# DEVELOPING CONSENSUS LANGUAGE FOR DESCRIBING DIFFERENCES IN AUDITORY IMAGERY ASSOCIATED WITH FOUR MULTICHANNEL MICROPHONE TECHNIQUES

**Sungyoung Kim and William L. Martens**
**Schulich School of Music, McGill University, 555 Sherbrooke Street W., Montreal, QC, Canada H3A 1E3**
sungyoung.kim@mail.mcgill.ca

## 1. INTRODUCTION

Eight listeners who had already completed a series of descriptive analysis (DA) sessions on the stimuli employed in this study (Martens & Kim, 2007) made ratings on the stimuli during two experimental test sessions separated in time by 6 months using the bipolar adjective scales that resulted from those DA sessions. This retest was initiated to examine the consistency with which the eight listeners could make ratings on three of the bipolar adjective scales that had been used to describe perceptual differences within a set of 32 stimuli comprising solo piano performances captured using four different multichannel microphone techniques. It was hypothesized that the auditory attributes associated with these bipolar adjective scales may represent relatively permanent perceptual characteristics of reproduced musical sound that might be assessed subjectively in a similar manner over time for a given stimulus domain. If this were so, then the ratings that a given listener produces on one occasion for a restricted set of stimuli should be highly correlated with the ratings that same listener produces on another occasion, removed in time from the first so as to represent an independent assessment of the characteristics of those stimuli. The reliability of the listeners' ratings was examined in a number of ways. First, inter-subject consistency was examined using Procrustes Analysis (Dijksterhuis ,1996). Then, for a selected subset of listeners showing highest agreement with each other, intra-subject consistency was examined through comparison of ratings made by each of these eight listeners before and after the 6-month break between rating sessions. Finally, Principal Component Analysis was employed in order to find relatively stable perceptual components underlying the listener's ratings.

In the study to be described in this paper, the experimental variable that was under direct control was the multichannel microphone technique that was used to record a selection of solo piano performances. Another important factor here was the selection of musical program material to be used in evaluating the results of using the microphone techniques to be evaluated. Previous reports on this project have already given more in-depth introduction to these issues (Martens & Kim, 2007), and only details relevant to this particular longitudinal study will be provided here.

## 2. METHOD

### 2.1 Listeners

A total of eight masters students in the Sound Recording program of McGill University participated in the listening experiments. While these students could not be regarded as experts either in sensory evaluation nor in sound recording practice, they were all engaged in the training that follows the *Tonmeister* tradition, in which they develop skills in microphone placement and in aural evaluation of the results.

### 2.2 Attribute Ratings

The same eight individuals who participated in this longitudinal attribute rating study completed a verbal elicitation task using a triadic comparison method to explore the adjectives that could be used to describe differences between solo piano performances captured using four different multichannel microphone techniques. The attribute rating scales employed in this study were anchored by the following pairs of bipolar adjectives upon which the eight listeners had reached some consensus: Wide–Narrow, Focused–Diffused, and Tight-Bass–Muddy-Bass.

## 3. RESULTS

The mean ratings over the two sessions for each individual listener were submitted to Procrustes Analysis (Dijksterhuis, 1996) in order to gauge how similarly the attributes scales were being used by the eight listeners. First the centroid response dataset was calculated from the combined ratings of all eight listeners. This centroid was used as a basis for comparison, and Procrustes Analysis (PA) was used to determine a linear transformation (translation, reflection, orthogonal rotation, and scaling) of the points in each individual dataset to best conform them to the points in this centroid dataset. The "goodness-of-fit" criterion is a dissimilarity measure D, which is the sum of squared deviations of the individual dataset from the centroid response dataset. For each of the listeners the analysis produced a minimized value of D, standardized by a measure of the scale of the centroid (i.e., the sum of squared elements of a centered version of the centroid response dataset). The average obtained dissimilarity for the five listeners who were in best agreement was D = 0.390,

while the average dissimilarity for the three listeners who were judged to be outliers was D = 0.534.

Pearson correlation coefficients between the first and second ratings on each of the three attributes were calculated for each of the five selected listeners (who were in best agreement according to PA results). The observed correlations are plotted in Figure 1, with the correlation coefficient value of r=.345 marked by the horizontal dashed line. Correlation coefficients smaller than criterion might indicate either that listeners in these cases were inconsistent in how they understood the attributes on which they were required to make their ratings, or that they simply were not able to make consistent magnitude estimates for these attributes, though they might have understood well the meaning of the anchors defining the extremes for each attribute scale. Regardless of the reason, three subjects, were separated out from the other five as relatively poor in producing ratings that matched their previous ratings. Since combining such inconsistent perceptual responses would provide a poor definition of the auditory attributes under investigation here, the ratings from these three relatively inconsistent subjects were excluded from the subsequent Principal Components Analysis (PCA) designed to examine the underlying perceptual structure for the obtained attribute ratings.

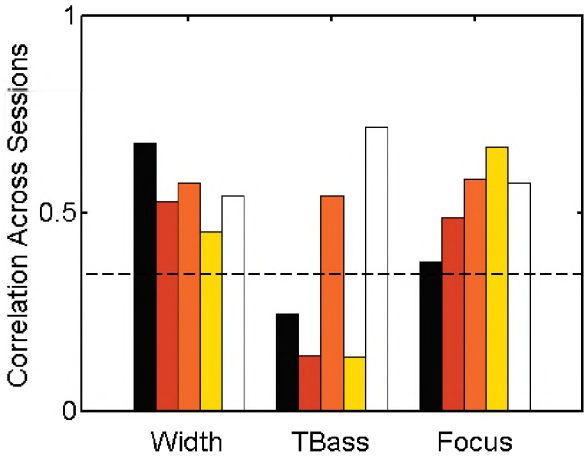


**Fig. 1. Pearson correlation coefficients between the first and second ratings on each of the three attributes were calculated for each of the five selected listeners. The horizontal dashed line indicates the criterion for statistical significance (at probability alpha < .05 of incorrectly retaining the null hypothesis).**

The PCA results shown in Figure 2 reveal the relationship between these three sets of collected attribute ratings, with ratings from the two separate sessions treated as distinct variates (i.e., a matrix of 6 columns was submitted to PCA, with a total of 160 rows containing ratings on 32 stimuli from each of 5 listeners). Loadings on each PC are plotted separately for each rating session, so there are two corresponding symbols for each of the three attributes.

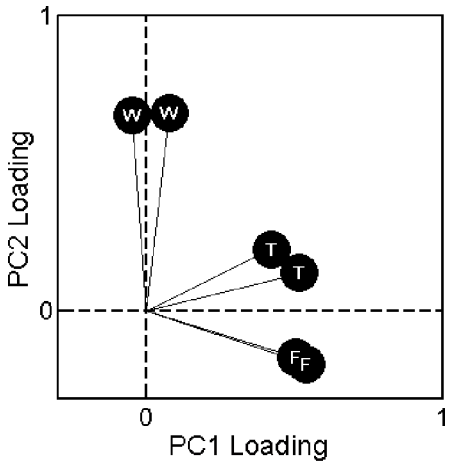


**Fig.2. Attribute loadings on the first two principal components (PCs) plotted separately for each of two rating sessions, with symbols corresponding to each of the three attributes labeled as follows: W for Width, T for Tight-Bass, and F for Focus.**

## 4. DISCUSSION

It was shown that the attributes found most salient in a previous study could be reliably rated in well separated sessions by 5 out of the 8 listener tested here. These 5 listeners were also in good agreement on how to use the attribute scales in describing the restricted set of 32 stimuli presented here. Only for one of the attribute scales (Tight-Bass) were ratings not significantly correlated between first and second sessions. For the other two attributes, (Width and Focus), ratings produced by each of the five selected listeners always were found to be correlated significantly with the ratings that the same listener produced on another occasion 6 months later. The PCA results indicate that a relatively stable perceptual structure may be said to underlie the attribute ratings obtained on such temporally separated occasions.

## REFERENCES

Martens, W. L., and Kim, S. (2007) Verbal Elicitation and Scale Construction for Evaluating Perceptual Differences between Four Multichannel Microphone Techniques. In Proc. Audio Engineering Society 122nd Int. Conv., (Vienna, Austria), Preprint 7043.

Dijksterhuis, G. (1996) Multivariate analysis of data in sensory science, volume 16 of Data handling in science and technology, chapter 7: Procrustes analysis in sensory research, pages 185–217. Elsevier, Amsterdam, The Netherlands.

## ACKNOWLEDGEMENTS

This investigation was supported by grant No. 110403 awarded to the first author by the FQRSC. Additional support was provided by the CFI New Opportunities Program, and also by McGill University's Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT).