AES Surround Study Group

Evaluation Tests Report Part3

Presented at the AES Surround Recording Experiments Project Report
2006-2007 AES Japan.

Subjective Impression of Surround Sound Microphone Arrays

MARUI, Atsushi[1]; KAMEKAWA, Toru[1]; and IRIMAJIRI, Hideo[2]

[1] Department of Musical Creativity and the Environment, Faculty of Music,
Tokyo National University of Fine Arts and Music, Tokyo, Japan
marui@ms.geidai.ac.jp
kamekawa@ms.geidai.ac.jp

[2] Mainichi Broadcasting System, Inc. Osaka, Japan
iririn@mbs.co.jp

ABSTRACT

In order to investigate how recorded sounds of different surround sound microphone arrays are perceived, subjective evaluation of two classical music pieces recorded using seven microphone arrays were conducted in Fukuoka, Osaka, Tokyo, and Vienna. The listening tests were done using Scheffé's pairwise rating method on *powerfulness*, *width*, *softness*, and *preference*. Although, in general, the ratings were highly dependent on a choice of music piece, high and low ratings on powerfulness were seen for Decca Tree with Omni Square array and Double MS array, respectively for both of the pieces used in the experiments. Also, prediction equation of *preference* was formulated from perceptual attributes, and the equation showed that higher *powerfulness*, higher *width*, and lower *softness* yielded higher *preference*.

1. Subjective Evaluation on Impression of Recorded Sounds from Surround Sound Micrphone Arrays (in Japan)

Subjective evaluation tests were conducted at Fukuoka, Osaka, and Tokyo, Japan, in March 2007, to investigate the perceptual impression of music recorded using multiple surround sound microphone arrays. This paper reports the methods, analyses, and results of the subjective evaluation.

The motivation of the study was to investigate how listeners' impression are affected from differences in surround sound microphone arrays. Also, the authors' interest was in how individual background (occupation, listening position, age, etc.) of the listeners has influence on the impression. The analyses included the prediction of affective sentiments (preference) from perceptual judgments (powerfulness, spaciousness, and softness).

1.1. Sound Stimuli

Two music programs were recorded using seven surround sound microphone arrays making 14 stimuli. Two programs were "Pines of Rome" by Respighi and "Wellington's Victory" by Beethoven. Refer to the part one of the report for the selection of the programs.

The stimuli used in the experiment were recorded using seven microphone arrays. The seven microphone array were carefully chosen by the authors to have different characteristics. The selection were done considering the criteria including salient spatial/timbral characteristics, little history of

being evaluated, and reputation in the industry. The seven microphone arrays were:

- Fukada Tree
- Decca Tree & Omni Square Mid
- 3 Omni & IRT Cross
- 5 Cardioid & Hamasaki Square Near
- Omni 8
- INA 5
- Double MS

The detailed settings of microphone arrays are shown in the Part 1 of this document, and therefore omitted here.

1.2. Method

Scheffé's pairwise comparison method was employed for the experiments. Pairwise comparison is a method that a participant is asked to compare a pair of two stimuli out of all stimuli. All possible combination of pairs should be compared on attribute scales usually anchored with bipolar adjective pairs [1].

One weakness of the pairwise comparison, which is not only limited to the Scheffé's, is that the number of combination may become undesirably large. For example, if k stimuli are to be compared in a pair (in both direction of A-B and B-A) number of all combinations is $k\times(k-1)$, which grows polynomially larger for the increase of k. Since there are seven surround sound microphone arrays in the current investigation, $7\times(7-1)=42$ comparison was necessary to span the linear combination of all stimuli. However, the decision was made to spread combinations over multiple participants, because the number of participants was decent while there were time constraints for each participant. Specifically, each participant was asked to compare either A-to-B or B-to-A (upper or lower half of the pairwise comparison matrix) and each participant was given three out of the six attributes for rating.

Maximum of 14 participants sat in one run of the experiment around the sweet spot. Therefore, the authors are aware that some participants were in the non-ideal listening spot for some of the evaluations.

1.3. Participants

The experiment was done in three different locations in Japan, namely, Fukuoka, Osaka, and Tokyo. Numbers of participants were 54 in Fukuoka, 87 in Osaka, and 111 in Tokyo. From the clear mistakes and no answers in 11 of the participants' data, 241 out of 252 data were used for the further analyses.

Age of participants ranged from 19 to 76 (with mean of 39.3 and standard deviation of 11.1). Ratio of the gender was 87 to 13.

Self-described occupation varied in wide range from students and householders to broadcasting engineers and researchers of recording equipments. In this report, the occupations are categorized to three groups: audio specialists in broadcasting industry, audio specialists in non-broadcasting industry, and people in non-audio industry (including naïve participants).

1.4. Instruction to Participants

In Scheffé's pairwise comparison, scales of bipolar attribute pairs are used to specify what in a stimulus to pay attention to. Thus, the selection of attributes becomes one of the important factors. In the experiment, four attributes "迫力" (*hakuryoku*; powerfulness), "拡がり" (*hirogari*; spatial width), "柔らかさ" (*yawarakasa*; softness), and "好み" (*konomi*; preference) were chosen. *Powerfulness*, *softness*, and *preference* were chosen to refer to three central factors of timbral impression of musical instruments (*powerful*, *metallic*, and *beautiful*, respectively). *Spatial width* was added to represent spatial aspects of the recorded materials.

When a human is comparing a pair of stimuli on a given scale, there are two distinct worlds of local and global. Evaluation on a specific perceptual attribute such as *brightness* and *size* is based on a single perception of the target stimulus, and evaluation on an affective attribute such as *preference* and *suitability* is based on a global impression of the stimulus [2]. In the attributes used in this current investigation, *preference* is the attribute for asking global impression and the other three are for asking single perception of the stimuli. Scheffé's method can be used for both local perception attributes and global affection attributes, but in order to avoid the bias caused by rating on affection before the single perception, the attribute *preference* was always asked as the last attribute.

Following instruction was given to each participant:

In this section, two sounds are going to be presented per question. Please listen carefully and answer "how" the second sound is different from the first. The answer should be chosen from 1 to 5.

1. The second sound is more POWERFUL than the first
   (1: strongly disagree, 2: disagree, 3: neither, 4: agree, 5: strongly disagree)

2. The second sound is more POWERFUL than the first
   (1: strongly disagree, 2: disagree, 3: neither, 4: agree, 5: strongly disagree)

.
.
.

The above instruction only showed the part regarding powerfulness, but two attributes were presented per participant and all four attributes were balanced over the participants. Note that although above instruction and attributes are written in English, the actual instruction was written and given in Japanese.

1.5. Analyses

Scheffé's pairwise comparison method is commonly associated with ANOVA (analysis of variance). From the ANOVA result, microphone arrays had significant differences in mean responses with 1% significance level on all adjectives in both of the two programs. The results are shown in Figure 1 (for "Pines of Rome") and Figure 2 (for "Wellington's Victory"). Colored bars show mean scores and whiskers show 95% confidence intervals around the means.

Confidence interval gives us an interval estimate that is known for good compatibility with the scale values used in the experiment, and a relationship with *p*-value in null-hypothesis significance testing is also known. The use of confidence interval is recommended over *p*-values in *APA Style Manual* for better explanation of data [4]. The meaning of "95% confidence interval" is that "there is a 95% chance that the interval contains the population mean." In this report, a simpler method based on confidence interval yet having as much testing power as a null-hypothesis significance testing was used to analyze the results of Scheffé's pairwise comparison.

The method proposed by Cumming and Finch [3] is a method that allows us to read confidence intervals and to obtain the result as powerful as typical null-hypothesis significance testing. Among seven *rules of eye* that they propose, one that is suitable for the current investigation is quoted here. A phrase "statistically significant" is used as *p*-value being less than .05, hereafter, noted otherwise.

> *Rule of Eye 4: For a comparison of two independent means, p≤05 when the overlap of the 95% CIs is no more than about half the average margin of error, that is, when proportion overlap is about .50 or less. In addition, p≤01 when the two CIs do not overlap, that is, when proportion overlap is about 0 or there is a positive gap. These relationships are sufficiently accurate when both sample sizes are at least 10, and the margins of error do not differ by more than a factor of 2.*

According to the rule, there were higher *powerfulness* with Decca Tree & Omni Square Mid and 5 Cardioid & Hamasaki Square Near, and conversely lower *powerfulness* with Double MS in "Pines of Rome" (Figure 1) . As for *spatial width*, no statistically significant difference was seen for all microphone array pairs excluding 5 Cardioid & Hamasaki Square Near and Double MS. For *softness*, significant differences against Double MS were seen in Fukada Tree, 5 Cardioid & Hamasaki Square, Omni 8, and INA 5. All microphones except for Fukada Tree were *preferred* over the recording of Double MS.

In "Wellington's Victory" (Figure 2), more clear results were obtained. Decca Tree & Omni Square Mid had high *powerfulness*, and 3 Omni & IRT Cross and Omni 8 also had statistically significant differences in *powerfulness*. 5 Cardioids & Hamasaki Square Near and Double MS methods did not have significant difference between the two in terms of *powerfulness* but had significant difference against other microphone arrays.

Similarly, by comparing means and confidence intervals between the surround sound microphone arrays, Decca Tree & Omni Square Mid and 3 Omni & IRT Cross had significant difference in *width* against 5 Cardioids & Hamasaki Square Near and Double MS. In terms of *softness*, although small but significant difference was seen between 3 Omni & IRT Cross と Double MS, other microphones were not clearly distinguished in *softness*.

While no significant difference between 5 Cardioids & Hamasaki Square Near and Double MS were found in *preferences*, the two arrays had significant difference against the other microphone arrays. One

thing to mention is that the difference here is somewhat similar to that of *powerfulness* and *width*, and it is tempting to say that *preference*, *powerfulness*, and *width* are correlated in some way.

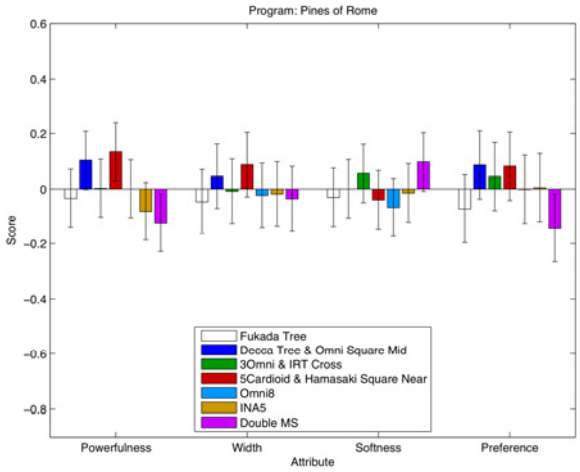


Figure 1: Scheffé's pairwise comparison analysis results from "Pines of Rome" for Japanese participants. Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.

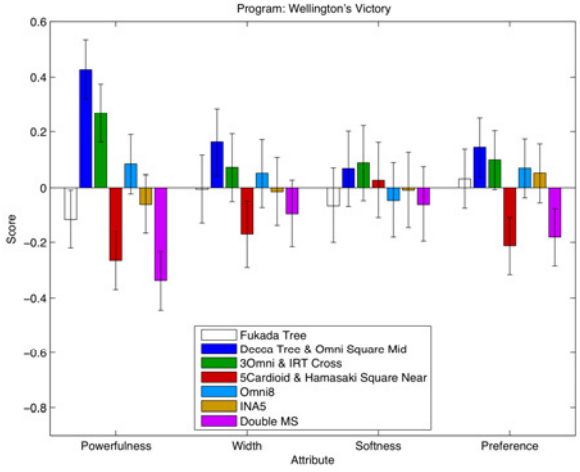


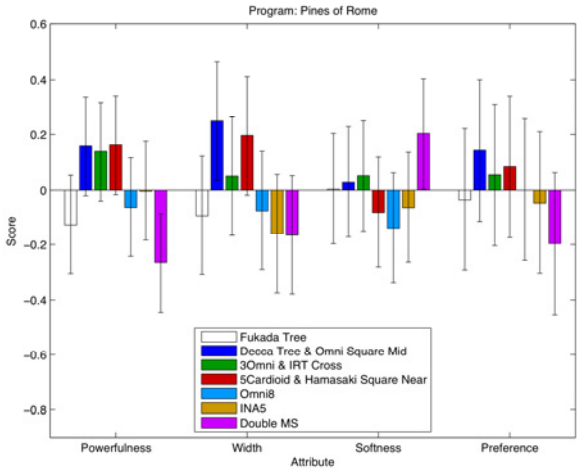
Figure 2: Scheffé's pairwise comparison analysis results from "Wellington's Victory" for Japanese participants. Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.

1.6. Differences in Listener Occupations and Seating Positions

Listeners were categorized into three groups—audio specialists in broadcasting industry, audio specialists in non-broadcasting industry, and people in non-audio industry (including naïve participants)—and analyses were done to compare between the groups. However, there were no significant difference between the occupation groups. Probable reason for this may be that the listeners from non-audio industry had high variance within the group. Also, another factor was that the number of the members in the group (approximately 20) was less than the number of the members in the other groups (approximately 35).

In the next analysis, the listeners were categorized with their listening positions (Figure 3 and Figure 4). In the case of "Pines of Rome," no significant differences between the microphone arrays were found when the listeners were seated in the back rows (Lower panel of Figure 3). Again, the number of the listeners was not enough to see any significant difference. The results from listeners positioned in the front rows (Upper panel of Figure 3) showed significant differences in *powerfulness* responses. Decca Tree & Omni Square Mid, 3 Omni & IRT Cross, and 5 Cardioids & Hamasaki Square Near were significantly different from Fukada Tree, Omni 8, and Double MS. Possible explanation is that the listeners seated in the frontal rows were more sensible to the powerfulness of the instruments, because orchestral instruments were mainly reproduced from the frontal speakers and ambience of the hall was mainly reproduced from the rear speakers in this particular piece. As for *width*, Decca Tree & Omni Square Mid and 5 Cardioids & Hamasaki Square Near were rated relatively *wide* and Fukada Tree, Omni 8, INA 5, and Double MS were rated relatively *narrow*. For *softness*, Double MS was significantly salient with comparison to 5 Cardioids & Hamasaki Square Near, Omni 8, and INA 5. Seating position did not show significant difference in *preference*.

Furthermore, in "Wellington's Victory" (Figure 4), *powerfulness* and *preference* seemed to be affected little from the differences in seating position. However, similarly to the case of "Pine of Rome," it is not possible to draw a significant conclusion on the matter from the limited number of listeners seated in the back row (lower panel of Figure 4) resulting in wider confidence interval.



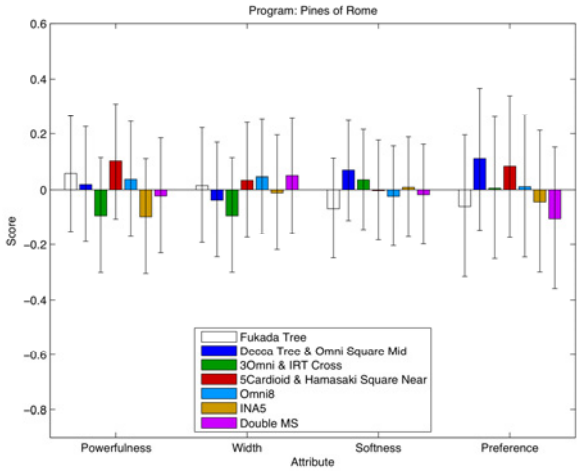
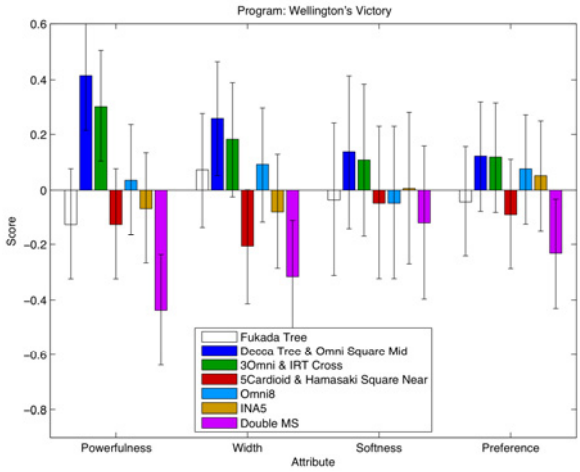


Figure 3: Results from participants seated in the front area (upper panel) , and the rear area (lower panel) , for "Pines of Rome." Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.



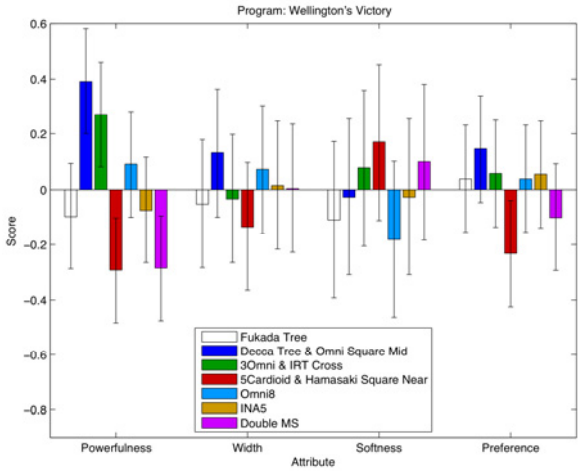


Figure 4: Results from participants seated in the front area (upper panel) , and the rear area (lower panel) , for "Wellington's Victory." Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.

In both cases of "Pines of Rome" and "Wellington's Victory," left and right seating positions showed little differences. The notable difference common to the left and right positions compared with center seats is that the confidence interval for *width* response was larger, suggesting the width evaluation became more difficult and inconsistent at the non-center seating positions.

2. Subjective Evaluation on Impression of Recorded Sounds from Surround Sound Micrphone Arrays (in Vienna)

2.1. Method and Participants

Experiment similar to the one done in Japan was conducted at the venue of 122nd AES Convention at Vienna, Austria. Sixty-five people from across the nations volunteered to participate in the experiment. However, from the experimental constraints including reproduction setup, mother tongue used by the participants, and the language used in the experiment, direct comparison with the results obtained in Japanese experiment is difficult.

Stimuli used in the listening experiment was the same as the Japanese experiment; "Pines of Rome" by Respighi and "Wellington's Victory" by Beethoven. Subjective rating on 5-point bipolar scales were done for the adjective scales *powerfulness*, *width*, and *softness*. For the sake of reducing the time for the experiment, *preference* scale was omitted from the Vienna experiment.

The microphone arrays presented were as follows:

- Fukada Tree
- Decca Tree & Omni Square Mid
- INA 5
- Double MS

All the instruction was given in English to all participants including non-native English speakers, to match the common language used in the convention. Presentation and data collection were done the same way as the Japanese experiment using Scheffé's pairwise comparison method.

2.2. Analyses and Results

The same analysis method was used as the Japanese experiment. From the ANOVA result, microphone arrays had significant differences in mean responses with 1% significance level on all adjectives in both of the two programs. The results are shown in Figure 5 ("Pines of Rome") and Figure 6 ("Wellington's Victory"). The bars show mean values and whiskers show 95% confidence intervals about the means.

From the "Pines of Rome" results, Double MS had significant difference against three other microphone arrays (Fukada Tree, Decca Tree & Omni Square Mid, and INA 5) in *powerfulness*. Fukada Tree was rated higher than three other microphone arrays in *width*. For *softness*, Fukada Tree and Double MS were rated on either end of the scale and the other two were rated in similar level.

For "Wellington's Victory," Decca Tree & Omni Square Mid were rated significantly higher and Double MS was rated lower in *powerfulness*. Although there were statistically significant pairs of microphones for *width*, overall differences were not so large. Decca Tree & Omni Square Mid were rated significantly lower in *softness* relative to the other microphone arrays.

Analyses for different occupation and seating position were done for the Vienna data following the Japanese experiment analyses. However, the number of participants were too small according to Cumming & Finch (there must be more than 10 samples to apply *Rule of Eye 4*), thus no significant differences were seen between the occupation group and seating position.

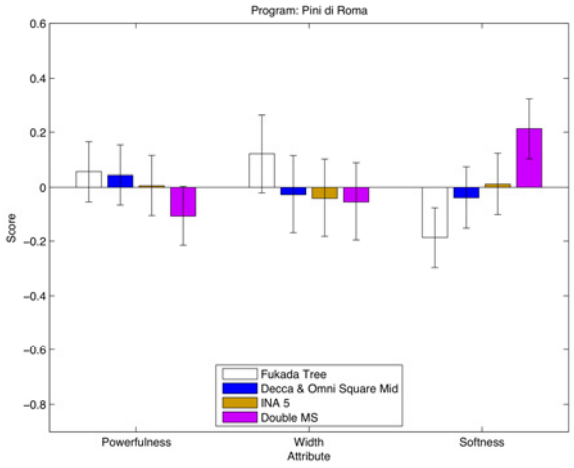


Figure 5: Scheffé's pairwise comparison analysis results from "Pines of Rome" for participants in Vienna. Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.

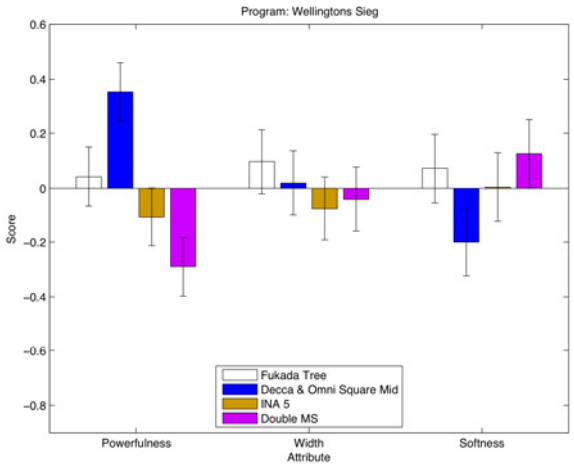


Figure 6: Scheffé's pairwise comparison analysis results from "Wellington's Victory" for participants in Vienna. Colored bars show relative mean ratings and whiskers show 95% confidence intervals around the mean.

3. Relation between Perceptual Attributes and Preference

It was mentioned that the perceptual attributes *powerfulness*, *width*, and *softness* are all based on a single perception of the target stimulus, and *preference* is based on a global impression of the stimulus in Section 1.4. According to Filter Model described in [2], attribute ratings on a global impression can be predicted from multiple local impressions (related to single perceptions). The prediction equation can be formulated by using multiple regression analysis with ratings from attributes on single perception as independent variables and ratings from an attribute on a global impression as dependent variable.

The following prediction equation was obtained from multiple regression analysis applied to the data in Japanese experiment for "Pines of Rome."

$$y_{\text{pref}} = .56x_{\text{power}} + .37x_{\text{width}} - .12x_{\text{softness}}$$

The fit was R^2=.74 and for "Wellington's Victory,"

$$y_{\text{pref}} = .37x_{\text{power}} + .46x_{\text{width}} - .69x_{\text{softness}}$$

was obtained with R^2=.89. Here, $x_{\text{attribute}}$s are the independent variables (ratings from each attribute) and y_{pref} is the dependent variable.

Although the coefficients in the preference prediction equation will vary for different musical programs, it if speculated that higher *preference* can be predicted from higher *powerfulness* and *width* with lower *softness* in a recording of multichannel surround sound.

However, since this result was derived from the limited number of musical programs and microphone arrays, it may or may not work for different programs or microphone arrays. In addition, it must be noted that, although high R^2 value was obtained to show a goodness-of-fit of the models, this may be from comparatively larger number of independent variables for the number of samples in the dependent variable.

Furthermore, similar analysis was not possible for the experimental results from Vienna experiment, for not having collected the ratings on listener preferences for microphone arrays.

4. Comparison between 5.1ch Surround and 2ch Stereo

Pairwise comparison of a 2 channel stereophonic reproduction and a 5.1 channel surround sound reproduction was done as a part of the experiment both in Japan and in Vienna. The two stimuli (2ch and 5.1ch) were independently mixed, namely, 2ch version was not a simple downmix from 5.1ch version. Because the constraints for the experiments in Japan and in Vienna were different, numbers of attributes used in the experiments were different. Seven attributes, "迫力" (*hakuryoku*; powerfulness), "奥行き"(*okuyuki*; depth), "左右への拡がり" (*sayuu e no hirogari*; spread in width direction), "上下への拡がり" (*jouge e no hirogari*; spread in height direction), "包まれ感の自然さ" (*tsutsumare-kan no shizen-sa*; naturalness in envelopment) "余韻の自然さ" (*yoin no shizen-sa*; naturalness in reverberation), and "好み" (*konomi*; preference) in Japan, and six attributes, *powerfulness*, *depth*, *width*, *envelopment*, *natural reverberation*, and *preference* in Vienna were used as pairwise comparison attribute scales.

The result is shown in Figure 7. Crosses show mean ratings and whiskers show 95% confidence intervals around the mean. Dashed line shows the neutral point where no significant difference was observed between 2 channel stereo and 5.1 channel surround if a confidence interval do not cross below and above it.

Overall, 5.1 channel surround was rated higher in most of the attributes. *Powerfulness* rating by Japanese listeners was not significantly different between the two reproduction setups. Also, although *height* was statistically significant, the mean rating was close to the middle point of the scale. First thing to note in Vienna result is that confidence intervals are wider than the Japanese result. This is due to less number of participants in the experiment. Except for *natural reverberation* rated rather close to the middle point of the scale, all attributes were statistically significant in supporting high ratings for the 5.1 channel surround sound reproduction.

5. Conclusion

In order to investigate the psychological impression of sounds recorded using different surround microphone arrays, subjective listening experiments were conducted at three locations in Japan and at Vienna using the method of pairwise comparison by Scheffé.

From the analyses, it was observed that microphone arrays had significant differences in effects on the listeners' subjective perceptual responses. As for the microphone arrays with higher *powerfulness* and *width*, Decca Tree & Omni Square Mid was commonly chosen for two musical programs. Double MS was chosen to have less *powerfulness* and *preference* among the microphone arrays. The other microphone arrays had too much variation between the two music programs, and the subjective impression for them seemed to be largely dependent on the choice of music programs, orchestra and performance, and the hall to be played in.

Differences in subjective impression between occupations were not confirmed. Although there were microphone arrays that were significantly different from the others in different seating positions, this also is strongly dependent on differences in music programs.

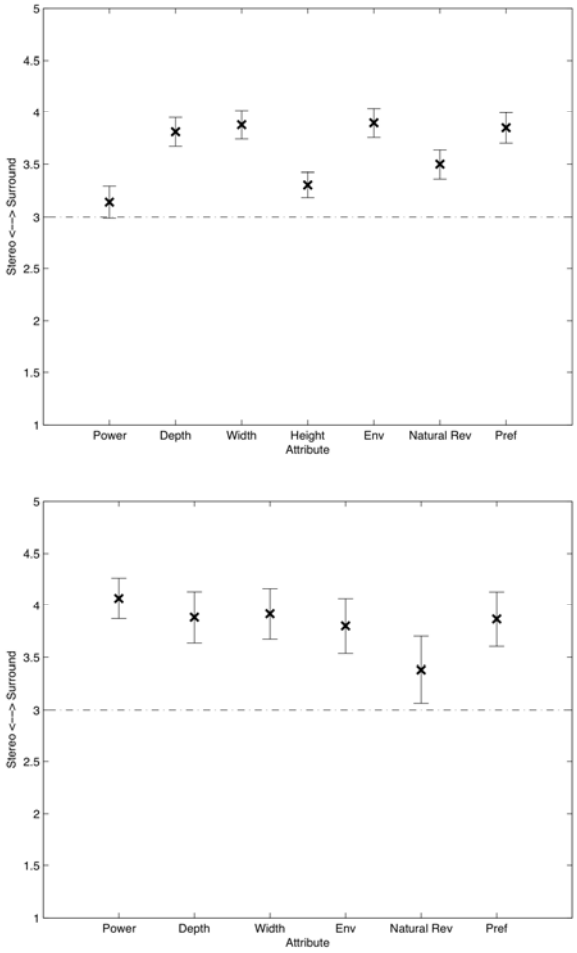


Figure 7: Comparison of subjective impression regarding 5.1 channel surround playback and 2 channel stereo playback. Crosses show mean ratings and whiskers show 95% confidence intervals around the mean. The results were collected in Japan (upper panel) and Vienna (lower panel).

An attempt of predicting global impression of the recordings from the local impressions with Multiple Regression Analysis was done. The result suggests that high *preference* is related to low softness of the recording.

Subjective comparison of 2 channel stereophonic reproduction and 5.1 channel surround sound reproduction was done as well. Statistically significant differences were observed on most of the attribute scales used. Especially, listeners heard 5.1 channel surround sound as having larger horizontal width and natural envelopment.

6. References

[1] SATOH, Shin. *Statistical Sensory Testing*. Union of Japanese Scientists and Engineers Publishing. 1985. (Available only in Japanese language)

[2] Søren Bech and Nick Zacharov. *Perceptual Audio Evaluation — Theory, Method and Application*. John Wiley & Sons. 2006.

[3] Geoff Cumming and Sue Finch. Inference by Eye: Confidence Intervals and How to Read Pictures of Data. In *American Psychologist*, vol.60, no.2, pp.170–180. 2006.

[4] American Psychology Association. *APA Style Manual*. 2004.