RESULTS OF A ROUND ROBIN SUBJECTIVE EVALUATION OF VIRTUAL HOME THEATRE SOUND SYSTEMS

NICK ZACHAROV† AND JYRI HUOPANIEMI‡

Nokia Research Center, Speech and Audio Systems Laboratory, Finland.
†Nick.Zacharov@Nokia.com
‡Jyri.Huopaniemi@Nokia.com

Virtual home theatre systems provide an efficient means of multichannel sound reproduction with two loudspeakers. Six commercial and academically developed systems have been collected and subjectively evaluated under controlled conditions against discrete 5-channel reproduction. Tests have been carried out in two test rooms. The experimental design and the detailed analysis of results are presented in this paper.

INTRODUCTION

This paper is concerned with the quality evaluation of multichannel sound reproduction using cross-talk cancelled binaural technologies. As the development of high quality multichannel audio encoding, storage media and broadcast techniques evolves, the availability of multichannel audio has become widespread. While matrixed multichannel methods (e.g. Dolby Surround) have been widely available for some time, the benefits of discrete multichannel reproduction are finding increasing favor in professional and consumer markets, for example, with new audio-visual storage formats such as DVD.

The so-called *virtual surround* or *virtual home theater (VHT)* aim to faithfully reproduce the spatial sound qualities of 5.1 channel audio systems using only two loudspeakers or headphones. Virtual loudspeaker technology enables the virtualisation of surround left and right channel information. Thus the surround channels appear to emanate from virtual sources placed outside the physical stereo setup of the two loudspeakers. Virtualisation can also be created for the left and right signals in order to, for example, expand the stereo base. Furthermore, virtualisation may be applied for the center channel, or it can be left unprocessed (creating a traditional phantom image)[1].

The round robin[2] experiment reported here was designed to formally evaluate the relative performance of state of the art virtual home theatre algorithms. Both commercial and academic algorithms are tested and compared against each other and also against a discrete 5-channel system. Whilst the latter is now widely available, no assumptions regarding its relative quality is assumed for this test.

[1] It should be noted that all virtual surround or virtual home theater concepts discussed in this experiment refer to 2-loudspeaker reproduction of 5.1 audio material in such a way that at least the surround (and possibly also the center, left and right) audio channels are processed virtually.

[2] *Written petition with signatures radiating from circles to conceal the order of writing* - The Oxford English dictionary.

0.1. Background

The theory of presenting crosstalk-compensated binaural information over a pair of loudspeakers was first formulated by Bauer [1], and Schroeder and Atal almost 40 years ago [2, 3]. Schroeder and Atal described the use of a crosstalk cancellation filter for converting binaural recordings made in concert halls for loudspeaker listening. Their impressions of listening to loudspeaker reproduced dummy-head recordings were "nothing less than amazing". However, they observed the limitations of the listening area, the "sweet spot", which has remained a difficult problem and limitation in loudspeaker binaural reproduction ever since.

The psychophysical and acoustical basis for manipulating stereophonic signals has been studied in, e.g., [1, 4]. Damaske studied loudspeaker reproduction issues and formulated the basic theories further in the TRADIS project (True Reproduction of All Directional Information by Stereophony) [5]. He conducted studies on sound image quality deterioration as a function of listener placement. The cross-talk canceling theory was refined by Cooper and Bauck [6]. They applied a well-known shuffler structure to cross-talk cancellation, and a new term to cross-talk canceled binaural presentation was introduced: *transaural stereo*. The transaural stereo concept originally applied shuffler structures and simplified head models for cross-talk canceling. These techniques have been further developed by, for example, [7, 8, 9] and by Cooper and Bauck [10] to include improved head models, varying loudspeaker schemes, and more sophisticated signal processing techniques (see [11] for a review of signal processing techniques for cross-talk canceling). Recently, concepts of using closely spaced loudspeakers to generate virtual sources have been introduced [12, 13]. The close spacing of speakers has been shown to be robust to head movements and to exhibit a wide sweet spot [14, 15].

Commercial interest in the field of cross-talk canceled binaural processing has grown very rapidly in the past few years. Studio and multimedia systems have in recent years adopted the use of positional 3-D audio and stereo widening for many applications, ranging from digital mixing consoles to computer sound cards, audio software, games, virtual environment simulation and telecommunications [16]. However, it has been the rapid emergence of discrete multichannel audio formats and associated technology that in recent years has boosted the research and industry interest in virtual and 3-D sound.

0.2. Methods for Virtual Surround Processing

Virtual surround processing can be divided into two parts:

- 3-D source positioning
- Cross-talk canceling

These procedures can be carried out in cascade so that the binaural filtering is carried out first and after that the cross-talk canceling network is applied. Another alternative is to integrate both the positional and cross-talk canceling filtering into the same process (see e.g. [17] for discussion on binaural format conversions). The latter approach is often called "virtual loudspeaker" processing [6].

An example of 3-D source positioning is illustrated in Figure 1. In the figure, a monophonic time-domain signal $x_m(n)$ is filtered with two HRTF or room impulse response approximations $H_l(z)$ and $H_r(z)$ to create an image of a single virtual source.

In loudspeaker listening, the signals are exposed to *crosstalk* as seen in Figure 2. The inversion of direction-dependent loudspeaker-to-ear transfer functions $H_{ll}(z)$, $H_{lr}(z)$, $H_{rl}(z)$, and $H_{rr}(z)$ has to be carried out in order to deliver the binaural signals to the ears of the listener in the same manner as in headphone binaural listening. As previously discussed, the filtering can here be understood as a cascaded process, in which HRTF filters are designed and implemented separately from the crosstalk

canceling filters. Another alternative is to combine these processes and directly design virtual speaker filters.

1. EXPERIMENTAL DESIGN AND PROCEDURE

This section will briefly summarise the experimental procedure and setup for the experiment performed in the two sites. At each site a different listening room and listening panel were employed as described below. In all other respects the experiments performed at each site were identical. For a full description of the design, the interested reader is referred to [18]

For the statistical analysis, the null hypothesis, H0, for these experiments can be defined as follows: no difference exist between compared VHT systems in-terms of spatial or timbral sound quality.

1.1. Listening rooms

The first of the two experiments was performed in the new Nokia Research Center (NRC), Speech and Audio Systems Laboratory listening room, which is fully conformant with ITU-R BS.1116-1 [19] as illustrated in Figure 3.

The second experiment was performed at the AES 16^{th} International Conference in a smaller meeting room as illustrated in Figure 4. The sparsely furnished room provided symmetrical geometry, with the exception of the door alcove.

1.2. Reproduction system

Genelec 1030A speakers were employed for all of the experiments consisting of a two-way system with a frequency response of 52 Hz – 20 kHz (-3 dB). Units were anechoically measured to ensure that they are well matched. The setup for the 5-channel reproduction was in accordance with ITU-R BS.775-1 [20] as far as possible and thus speakers were placed at the normal $0°$, $\pm 30°$, $\pm 110°$ angles, with the speaker's axis at average ear height. To allow proponents of virtual home theatre systems some flexibility, additional speakers were setup at an angle of $\pm 5°$. Loudspeakers were be placed behind an acoustically transparent screen to limit any bias effects. As only discrete 5-channel material was employed for this experiment, no low frequency energy (LFE) channel or loudspeaker was available.

1.3. Listeners

In the first, NRC, experiment a mixed panel of listeners was employed including experienced and trained listeners from the permanent NRC listening panel. These listeners have been selected for their listening capabilities using the GLS procedure [18, 21] and trained with tools such as the timbral ear training system [22, 23]. In total 15 persons took part (7 members of the permanent listening panel and 8 members of the NRC Speech and Audio Systems Laboratory). Whilst all members have listening experience, they cannot be categorized as "expert" in accordance to the ITU-R BS.1116-1 for this task. Listeners did not undergo any training or familiarisation prior to the experiment, nor was any post evaluation of listener performance performed.

For the AES16 experiment delegates of the conference were invited to participate. In total 29 persons completed the test, six of which were also proponents of VHT systems. In the final analysis, only two corrupt (incomplete) data sets were dropped. All listeners are considered naive in this case due to the lack of knowledge of their experience, hearing capabilities, etc. Listeners did not undergo any training or familiarisation prior to the experiment, nor was any post evaluation of listener performance possible.

1.4. Program material

Four program items were selected for this experiment. These items were selected to provide a range of spatial sound cues and different types of program.

- Blue bell railway [24], BBC. Scene consisting of steam training pulling away from station and approaching bridge. Contains country atmosphere with directional cues, and panning effects.
- Tipsy Gypsy, BBC. A concert at the Albert hall, consisting of audience cheering, applause and the conductor talking.
- Rain storm. Sample consisting of a thunder roll followed by the sound of hard rain [25].
- Felix Mendelsohn-Bartoldy, Symphony No. 4. Live recording at the Neues Gewandhaus, Leipzig, Deutsche Telekom, 1993 [25].

1.5. VHT systems

Seven systems in total were tested of which one was a discrete 5-channel setup. No restrictions were given for characteristics of the *virtual home theatre* systems in terms of processing requirements, filter lengths, additional processing, etc. However, in the interests of further understanding these systems relative performance, proponents were asked to provide information regarding the algorithms provided. A summary of these details is provided in Table 1.

Each tested system was given an arbitrary index number to be employed in the analysis and the presentation of results. The index number was different for each experiment due to the differing number of participating systems.

1.6. Test paradigm and psychometric task

A rank order procedure ([26], pp. 691–700) was employed for the test and implemented on GP2 [27]. All experiments were completely double blind, in that neither the listeners nor the experimenter knew the order of presentation of the systems. The presentation order only became clear to the experimenter at the data analysis stage.

Two grading scales were employed to evaluate the perceived spatial and timbral quality of reproduction of the seven systems under test, as illustrated in Figure 5. The following questions were posed to each listener.

- Rank order the samples by spatial sound quality (1 = lowest rank, 7 = highest rank)

  When evaluating the spatial sound quality, please consider all aspects of spatial sound reproduction. This might include the locatedness or localisation of the sound, how enveloping it is, it's naturalness and depth [28].

- Rank order the samples by timbral quality (1 = lowest rank, 7 = highest rank)

  When considering aspects of the timbre quality, please consider timbre as a measure of the tone colour. Timbre can be considered as the sensory attribute which allows two similar sounds, of the same pitch and loudness, to be different, for example a clarinet and a cello. Any audible distortions can also be considered as an aspect of the timbral quality.

The comparable quality of the virtual home theatre systems under test could not be prejudged. For this reason, ties were allowed. All systems must be ranked to complete the test.

1.7. Calibration

To ensure that all systems were evaluated on an equal basis a loudness calibration was developed. In brief, a 5 channel decorrelated set of pink noise signals was provided to each proponent and filtering by their VHT algorithm. Once filtered, the diffuse field Zwicker loudness [29] of this signals, replayed via all channels, was aligned to be equal.

2. RESULTS AND DISCUSSION

As the first step towards a thorough statistical analysis a suitable model must be applied and its validity tested. In an effort to analyse the data in an in-depth manner, the analysis of variance (ANOVA) model was considered. As the data is categorical in nature, it cannot be assumed that the ANOVA model can be directly applied due to the fact that an interval scale has been employed. To study the validity of applying the ANOVA model, some basic model assumptions were tested and compared against non-parametric methods (i.e. Kruskal-Wallis test). Prior to this the data was test for whether it conforms with the basic ANOVA assumptions for normal distribution of sampling, homogeneity of variance across groups and the normal distribution of residuals. These tests were applied for each dependent variable (spatial and timbral) and for each experiment. The results are summarised in Table 2, with the last column indicating whether the assumption are met. In all cases they are, with only one marginally significant aspect.

The comparisons of the one-way ANOVA and the one-way non-parametric test are shown in Tables 3 and 4 for each factor (SYSTEM, SUBJECT and PROGRAM). Whilst the level of the statistics for both tests are not identical, it can be said that the rank order of the factors is the same, as well as the significance levels.

Based upon these findings, the ANOVA model was considered a reasonable model for the data and it was used to evaluate the data in greater depth, as discussed below.

2.1. NRC experiment

Having illustrated the applicability of the ANOVA model in principle, a full analysis was performed for each dependent variable, the results of which can be seen in Tables 5 and 6.

The data were analysed using a type III sum of squares general linear model (GLM) procedure with the fixed factors PROGRAM, SUBJECT, and SYSTEM and all main and two-way interactions. For both dependent variables the model is significant (Spatial: $F(149, 210) : 3.400, p < 0.000$; Timbral: $F(149, 210) : 4.831, p < 0.000$) suggesting the model is valid.

In practice some similarity can be observed between the two tables. In both cases the most significant factor is SYSTEM (Spatial: $F(5, 210) : 62.021, p < 0.000$; Timbral: $F(5, 210) : 105.016, p < 0.000$), which allows us to reject the null hypothesis. The second most significant factor is the two-way interaction of PROGRAM*SYSTEM (Spatial: $F(15, 210) : 2.830, p < 0.000$; Timbral: $F(15, 210) : 3.097, p < 0.000$), implying that systems are rated differently for each program item.

All other factors and interactions are found insignificant. It is interesting to note that SUBJECT is insignificant, implying that listeners tend to rank the systems in a similar fashion, i.e. there is consensus. Alternatively, this may be due to the listeners performing the ranking based upon identification. Also PROGRAM is not significant, suggesting no difference between the items.

Having performed the ANOVA's it is now reasonable to study the results in detail. Figures 6 and 7 illustrate the means and 95% confidence intervals for all listeners for the spatial and timbral variables respectively. As the PROGRAM factor is not considered significant it is reasonable to average across it, as illustrated in Figure 8.

It can be noted that for both dependent variables the grading are very similar in nature. In all cases

the 5 channel system is consistently and significantly superior than any of the VHT systems, whilst system 1 is considered the most inferior in all cases. For systems 2-5 no significant differences are perceived for the spatial variable. For the timbral variable large differences are found, with system 5 fairing only slightly worse than the discrete 5 channel system and system 2 only marginally below that.

It can be noted that overall the difference between the discrete 5 channel system and the best VHT system is 2 ranks for the spatial variable and less than one ranking for the timbral variable.

It is noted that there is a wider error variance for Blue Bell item compared to the rain item. Many listeners commented on the complexity of grading the former item due to its very time varying characteristics. The rain sample, by comparison, was very time invariant. It might be concluded that time invariant test items are more suitable in this type of test, confirming earlier findings [30].

2.2. AES16 experiment

Once again the data were analysed using a type III sum of squares GLM procedure with the fixed factors PROGRAM, SUBJECT, and SYSTEM and all main and two-way interactions. As the number of listeners was substantially larger than in the NRC experiment, the statistical power of the AES16 experiment was much greater as if reflected in the statistics, which are found in Tables 7 and 8. For both dependent variables the models is significant (Spatial: $F(287, 468) : 5.622, p < 0.000$; Timbral: $F(287, 468) : 6.764, p < 0.000$) suggesting the model is valid.

As in the NRC experiments, similar factors are significant, with SYSTEM dominating, with F-statistics more than 40 times higher than any other factor (Spatial: $F(6, 468) : 173.703, p < 0.000$; Timbral: $F(6, 468) : 209.638, p < 0.000$). The second most significant factor is the interaction between SYSTEM*PROGRAM $F(18, 468) : 5.331, p < 0.000$; Timbral: $F(18, 468) : 5.539, p < 0.000$) implying that systems are graded differently for different program items. The third significant factor is now SUBJECT*SYSTEM $F(156, 468) : 2.589, p < 0.000$; Timbral: $F(6, 468) : 3.091, p < 0.000$). This suggests that for the AES16 experiment some listeners graded system differently than others.

The means and confidence intervals for each dependent variable are illustrated in Figures 9 and 10 with the results averaged over program shown in Figure 11.

Again we find supporting evidence that the discrete 5 channel system is significantly superior to any of the tested VHT systems. Three systems, namely 25, 27 and 28 all fall in a similar range around the 4-5 ranking. System 23 has a slightly poorer performance in the 2-3 range and system 26 is ranked lowest.

System 27 shows very high performance for one program item, namely the Mendelsohn item, with a comparable performance to the 5 channel systems both spatially and timbrally. It can also be noted that there is a large spread of variance as a function of program compared to the NRC experiment.

In addition to the overall analysis, several listeners (N=6) were members of the proponent organisations. For interest the data for this group has been presented averaged across program in Figure 12. Whilst it can clearly be seen that the error variances are slightly wider, due to the limited size of the group, significant difference can be found between systems, as predicted in the ANOVA tables. It can be noted that generally the ratings of this group are very similar to that of the overall population, both in terms of the ranking and the comparison between systems.

2.3. Correlation analysis

Based upon the results of the ANOVA tables and the presented graphs, it becomes very apparent that there might be a correlation between the two dependent variables. To test this a Pearson correlation was performed for all the data, the results of which are to be found in Table 9. In all cases we can see a

fair degree of correlation, in the order of 61.1 – 85.1%, with the higher values apparent in the AES16 data. This type of results has been noted in previous studies [31], where similar grading scales were employed with an untrained listening panel. This relatively high level of correlation can be attributed to a number of causes. Firstly, when employing 3D sound techniques the psychoacoustic techniques implicitly link timbral and spatial cues. Secondly, untrained listeners may have trouble differentiating multiple rating scales.

3. CONCLUSIONS

Based upon the discussion of the results above, the following conclusions may be drawn for this experiment.

- Results are statistically significant and considered meaningful
- Results are similar between both sites
- Discrete 5 channel is significantly superior. It has been suggested that this is due to listeners being so familiar with discrete 5 channel reproduction, that it is thus preferred. This concept has been disregarded due to the following observations
  - The auditory memory is short.
  - Most NRC listeners are not familiar with surround systems nor have they performed any other surround sound experiments.
  - On average the proponents who participated in the test, who are very familiar with the sound of VHT systems, also preferred the discrete 5 channel system consistently.
- There are large and significant differences in the perceived performance of different VHT systems both spatially and timbrally
- Proponents participating in the test (N = 6) graded in a similar fashion to the average population of the study (N = 27 in total)
- There are strong correlations between the ranking for both spatial and timbre quality for all program items. In general the blue bell railway items has wider error variance than other program items and was commented to be very difficult to evaluate. This sample is very time variant in nature and, in retrospect, considered to be a non-ideal sample
- The rank order method is very fast and provides very useful results

4. ACKNOWLEDGEMENTS

This work forms part of the studies of the Eureka 1653 Medusa (Multichannel Enhancement of Domestic User Stereo Applications) project[3] . All members of the Medusa project are thanked for their comments and discussion throughout the project so far. Søren Bech is thanked for his critical and constructive comments regarding experimental design.

The proponents of virtual home theatre systems are thanked for their support in providing systems for evaluation. The British Broadcasting Corporation and the German Surround Sound Forum are

[3] The Medusa project is a 3.5 years joint research project with the following partners: British Broadcasting Corporation, The Music Department of the University of Surrey, Nokia Research Centre, Genelec Oy, and Bang & Olufsen A/S.

thanked to providing discrete 5-channel test material. The authors would like to thank Kalle Koivuniemi, Jukka Holm and Gaetan Lorho (Nokia Research Center) for assisting in the preparation and running of these experiments. The NRC listening panel and delegates of the AES 16^{th} International Conference are thanked for participating in this evaluation. We are endebted to the funding bodies for supporting the Eureka Medusa project, including Tekes (Technology Development Centre of Finland).

REFERENCES

[1] B. B. Bass, "Stereophonic earphones and binaural loudspeakers," *Journal of the Audio Engineering Society*, vol. 9, pp. 148–151, 1961.

[2] M. Schroeder and B. Atal, "Computer simulation of sound transmission in rooms," *IEEE Conv. Record, pt. 7*, pp. 150–155, 1963.

[3] B. S. Atal and M. R. Schroeder, "Apparent sound source translator." U.S. patent no. 3,236,949, Feb. 1966.

[4] J. Blauert, *Spatial hearing. The psychophysics of human sound localization*. Cambridge, MA, USA: MIT Press, 1997.

[5] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction," *Journal of the Acoustical Society of America*, vol. 50, no. 4, pt. 2, pp. 1109–1115, 1971.

[6] D. H. Cooper and J. L. Bauck, "Prospects for transaural recording," *Journal of the Audio Engineering Society*, vol. 37, pp. 3–19, Jan./Feb. 1989.

[7] K. Kotorynski, "Digital binaural/stereo conversion and crosstalk canceling," in *Proceedings of the 89$^{\text{th}}$ Convention of the Audio Engineering Society, Preprint 2949*, (Los Angeles, USA), 1990.

[8] C. J. MacCabe and D. J. Furlong, "Spectral stereo surround pan-pot," in *Proceedings of the 90$^{\text{th}}$ Convention of the Audio Engineering Society, Preprint 3067*, (Paris, France), 1991.

[9] M. J. Walsh and D. J. Furlong, "Improved spectral stereo head model," in *Proceedings of the 99th Convention of the Audio Engineering Society*, (New York, USA), preprint 4128, 1995.

[10] J. L. Bauck and D. H. Cooper, "Generalized transaural stereo and applications," *Journal of the Audio Engineering Society*, vol. 44, no. 9, pp. 683–705, 1996.

[11] W. G. Gardner, *3-D Audio Using Loudspeakers*. PhD thesis, MIT Media Lab, September 1997.

[12] J. L. Bauck, "A new loudspeaker technique for improved 3-d audio," in *Proc. AES 14th International Conference*, (Seattle, WA, USA), June 1997.

[13] O. Kirkeby, P. Nelson, and H. Hamada, "The stereo dipole - binaural sound reproduction using two closely spaced loudspeakers," *Journal of the Audio Engineering Society*, vol. 46, no. 5, pp. 387–395, 1998.

[14] D. Ward and G. Elko, "Optimum loudspeaker spacing for robust crosstalk cancellation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Los Alamitos, California), The Institute of Electrical and Electronics Engineers, IEEE Computer Society Press, 1998.

[15] D. Ward and G. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Processing Letters*, vol. 6, no. 5, pp. 106–108, 1999.

[16] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Cambridge, MA, USA: Academic Press, 1994.

[17] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proceedings of the 98th Convention of the Audio Engineering Society*, (Paris, France), preprint 3980, 1995.

[18] N. Zacharov, J. Huopaniemi, and M. Hämäläinen, "Round robin subjective evaluation of virtual home theatre sound systems at the aes 16^{th} international conference," in *Proc. of the AES 16^{th} International Conference on Spatial Sound Reproduction*, pp. 544–556, Audio Eng. Soc., 1999.

[19] ITU-R, *Recommendation BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Telcommunications Union Radiocommunication Assembly, 1997.

[20] ITU-R, *Recommendation BS.775-1, Multichannel stereophonic sound system with and without accompanying picture*. International Telcommunications Union Radiocommunication Assembly, 1994.

[21] N. Zacharov and V. V. Mattila, "Gls - a generalised listener selection proceedure." unpublished, 1998.

[22] R. Quesnel and W. R. Woszczyk, "A computer-aided system for timbral ear training," in *AES Int. Conv.*, Audio Eng. Soc., 1994.

[23] R. Quesnel, "Timbral ear trainer: Adaptive, interactive training of listening skills for evaluation of timbre difference," in *AES Int. Conv.*, Audio Eng. Soc., 1996.

[24] D. G. Kirby, N. A. F. Cutmore, and J. A. Fletcher, "Programme origination of 5-channel surround sound," in *AES Int Conf.*, Audio Eng. Soc., 1997.

[25] "Multichannel universe." DVD, Balance München and GLS München, 1998.

[26] H. T. Lawless and H. Heyman, *Sensory evaluation of food*. Chapman and Hall, 1998.

[27] J. Hynninen and N. Zacharov, "Guineapig - a generic subjective test system for multichannel audio," in *Proc. of the AES 106^{th} Int. Conv.*, Audio Eng. Soc., 1999.

[28] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Proceeding of the AES 16^{th} International Conference.*, Audio Eng. Soc., 1999.

[29] E. Paulus and E. Zwicker, "Programme zur automatischen bestimmung der lautheit aus terzpegeln oder frequenzgruppenpegeln," *Acustica*, vol. 27, pp. 253–266, 1972.

[30] N. Zacharov, "On the loudspeaker directivity considerations for 5.1 channel audio-visual reproduction: A subjective appraisal," in *Proceedings of the 102^{nd} Convention of the Audio Engineering Society, Preprint 4459*, pp. 288–303, Audio Eng. Soc., 1997.

[31] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and subjective evaluation of head-related transfer function filter design," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 218–239, 1999.

[32] URL: http://www.harman.com/?innovation/featuredtech/vmax.ptml, 1999.

[33] URL: http://www.srslabs.com/technology/trusurround.html, 1999.

[34] J. Huopaniemi, *Virtual acoustics and 3-D sound in multimedia signal processing*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, to be published in 1999.

[35] URL: http://www.sensaura.co.uk, 1999.

[36] URL: http://www.dolby.com, 1999.

FIGURES

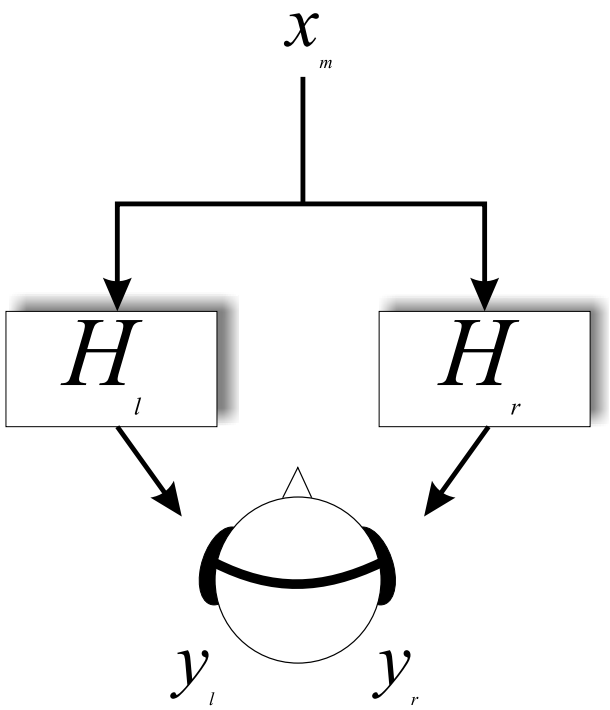

Figure 1: Signal flowgraph of 3-D audio positioning.

[4] Certain characteristics of these systems have been corrected since the original paper [18].

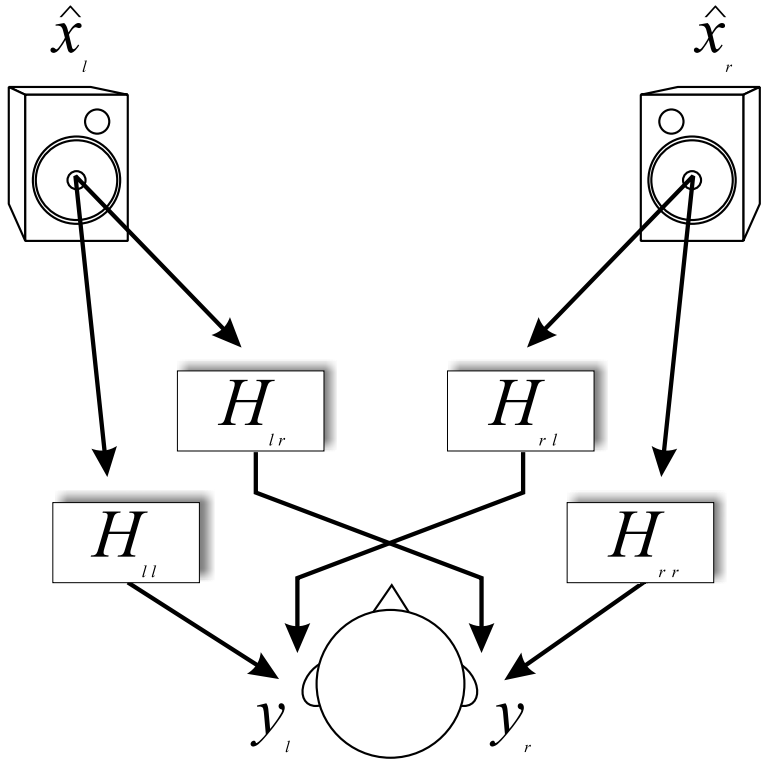

Figure 2: Signal flowgraph in loudspeaker listening.

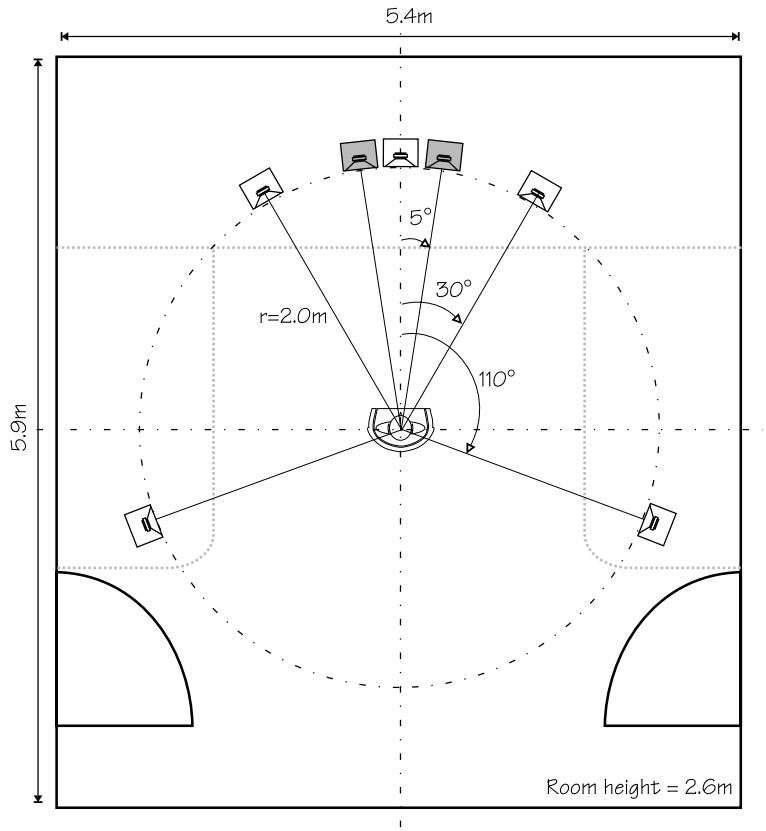

Figure 3: Layout of the new NRC ITU-R BS.1116-1 [19] compliant listening room and loudspeaker setup.

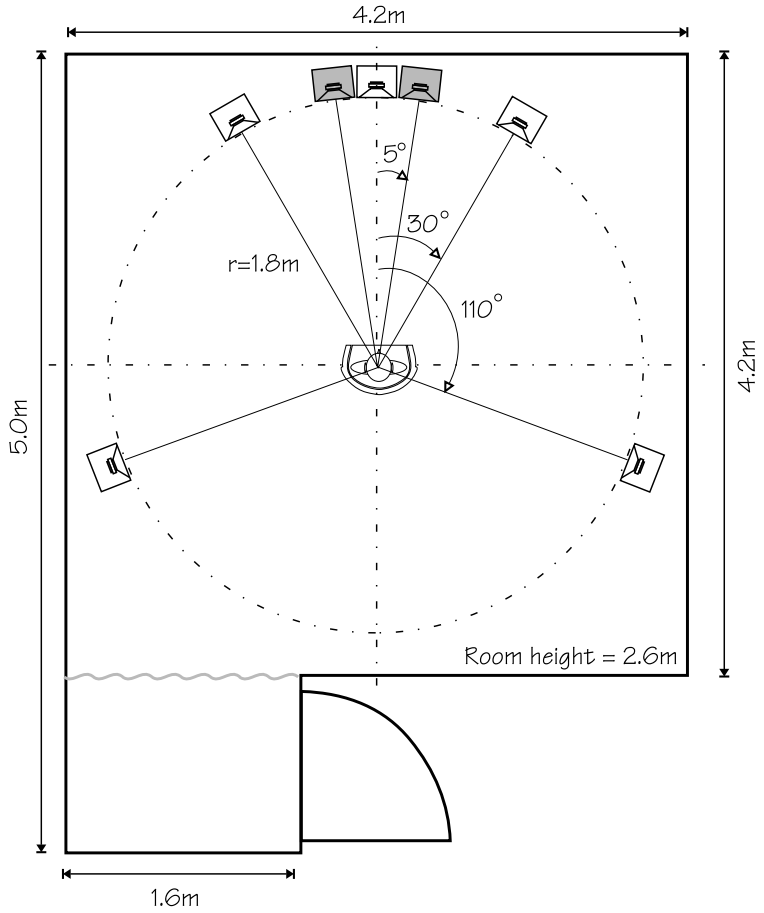

Figure 4: Layout of the AES16 test room and loudspeaker setup.

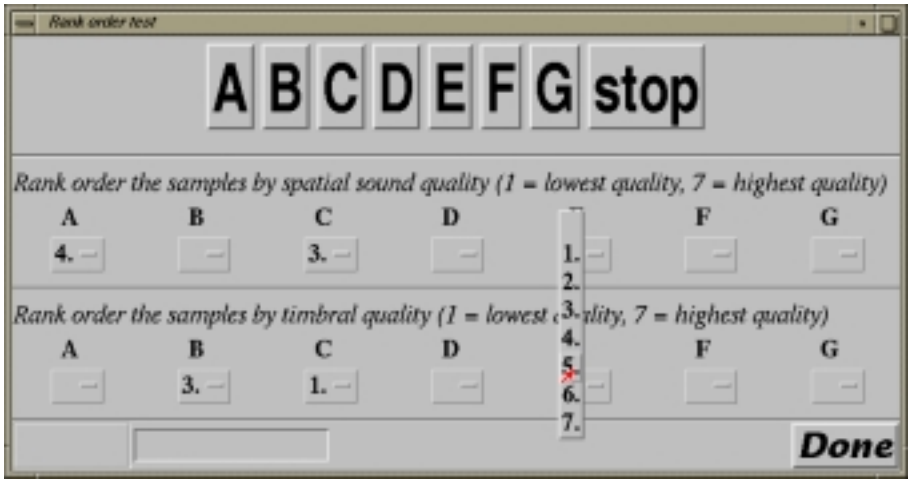

Figure 5: The rank order test user interface.

Proponent	**System**	**Reproduction angle**	**Center channel**	**MIPS @ 48 kHz**	**System description**
Harman Multimedia	VMAx 3D Virtual Theatre [32]	5°	Virtual	6	85 filter coefficients, memory 400×(word size)
Aureal Semiconductors	A3DS	30°	Phantom	10	Virtual front channels and decorrelation processing, coefficient memory: 150 words, delay memory: 1k words (inc. decorrelator)
SRS Labs, Inc.	TruSurround [33]	30°	Phantom	4.7	IIR filters with 18 filter coefficients / Total ROM code + filter coefficients = 306, RAM=36
Helsinki University of Technology	PBVS [34]	30°	Phantom	6.4	20^{th} order warped IIR filters
Sensaura Ltd	Sensaura Virtual Surround [35]	30°	Phantom	5.2	25 coefficients per filter, no additional memory required
Dolby Laboratories [36]	Dolby Virtual Surround	5°	Phantom	∼ 5	10 filters and 10 delays

Table 1: Summary of proponent systems, based upon the information provided by manufacturers.

	Exp.	**Spatial**		**Timbral**		**Assump.**
		statistic	*sig.*	*statistic*	*sig.*	
Normal distribution[a]	NRC	0.158	0.000	0.153	0.000	ok
	AES16	0.135	0.000	0.132	0.000	ok
Homogeneity of variance[b]	NRC	0.920	0.431	1.343	0.259	ok
	AES16	0.087	0.967	0.121	0.948	ok
Normal distribution of residuals[c]	NRC	0.050	0.031	0.048	0.049	ok
	AES16	0.040	0.007	0.024	*0.200*[d]	∼ ok

Table 2: Testing ANOVA assumptions for all data.

[a] Kolmogorov-Smirnow test with Lilliefors significance correction
[b] Levene's test, based upon the mean
[c] Kolmogorov-Smirnow test with Lilliefors significance correction
[d] Lower bound of significance

Dependent variable	**System**	**Subject**	**Program**
Spatial	$F = 54.014, p < 0.000$ *(153.425, p<0.000)*	$F = 0.370, p < 0.982$ *(5.081, p<0.985)*	$F = 0.574, p < 0.632$ *(1.755, p<0.625)*
Timbral	$F = 91.849, p < 0.000$ *(203.093, p<0.000)*	$F = 0.217, p < 0.999$ *(3.094, p<0.999)*	$F = 0.211, p < 0.889$ *(0.648, p<0.885)*

Table 3: Comparison of one-way ANOVA and Kruskal-Wallis test for NRC experiment. *Chi-square statistics and significance in italic.*

Dependent variable	**System**	**Subject**	**Program**
Spatial	$F = 125.195, p < 0.000$ *(382.540, p<0.000)*	$F = 0.352, p < 0.999$ *(9.185, p<0.999)*	$F = 0.201, p < 0.896$ *(0.688, p<0.876)*
Timbral	$F = 136.384, p < 0.000$ *(392.690, p<0.000)*	$F = 0.571, p < 0.958$ *(15.137, p<0.955)*	$F = 0.694, p < 0.556$ *(2.055, p<0.561)*

Table 4: Comparison of one-way ANOVA and Kruskal-Wallis test for AES16 experiment. *Chi-square statistics and significance in italic.*

Tests of Between-Subjects Effects

Dependent Variable: SPATIAL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	*718.958*[b]	*149*	*4.825*	*3.400*	*.000*	*506.582*	*1.000*
Intercept	*4278.003*	*1*	*4278.003*	*3014.307*	*.000*	*3014.307*	*1.000*
PROGRAM	*4.897*	*3*	*1.632*	*1.150*	*.330*	*3.451*	*.307*
SUBJID	*15.039*	*14*	*1.074*	*.757*	*.715*	*10.596*	*.475*
SYSTEM	*440.114*	*5*	*88.023*	*62.021*	*.000*	*310.107*	*1.000*
PROGRAM * SUBJID	*21.561*	*42*	*.513*	*.362*	*1.000*	*15.192*	*.383*
PROGRAM * SYSTEM	*60.253*	*15*	*4.017*	*2.830*	*.000*	*42.454*	*.995*
SUBJID * SYSTEM	*177.094*	*70*	*2.530*	*1.783*	*.001*	*124.782*	*1.000*
Error	*298.039*	*210*	*1.419*				
Total	*5295.000*	*360*					
Corrected Total	*1016.997*	*359*					

a. Computed using alpha = .05

b. R Squared = .707 (Adjusted R Squared = .499)

Table 5: ANOVA tables for the NRC experiment, for spatial rank.

Tests of Between-Subjects Effects

Dependent Variable: TIMBRE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	*820.522*[b]	*149*	*5.507*	*4.831*	*.000*	*719.823*	*1.000*
Intercept	*4452.100*	*1*	*4452.100*	*3905.713*	*.000*	*3905.713*	*1.000*
PROGRAM	*1.878*	*3*	*.626*	*.549*	*.649*	*1.647*	*.162*
SUBJID	*9.233*	*14*	*.660*	*.579*	*.880*	*8.100*	*.358*
SYSTEM	*598.533*	*5*	*119.707*	*105.016*	*.000*	*525.078*	*1.000*
PROGRAM * SUBJID	*14.789*	*42*	*.352*	*.309*	*1.000*	*12.974*	*.319*
PROGRAM * SYSTEM	*52.956*	*15*	*3.530*	*3.097*	*.000*	*46.457*	*.998*
SUBJID * SYSTEM	*143.133*	*70*	*2.045*	*1.794*	*.001*	*125.567*	*1.000*
Error	*239.378*	*210*	*1.140*				
Total	*5512.000*	*360*					
Corrected Total	*1059.900*	*359*					

a. Computed using alpha = .05

b. R Squared = .774 (Adjusted R Squared = .614)

Table 6: ANOVA tables for the NRC experiment, for timbral rank.

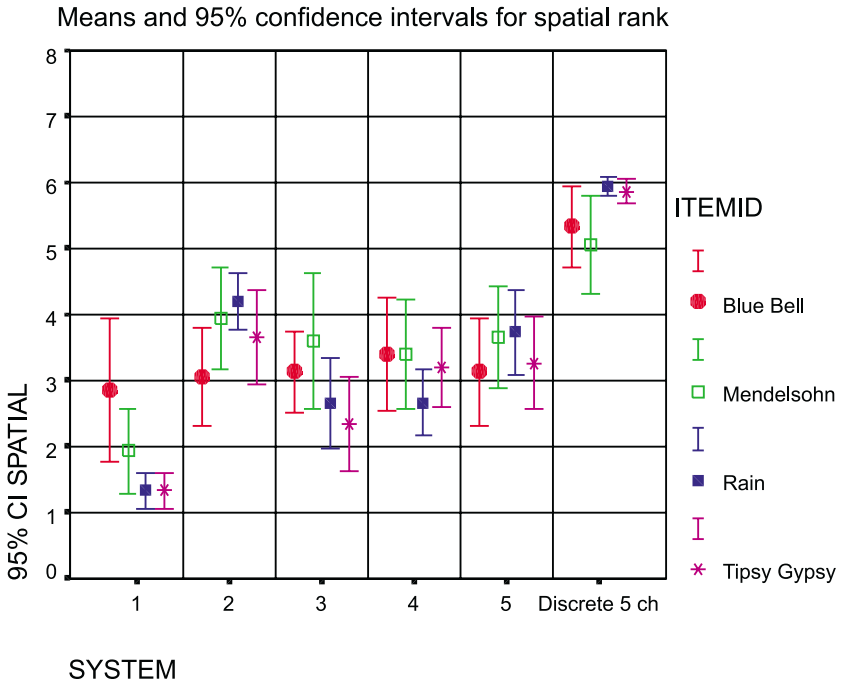

Figure 6: Means and 95% confidence intervals, for all listeners, for NRC experiment, for spatial rank.

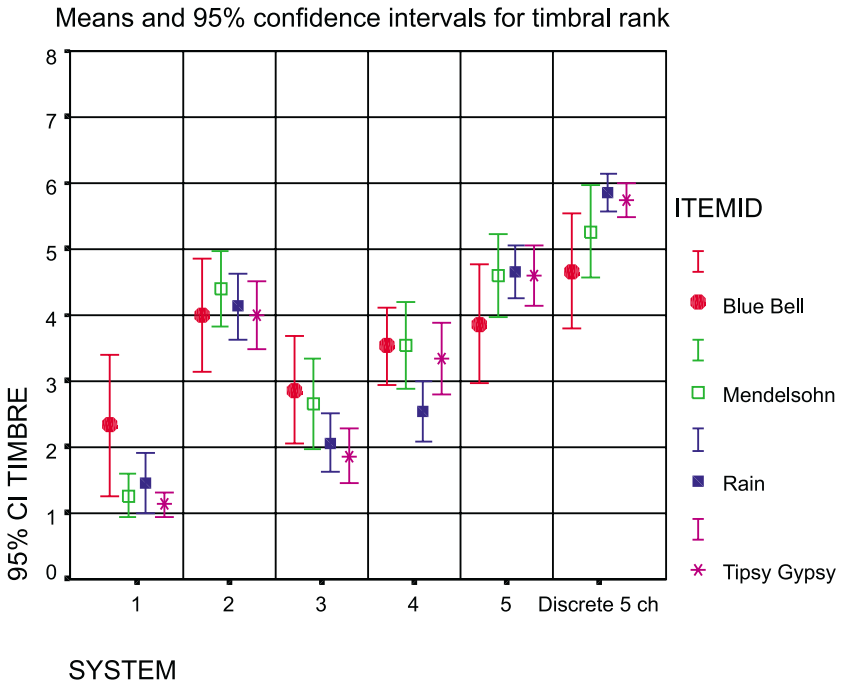

Figure 7: Means and 95% confidence intervals, for all listeners, for NRC experiment, for timbral rank.

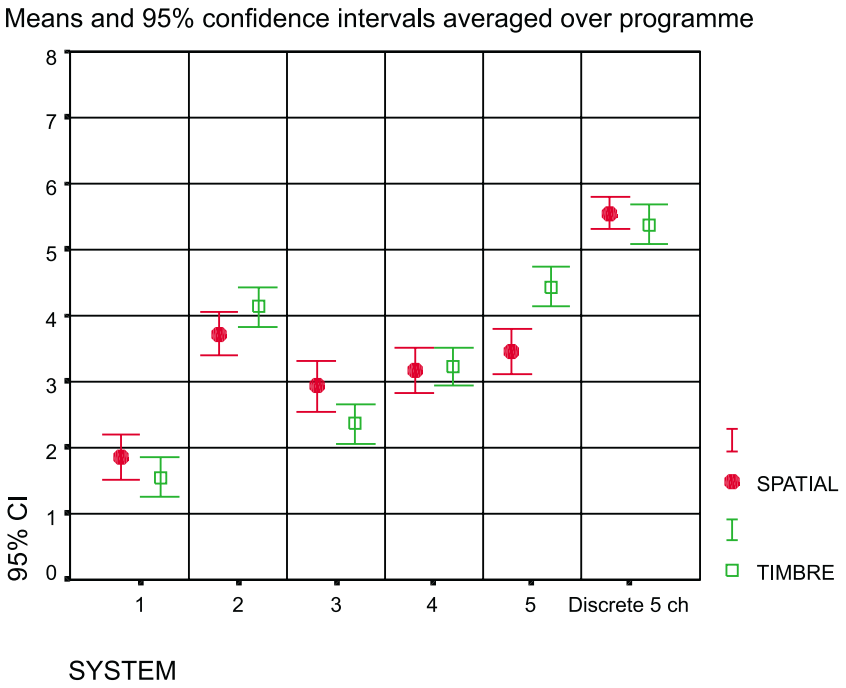

Figure 8: Means and 95% confidence intervals averaged over program, for all listeners, for NRC experiment, for both spatial and timbral rank.

Tests of Between-Subjects Effects

Dependent Variable: SPATIAL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	2224.308 [b]	287	7.750	5.622	.000	1613.420	1.000
Intercept	11301.493	1	11301.493	8197.632	.000	8197.632	1.000
SUBJID	35.614	26	1.370	.994	.475	25.833	.834
SYSTEM	1436.831	6	239.472	173.703	.000	1042.217	1.000
PROGRAM	2.300	3	.767	.556	.644	1.669	.165
SUBJID * SYSTEM	556.812	156	3.569	2.589	.000	403.888	1.000
SUBJID * PROGRAM	60.450	78	.775	.562	.999	43.848	.847
SYSTEM * PROGRAM	132.302	18	7.350	5.331	.000	95.966	1.000
Error	645.198	468	1.379				
Total	14171.000	756					
Corrected Total	2869.507	755					

a. Computed using alpha = .05

b. R Squared = .775 (Adjusted R Squared = .637)

Table 7: ANOVA tables for the AES16 experiment, for spatial rank.

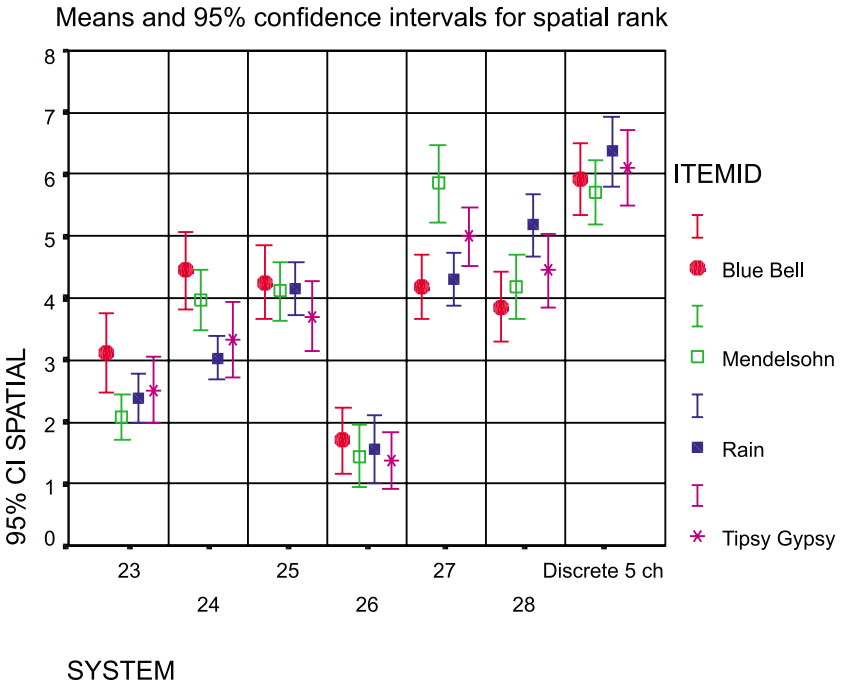

Figure 9: Means and 95% confidence intervals, for all listeners, for AES16 experiment, for spatial rank.

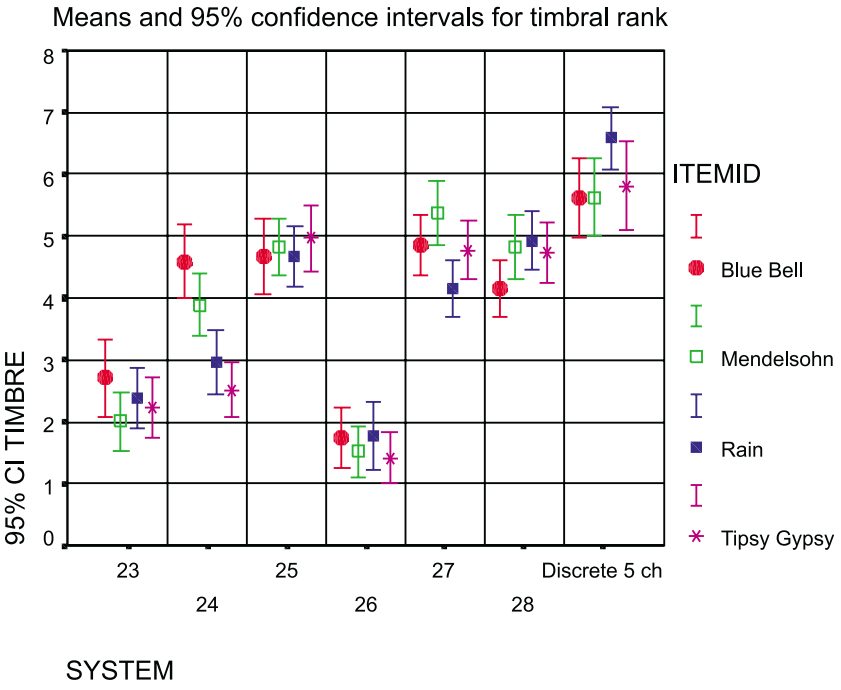

Figure 10: Means and 95% confidence intervals, for all listeners, for AES16 experiment, for timbral rank.

Tests of Between-Subjects Effects

Dependent Variable: TIMBRE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	2346.374 [b]	287	8.176	6.764	.000	1941.128	1.000
Intercept	11722.922	1	11722.922	9698.235	.000	9698.235	1.000
SUBJID	58.114	26	2.235	1.849	.007	48.077	.993
SYSTEM	1520.421	6	253.403	209.638	.000	1257.826	1.000
PROGRAM	8.036	3	2.679	2.216	.085	6.648	.561
SUBJID * SYSTEM	582.794	156	3.736	3.091	.000	482.138	1.000
SUBJID * PROGRAM	56.500	78	.724	.599	.997	46.742	.879
SYSTEM * PROGRAM	120.511	18	6.695	5.539	.000	99.697	1.000
Error	565.704	468	1.209				
Total	14635.000	756					
Corrected Total	2912.078	755					

a. Computed using alpha = .05

b. R Squared = .806 (Adjusted R Squared = .687)

Table 8: ANOVA tables for the AES16 experiment, for timbral rank.

	NRC		AES16	
	Pearson correlation	*2-tailed sig.*	*Pearson correlation*	*2-tailed sig.*
Blue bell railway	0.646	0.000	0.770	0.000
Mendelsohn	0.611	0.000	0.845	0.000
Rain	0.741	0.000	0.851	0.000
Tipsy gypsy	0.721	0.000	0.769	0.000
Overall	0.681	0.000	0.809	0.000

Table 9: Pearson correlation between spatial and timbral rank

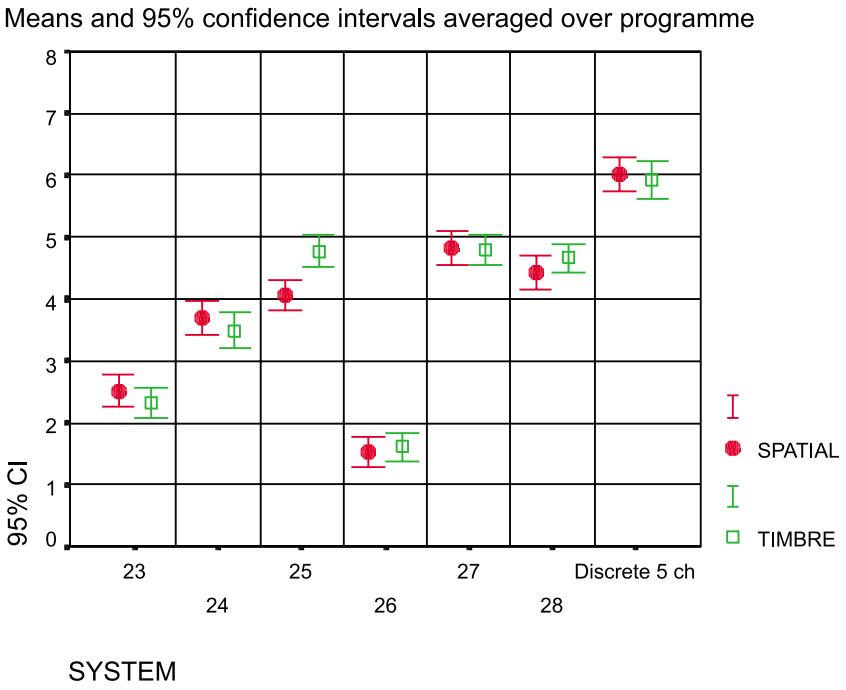

Figure 11: Means and 95% confidence intervals averaged over program, for all listeners (N=27), for AES16 experiment, for both spatial and timbral rank.

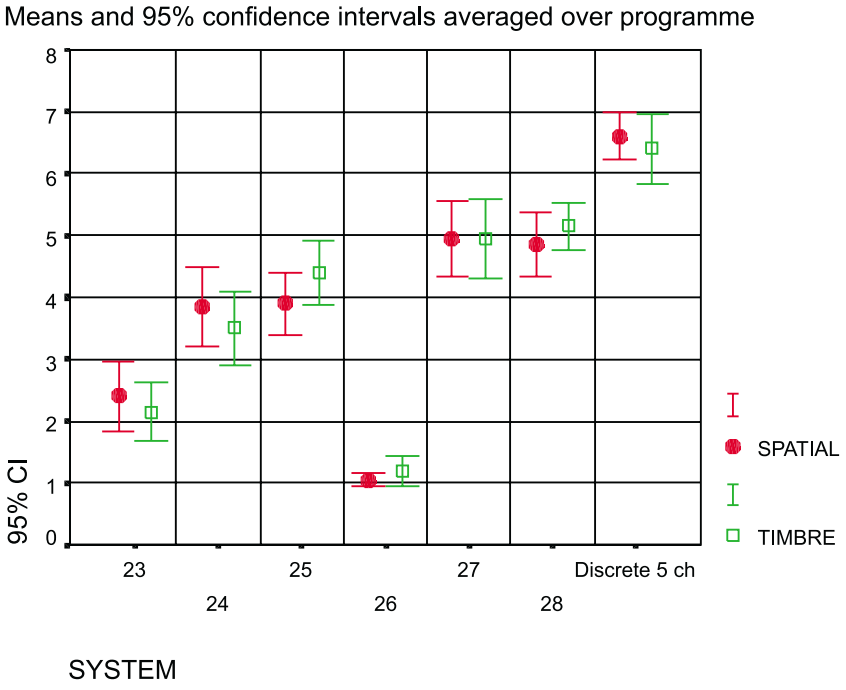

Figure 12: Means and 95% confidence intervals averaged over program, for system proponent listeners (N=6), for AES16 experiment, for both spatial and timbral rank.

