Proceedings of the Institute of Acoustics

MULTICHANNEL SPATIALIZATION TECHNIQUES FOR MUSICAL SYNTHESIS

J.M. Hirst University of Salford, School of Acoustics and Electronic Engineering, UK
W.J. Davies University of Salford, School of Acoustics and Electronic Engineering, UK
P.J. Philipson University of Salford, School of Acoustics and Electronic Engineering, UK

0. ABSTRACT

Until recently, spatial techniques for musical synthesis have usually been limited to stereo reproduction. With the advent of multichannel reproduction systems, techniques for creating more enveloping spatial effects have been investigated.

The spatializing techniques involved extracting the individual harmonics of a musical signal by means of narrow band pass filtering, assigning the harmonics into groups then distributing the groups to individual loudspeakers of a circular loudspeaker array. For comparison, the test signals were also presented through an ambisonic system. The techniques were subject to a psychoacoustic investigation by means of subjective testing.

The degree of spatial spread of the groups of harmonics around the listener was varied. The perceived degree of spatial impression was recorded for each variation by means of rank ordering. Results suggest that by spreading the groups of harmonics to an angle of 90° increased the perceived degree of perceived spatial impression. At angles above 90°, the perceived degree of spatial impression did not significantly increase. Conclusions are drawn on the utility of spatial reproduction systems and spatializing techniques for musical synthesis.

1. INTRODUCTION

At present, the audio outputs of a musical synthesizer are usually in the familiar stereo format. Stereo allows for the formation of phantom images in-between the loudspeakers, which gives rise to a limited perception of spaciousness. With the advent of cheaper digital storage, multichannel surround sound systems, which can produce a greater perception of spaciousness, have become more commonplace. Therefore, it is probable that synthesizers of the future will accommodate multichannel reproduction formats by increasing the number of audio outputs.

This report focuses on a non room reflection based multichannel spatializing technique for musical synthesis that involves decomposing a complex musical signal into it's individual harmonics, then spatially spreading the harmonics over a circular loudspeaker array. This presents the auditory system with a potential perceptual conflict. Due to the spatial separation of the signal, a number of sources may be localised, however, the harmonic relationship and synchronous onsets of the signals provides the auditory system with grouping cues.

The remainder of this paper includes a discussion concerning the theory behind the techniques followed by a description of a psychoacoustic subjective test that assessed the effectiveness of the techniques utilising both real and virtual sources.

2. THEORY

2.1 Localization

Localization, in terms of human spatial hearing, refers to the relationship between the physical location of a sound, (the sound event) and the perceived location of the sound event, the (auditory event) [1]. The auditory system utilises interaural time and level differences, which arise due to the path difference resulting from a sound event arriving at each ear, to localise a sound. By considering this auditory cue in isolation, it would be expected that the spatializing technique would result in the perception of a number of individual auditory events. As the spatializing technique presents the auditory system with multiple, harmonically related and temporally coincident sound events, localization cannot be considered as the only cue available to the auditory system.

2.2 Perception of Complex Tones

The perception of complex tones has been widely reported and is well summarised by Moore [2]. Moore reports on the hearing mechanism's ability to fuse complex tones, which consist of a number of harmonically related partials, into a single percept with a pitch equal to the fundamental frequency of the complex tone. Moore goes on to describe Schouten's work involving 'Residue pitch' or 'Fundamental tracking'. If the fundamental harmonic of a complex tone is removed, the perceived pitch of signal does not alter. Similarly, if all but few mid-frequency harmonics are removed, the perceived pitch remains the same, however the timbre of the signal is greatly changed.

Two main theories have been proposed to account for the phenomenon of pitch residue. Temporal theories propose that the pitch of a complex tone is related to time intervals between nerve firings emanating from a position on the basilar membrane where two partials are exciting the same critical band. Pattern recognition theories suggest that the complex tone is frequency analysed, then 'matched' to a pitch percept relating to the fundamental frequency of the matched pattern.

The fusion of complex tones and fundamental tracking suggest that harmonically related partials provide the auditory system with a strong grouping cue.

2.3 Auditory Scene Analysis

In a natural environment, the acoustic energy produced by a number of concurrent sound sources arriving at the ears of a listener is mixed. The auditory system first analyses this mixture into a large number of frequency components. As extensively reported by Bregman [3], the problem addressed by auditory scene analysis is which combination of frequency components should be attributed to each sound source? The analysis, which is dependent upon a number of cues, results in the perceptual fusion or segregation of sounds. The perceptual fusion or segregation of simultaneous components depends upon similarities or differences in harmonic relationships, regularity of spectral spacing, onset and offset synchrony, binaural frequency matching, parallel amplitude modulation, frequency, spectral envelope, amplitude and spatial location.

2.3.1 Interaction of Cues

Various auditory cues compete to form the perceptual grouping of sounds, however, cues do not operate in isolation. Interactions occur, with some cues reinforcing each other whilst other cues compete with

each other. Of interest to this study are the interactions of onset synchrony, harmonicity and localisation cues. By means of the rhythmic masking release paradigm, Turgeon [4] examined the interaction of these cues using spatially separated concurrent complex tones of a short duration. Her main findings were that temporal (onset) synchrony strongly contributes to sound source grouping, whilst spatial separation and harmonicity contributed only weakly or not at all to the perceptual organisation of sounds. This was in partial agreement with earlier work [5] that suggested that harmonic structure is more important than commonality of spatial position for the grouping of complex sounds.

Regarding spatial impression, it is interesting to note that Turgeon also reports that whilst sounds coming from different locations in space can be perceived as a single event, they are difficult to locate and can be described as 'diffuse'.

2.4 Experiments Involving Spatially Separated Complex Tones

Previous investigations involving the spatial separation of harmonic signals have concentrated mainly upon psychoacoustics rather than the creation of spatial impression. Bregman cites a few examples of similar experiments. An unreferenced example [6] involves a demonstration of how the hearing system fuses harmonically related signals. Two sets of partials of a synthesized speech sound, occupying different regions of the frequency spectrum were presented to different ears. When the two sets of partials shared the same fundamental frequency, the signal was perceived as being fused, when the sets of partials did not share the same fundamental frequency, two separate signals were perceived.

Another example [7] involved a sound pattern created for a piece of music by Reynolds and Lancino at IRCAM. An oboe tone was synthesized with the odd and even harmonics separated into two channels, which fed two loudspeakers on the left and right of the listener. Frequency micromodulation was applied to both signals. When the frequency fluctuations of the harmonics were synchronized, a single oboe was perceived in between the loudspeakers. When the fluctuations of the harmonics presented in the left loudspeaker were gradually made independent of those in the right loudspeaker, two separate sounds were perceived, one from each loudspeaker.

Bregman also describes an informal experiment he performed with Divenyi [8]. Two harmonic signals were created, one consisting of tones at frequencies of 200, 400, 600 and 800 Hz, the other at frequencies of 300, 600, 900, and 1200 Hz. The signals were presented through headphones, with one signal panned 45° to the left and the other, 45° to the right. The signals were played at irregular intervals so that they overlapped for some of the time, but did not start or finish at the same time. The experimenters expected that in addition to the two complex tones, a third tone at 600 Hz (common to both complex tones) would be perceived at a central location. However, only the two complex tones were perceived. Bregman suggests that this was due to the 600 Hz tones always going on or off in synchrony with one of the complex tones thus accounting for the assignment of the 600 Hz tones to both complex tones, simultaneously.

2.5 Summary

These experiments seem to suggest that unless there is some correlation between the spatially separated components of a harmonically related signal, fusion of the components will not occur. Harmonically related signals with synchronous onset times (such as the signals used in this experiment) provide the auditory system with a strong grouping cue even if the signals emanate from different spatial locations. In terms of spatial impression, a spatially separated signal is difficult to locate and can be described as diffuse. Therefore, it could be expected that the spatializing techniques discussed in this paper would result in the perception of a fused sound, which is spatially diffuse.

3 EXPERIMENTAL METHOD

3.1 Introduction

The premise behind the test design was that firstly, the techniques would deliver a spatial effect and secondly, as the spatial spread of harmonics common to a complex musical signal was increased around the listener, the perceived degree of spatial impression would also increase.

The test signals were decomposed into individual harmonics then spatially spread around the listener in increasing steps. The subjects were asked to rank order, in terms of spatial sound quality, four auditions of varying spatial spread and one audition comprising of the original signal presented through all eight loudspeakers. The procedure was repeated using ambisonic reproduction.

3.2 Program Material

Two standard format (16 bit, 44.1 kHz) stereo samples were used in the test, both of which were downloaded from the Internet [9]. The samples were of a 4.26s, G4 (f_0 = 392 Hz) string ensemble and a 4.10s, C4 (f_0 = 261 Hz, with a 130 Hz sub-harmonic also present) synthesizer sound. Using a sample editor (Cool Edit Pro), the samples were converted into mono then narrow band pass filtered, using a Butterworth sixth order filter, to extract each harmonic. This yielded 28 harmonics for the string ensemble and 40 for the synthesizer sound. The harmonics were then assigned into groups.

3.3 Rank Order Arrangement

The assignment of harmonics to each group was dependent upon how many loudspeakers were being used in a particular audition. For example, the synthesizer sound consisting of 40 harmonics was split into eight groups of five harmonics for an eight-loudspeaker (360° spread) audition and five groups of eight harmonics for a five-loudspeaker (180° spread) audition. The assignment of harmonics to loudspeakers, for each audition are shown in *Tables 1 and 2*. For all auditions, the fundamental harmonic was assigned to the loudspeaker directly in front of the listening position.

Each presentation compared five auditions. Four of the auditions consisted of a decomposed signal with angular spreads of 0°, 90°, 180° and 360°. The other audition consisted of the original (not decomposed) mono signal, simultaneously replayed through all eight loudspeakers (henceforth referred to as mo8). To negate any bias introduced by differences in perceived loudness, the overall level of this audition was adjusted to be of the same perceived loudness as the decomposed auditions. This procedure was performed by the experimenter and confirmed by one of the subjects. With two source materials and two reproduction methods, this resulted in four rank order presentations of five auditions.

For the ambisonic auditions the harmonics were assigned to the same groups as for the real sources then positioned around the listener, at the same angular positions as the real sources using the ambisonic encoding process.

3.4 Subjects

Twelve subjects, seven males and five females, participated in the experiment all of whom were either staff or students of The School of Acoustics and Electronic Engineering, University of Salford. The majority of the subjects had previously participated in other listening tests. All subjects attended a training session that involved an introduction to multichannel spatial audio and a trial run of the test procedure.

3.5 Test Room Configuration

The experiment was carried out in the anechoic chamber of the School of Acoustics and Electronic Engineering, University of Salford. The working dimensions of the room were measured as 3.6m in height by 5.5m in length by 3.2m in width. The inner chamber is lined with 0.9m long fibreglass wedges to give a cut-off frequency below 100 Hz.

Eight loudspeakers, arranged in a circular array, were attached to an octagonal metal frame, with the listening position in the centre. The loudspeakers were placed at a distance of 1.41m from the listening position and at an angular spacing of 45°. An acoustically transparent curtain was hung between the listening position and the loudspeakers to facilitate blind testing. A computer keyboard (which acted as a switching mechanism), a loudspeaker and microphone (to enable communication between the subjects and the experimenter), were also present in the chamber. A diagram of the test room configuration is shown in *Figure 1*.

3.6 Equipment Configuration

The test signals were recorded into a computer based audio sequencer (Cubase VST), the digital output of which (via a multichannel soundcard) was connected to an Alesis ADAT to allow for digital to analogue conversion. Balanced outputs from the ADAT fed the eight Genelec 1029A loudspeakers that were level aligned using pink noise and a sound level meter. By using the 'Cue Point' feature in Cubase and a computer keyboard acting as a remote control, the subjects could switch between each of the five auditions of a rank order presentation at will, thus enabling quick comparisons.

For ambisonic playback a square, pantaphonic four loudspeaker configuration was used. This utilised loudspeakers 8,2,4 and 6 (See *Table 1*). A diagram of the equipment configuration is shown in *Figure 2*.

3.7 Experimental Procedure

The blindfolded subjects were escorted into the anechoic chamber, seated at the listening position then un-blindfolded. For each of the four presentations, consisting of five auditions each, the subjects were asked to rank order the auditions in terms of spatial sound quality (1 = lowest rank, 5 = highest rank). In evaluating the spatial sound quality the subjects were asked to 'Consider all aspects of spatial sound reproduction. This might include the locatedness or localisation of he sound, the width of the sound, how enveloping it is or it's naturalness and depth' [10].

The subjects could freely switch between auditions and could take as long as they needed to determine the rank order. On average, the test took approximately twenty minutes to complete. When the rank order had been determined, the subjects verbally relayed their choice to the experimenter via a microphone.

4 RESULTS

During the training session the subjects performed a trial rank ordering of the synthesizer sample. The data collected from this were correlated with the actual test data. Consequently data collected from two of the subjects was rejected due to low correlation coefficients of 0.2 and 0.

For each of the four presentations the data collected from the remaining ten subjects was subject to the non-parametric Friedman analysis of variance test [11]. The analysis showed that the preference ranks for all four sets of data differed significantly at the $p < 0.01$ level (See *Table 3*). Graphs depicting the mean score for each rank ordering and the overall mean spatial rank can be seen in *Figures 3 to 7*. Overall, a spatial spread of 180° delivered the greatest degree of spatial impression followed by 360°, 90°, mo8 and 0°.

Having established that preference ranks for each presentation differed significantly, the least significant rank difference (LSRD) for the Friedman test [12] was calculated. This test determined which auditions were ranked significantly higher or lower in preference from one another for each presentation. The results can be seen in *Table 4*. In brief, the spatial spread of 0° was consistently ranked significantly lower than the 90°, 180° and 360° auditions. For the ambisonic synthesizer presentation, the 180° audition was ranked significantly higher than all auditions, apart from the 360° audition which was ranked significantly higher than the mo8 and 0° auditions. For the remaining presentations, the ranked differences between the 90°, 180° and 360° auditions were not significant.

5. DISCUSSION

From the LSRD test results shown in *Table 4* it can be seen that a spatial spread of 0° was ranked significantly lower than all other spatial spreads in all four presentations. Apart from one presentation, there was no significant difference between the spatial spreads of 90°, 180°, and 360°. This suggests that whilst the techniques deliver a spatial effect, the degree of spatial impression does not increase further as the spatial spread of harmonics is extended beyond 90°. Possible reasons for the differences between the 90°, 180° and 360° auditions being insignificant are that the subjects may have experienced difficulties in discriminating between the auditions due to their inexperience in listening to spatial audio and the need for more accurate descriptions of 'spatial impression' and experimental methods [10].

As similar results were found for both program materials and both reproduction methods, the techniques appear to be robust. For the ambisonic presentations, the results were very similar. In both presentations spatial spreads of 180° and 360° were ranked significantly higher than 0° and mo8. For one of the ambisonic presentations (Synthesizer), spatial spreads of 180° and 360° were ranked significantly higher than all other auditions. This may suggest that for virtual reproduction, extending the spatial spread beyond 90° results in an increase in the perceived degree of spatial impression.

6. CONCLUSIONS

The multichannel spatialization techniques for musical synthesis, which involved decomposing a complex musical signal into it's individual harmonics, then spatially spreading the harmonics over a circular loudspeaker array were subject to a psychoacoustic preference test by means of rank ordering. To summarise, the results suggest the following:

- The Friedman test has shown that the results are statistically significant and meaningful.
- A harmonic spatial spread of 90° resulted in a significantly higher degree of perceived spatial impression than a spatial spread of 0° for all presentations.
- In all but one presentation, increasing the spatial spread beyond 90° did not significantly increase the perceived degree of spatial impression.
- The technique appears to be robust as the results were found to be similar for both real and virtual presentations.
- For virtual reproduction, increasing the spatial spread beyond 90° may further increase the perceived degree of spatial impression.

7. FURTHER WORK

Further work may entail continued subjective testing in order to establish greater confidence in the techniques. In particular, repeating the test using only two loudspeakers in the standard stereo configuration. Other areas involve developing the techniques to optimise spatial impression (investigating the grouping and positioning of the harmonics and creating asynchronous onsets by introducing short time delays to groups of harmonics) and developing a method to objectively measure the degree of spatial impression delivered by the technique.

8. REFERENCES

[1] J.Blauert, "Spatial hearing" Revised Edition, pp 2–5, MIT Press (1997)

[2] B.C.J.Moore, "An Introduction to the Psychology of Hearing", Fourth Edition, pp 749-761, Academic Press (1997)

[3] A.S.Bregman, "Auditory Scene Analysis", pp 641-674, MIT Press (1999)

[4] M.Turgeon, "Cross-Spectral Auditory Grouping Using the Paradigm of Rhythmic Masking Release", Unpublished doctoral thesis, McGill University, Montreal (1999)

[5] T.N.Buell and E.R.Hafter, "Combination of Binaural Information Across Frequency Bands", J. Acoust. Soc. Am., **90**, pp 1894-1900 (1991)

[6] A.S.Bregman, "Auditory Scene Analysis", p 247, MIT Press (1999)

[7] A.S.Bregman, "Auditory Scene Analysis", p 296, MIT Press (1999)

[8] A.S.Bregman, "Auditory Scene Analysis", p 623, MIT Press (1999)

[9] URL: http://www.samplenet.co.uk, (2000)

[10] J.Berg and F.Rumsey, "Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods", in *Proceedings of the AES 16th International Conference*, pp 51-66, Audio Eng. Soc. (1999)

[11] H.T.Lawless and H.Heymann, "Sensory Evaluation of Food", pp 694-697, Chapel and Hall (1998)

[12] H.T.Lawless and H.Heymann, "Sensory Evaluation of Food", pp 446-449, Chapel and Hall (1998)

9. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Martine Turgeon of IRCAM for her assistance in the preparation of this paper, Francis Wooff, Technician with the School of Acoustics and Electronic Engineering, University of Salford for his assistance in constructing the loudspeaker framework and the subjects for their time and patience.

Angular spread of loudspeakers (degrees)	Harmonics assigned to each loudspeaker							
	LS1	LS2	LS3	LS4	LS5	LS6	LS7	LS8
0	All							
90	2 5 8 11 14 17 20 23 26 29 32 35 38	3 6 7 12 13 18 19 24 25 30 31 36 37						1 4 9 10 15 16 21 22 27 28 33 34 39 40
180	2 9 12 19 22 29 32 39	3 8 13 18 23 28 33 38	4 7 14 17 24 27 34 37				5 6 15 16 25 26 35 36	1 10 11 20 21 30 31 40
360	2 15 18 31 34	3 14 19 30 35	4 13 20 29 36	6 11 22 27 38	8 9 24 25 40	7 10 23 26 39	5 12 21 28 37	1 16 17 32 33

See diagram below for loudspeaker numbering and angular spread. Harmonic 1 refers to a sub-harmonic, Harmonic 2 refers f_0, Harmonic 3 refers to 2 x f_0 etc.

Angular Spread (Degrees)	Loudspeakers Active
0	1
90	8, 1 and 2
180	7, 8, 1, 2 and 3
360	All

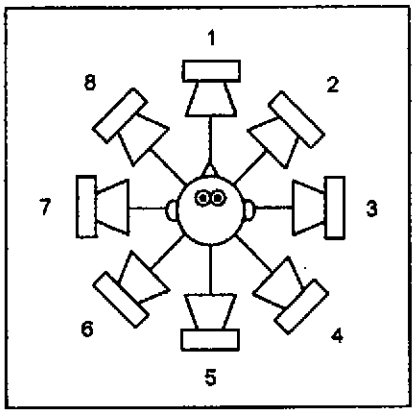

Table 1. Assignment of Harmonics to Loudspeakers (Synthesizer)

Angular spread of loudspeakers (degrees)	Harmonics inputted to each loudspeaker							
	LS1	LS2	LS3	LS4	LS5	LS6	LS7	LS8
0	All							
90	1 4 9 10 15 16 21 22 27	3 6 7 12 13 18 19 24 25 28						2 5 8 11 14 17 20 23 26
180	1 10 11 20 21	3 8 13 18 23 28	4 7 14 17 24 27				5 6 15 16 25 26	2 9 12 19 22
360	1 16 17	3 14 19	4 13 20	6 11 22 27	8 9 24 25	7 10 23 26	5 12 21 28	2 15 18

See diagrams below for loudspeaker numbering and angular spread. Harmonic 1 refers to f_0, Harmonic 2 refers to $2 \times f_0$ etc.

Angular Spread (Degrees)	Loudspeakers Active
0	1
90	8, 1 and 2
180	7, 8, 1, 2 and 3
360	All

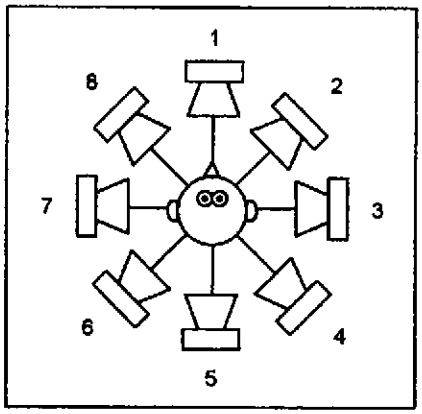

Table 2. Assignment of Harmonics to Loudspeakers (String)

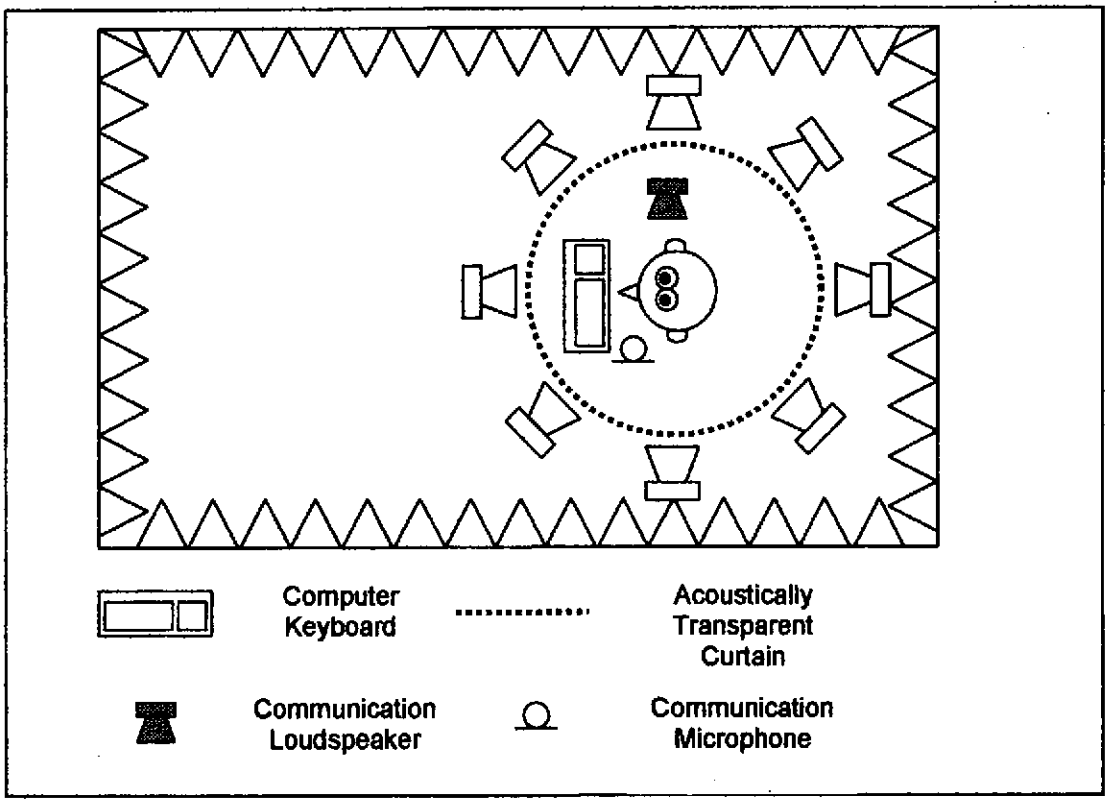

Figure 1. Test Room Configuration

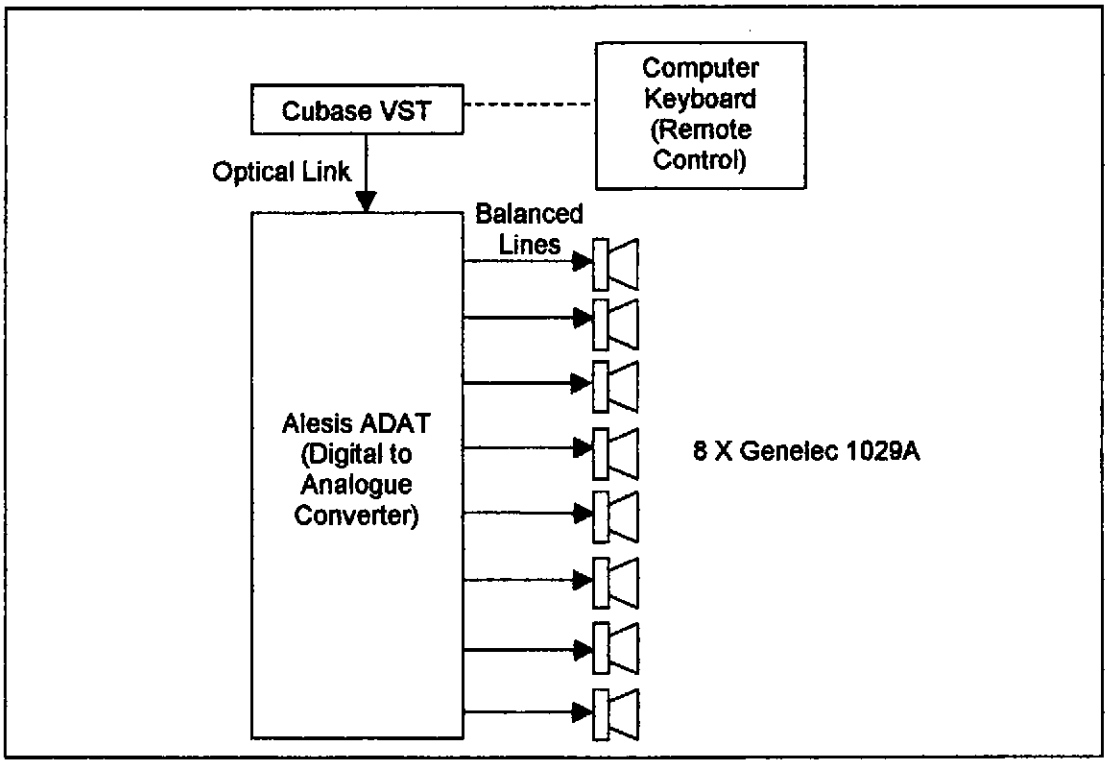

Figure 2. Equipment Configuration

Source Material	Friedman Test Value	Significant	p <
Synth	16.18	Yes	0.01
Ambisonic Synth	28.16	Yes	0.001
String	21.04	Yes	0.001
Ambisonic String	27.90	Yes	0.001

Table 3. Friedman Test Analysis of Data

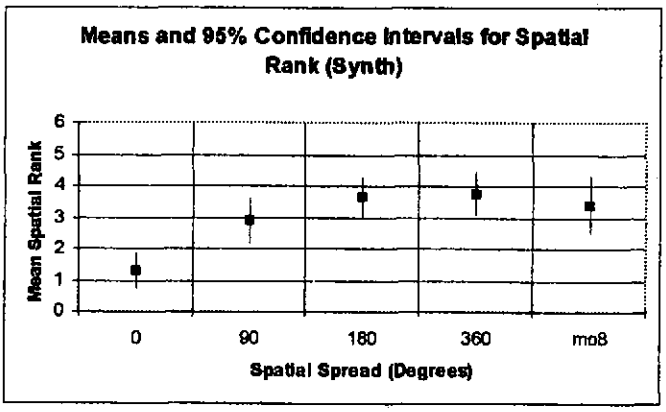

Figure 3. Mean Rank and 95% Confidence Intervals for Synthesizer Presentation

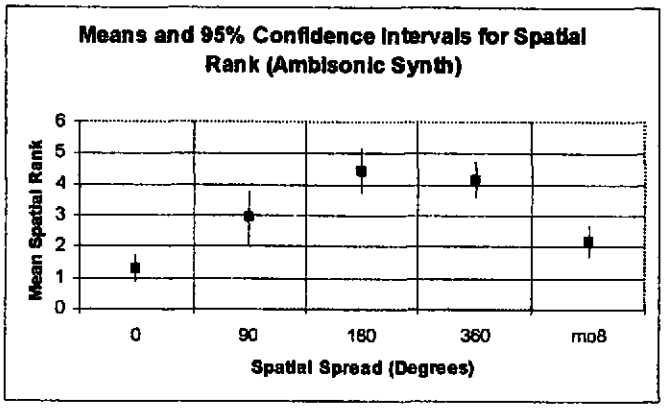

Figure 4. Mean Rank and 95% Confidence Intervals for Ambisonic Synthesizer Presentation

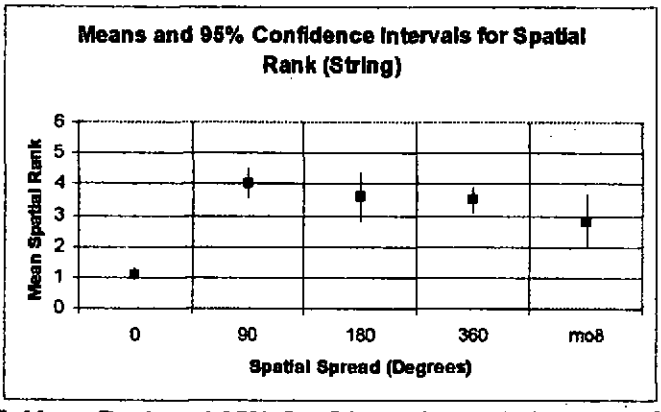

Figure 5. Mean Rank and 95% Confidence Intervals for String Presentation

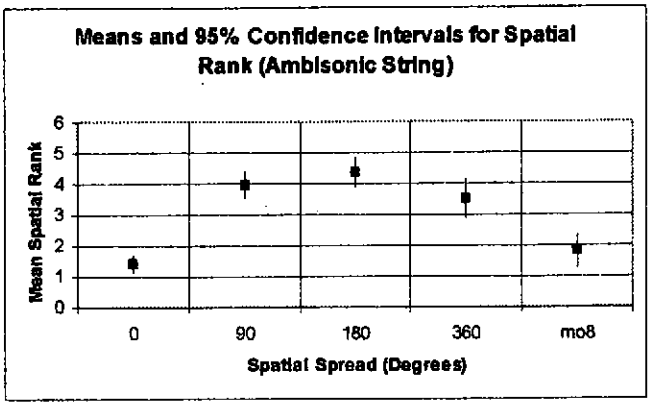

Figure 6. Mean Rank and 95% Confidence Intervals for Ambisonic String Presentation

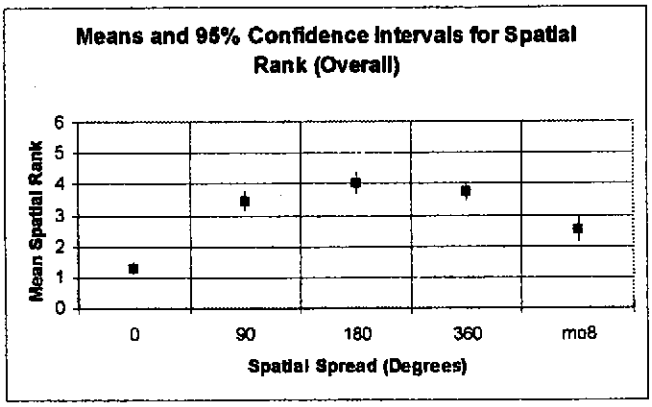

Figure 7. Mean Rank and 95% Confidence Intervals for Ambisonic String Presentation

	Spatial Spread (Degrees)				
Presentation	0	90	180	360	Mo8
Synth Rank Total	13	29	36.5	37.5	34
Significance group	a	b	b	b	b
Ambisonic Synth Rank Total	13	29.5	44.5	41.5	21.5
Significance group	a	bc	d	cd	ab
String Rank Total	11	40	36	35	28
Significance Group	a	b	b	b	b
Ambisonic String Rank Total	14	39.5	43.5	35	18
Significance Group	a	b	b	b	a

Table 4. Least Significant Rank Difference for the Friedman Test. (LSRD = 13.86). (Spatial spreads sharing the same significance group letter show no difference in ranked preference)