongues XVI: Genetic Algorithms and Their Application to FM Matching Synthesis. Computer Music Journal 17:4 pp 17-29: 17-29.
Hourdin, C., Charbonneau, G. and Moussa, T., 1997. A Multidimensional Scaling Analysis of Musical Instruments' Time Varying Spectra. Computer Music Journal, 21:2: 40-55.
Johnson, C.G., 1999. Exploring the sound-space of synthesis algorithms using interactive genetic algorithms, AISB'99 Symposium on Musical Creativity, Edinburgh.
Kaminskyj, I., 1999. Multidimensional scaling analysis of musical instrument sounds' spectra, Australasian Computer Music Conference (ACMC), Wellington, NZ, pp. 36-9.
Kewley-Port, D. and Watson, C.S., 1994. Formant-frequency discrimination for isolated English vowels. J. Acoust. Soc. Am, 95: 485-496.
Lyzenga, J. and Horst, J.W., 1997. Frequency discrimination of stylized harmonic synthetic vowels with a single formant. J. Acoust. Soc. Am, 102: 1755-1767.
McAdams, S., 1999. Perspectives on the Contribution of Timbre to Musical Structure. Computer Music Journal, 23:3: 85-102.
McAdams, S. and Cunible, J.C., 1992. Perception of Timbral Analogies. Philosophical Transactions of the Royal Society of London - Series B - Biological Sciences: 383-389.
Mermelstein, P., 1978. Difference limens for formant frequencies of steady-state and consonant-bound vowels. J. Acoust. Soc. Am, 63: 572-580.
Miranda, E.R., 1995. An Artificial Intelligence Approach to Sound Design. Computer Music Journal, 19:2: 59-75.
Misdariis, N., Smith, B.K., Pressnitzer, D., Susini, P. and McAdams, S., 1998. Validation of a Multidimensional Distance Model for Perceptual Dissimilarities among Musical Timbres, 16th International Congress on Acoustics and 135th Meeting Acoustical Society of America, Seattle, Washington, 20-26 June 1998, pp. 3005-3006.
Mynatt, E.D., 1994. Designing with auditory icons., CHI '94. ACM, Boston, MA, pp. 269 - 270.
Nakagawa, T., Saito, S. and Yoshino, T., 1982. Tonal differences limens for second formant frequencies of synthesized Japanese vowels. Ann. Bull. Res. Inst. Logoped. Phoniatrics 16, 81-88.
Nord, L. and Sventelius, 1979. Analysis and prediction of difference limen data for formant frequencies. Phon. Exp. Res. Inst. Ling. Univ. Stockholm, 1, 24-37.
Plomp, R., 1970. Timbre as a Multidimensional Attribute of Complex Tones. In: R. Plomp and G.F. Smoorenberg (Editors), Frequency Analysis and Periodicity Detection in Hearing. Suithoff, Leiden, pp. 397-414.
Seago, A., Holland, S. and Mulholland, P., 2004. A Critical Analysis of Synthesizer User Interfaces for Timbre. In: A. Dearden and L. Watts (Editors), HCI 2004: Design for Life. British HCI Group, Leeds.
Slawson, W., 1968. Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. Journal of the Acoustical Society of America, 43: 97-101.
Toiviainen, P., Kaipainen, M. and Louhivuori, J., 1995. Musical timbre: similarity ratings correlate with computational feature space distances. Journal of New Music Research (Netherlands), 24: 282-298.