Audio Engineering Society

Convention Paper 6190

Presented at the 117th Convention
2004 Oct 28-31 San Francisco, USA

A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model

Sean E. Olive, AES Fellow

Harman International Industries, Inc., Northridge, CA, 91329, USA
solive@harman.com

ABSTRACT

A new model is presented that accurately predicts listener preference ratings of loudspeakers based on anechoic measurements. The model was tested using 70 different loudspeakers evaluated in 19 different listening tests. Its performance was compared to 2 models based on in-room measurements with 1/3-octave and 1/20-octave resolution, and 2 models based on sound power measurements, including the Consumers Union (CU) model, tested in Part One. The correlations between predicted and measured preference ratings were: 1.0 (our model), 0.91 (in-room, 1/20th-octave), 0.87 (sound power model), 0.75 (in-room, 1/3-octave), and −0.22 (CU model). Models based on sound power are less accurate because they ignore the qualities of the perceptually important direct and early-reflected sounds. The premise of the CU model is that the sound power response of the loudspeaker should be flat, which we show is negatively correlated with preference rating. It is also based on 1/3-octave measurements that are shown to produce less accurate predictions of sound quality.

1 INTRODUCTION

Properly controlled loudspeaker listening tests are time-consuming, expensive and difficult to conduct. An alternative is to use a model that predicts listeners' subjective ratings based on objective measurements of the loudspeakers.

In Part One [1] we reviewed three loudspeaker models that predict sound quality ratings proposed by different authors. Three quite different approaches are taken in how and where the loudspeaker should be measured. One approach is to predict the sound quality using sound power measurements, with the underlying assumption being that the total radiated sound power largely determines the loudspeaker's

perceived quality in a room [2]-[5] The second approach favored by Staffeldt et al. [6]-[8] and Gabrielsonn et al. [9]-[11] is to model the loudspeaker's sound quality using in-room loudspeaker measurements. The third approach proposed by Toole [12]-[13] and tested in this paper, is to predict the loudspeaker's sound quality using a comprehensive set of anechoic measurements. Klippel's model uses a hybrid approach combining the free-field on-axis response with an in-room or predicted in-room response [14]-[15].

A characteristic common to all of the above models, except ours, is the use 1/3-octave loudspeaker measurements. Our hypothesis is that models based on 1/3-octave data are inherently less accurate because the human hearing has much better resolution than this. We test this by comparing the performances of two in-room models based on 1/20 and 1/3-octave data. The results have broad ramifications throughout the audio industry because 1/3-octave measurements are commonly used to diagnose and equalize loudspeakers in rooms, and are endorsed in many international standards [16].

Our second hypothesis is that sound power-based models are inherently less accurate because they do not include sufficient information to characterize the direct and early-reflected sounds at the listener's ears. The results from Part One of this paper support this hypothesis showing a correlation of -0.22 between the measured and predicted sound quality ratings based on the CU sound power model. To our knowledge, this is the first published test ever performed on the model, even though it has been used for over 30 years in Consumer Reports' loudspeaker reviews. We do not have legal permission from CU to republish any details on their proprietary, unpublished model. However, in this paper we explain why it fails by analyzing the objective measurements of loudspeakers previously tested by CU. To avoid indicting all models based on sound power, we develop a new one that is optimized to predict loudspeaker preference ratings. Our end goal is to give each model an equal chance and select the best one on the merits of how well it predicts loudspeaker preference ratings.

To clarify the goals of this paper we summarize the main research questions as follows:

1. Can a predictive model based on the anechoic measured frequency response of a loudspeaker accurately predict listener preference ratings in a typical listening room such as the Harman multi-channel listening lab?

2. What are the model's independent variables and their relative contribution to predicting listeners' preference ratings?

3. What is the relative accuracy of our model compared to predictive models based on sound power and in-room measurements?

4. Do predictive models based on 1/3-octave measurements provide more accurate or less accurate predictions of preference compared to 1/20-octave measurements?

2 MULTIPLE REGRESSION ANALYSIS

In this section a brief primer is given on multiple regression techniques so that the reader can better understand some of the terminology, principles and underlying statistical assumptions referred to in the following sections. Readers well versed in regression analysis can skip this section altogether.

2.1 A Primer on Multiple Regression

Regression analysis is a popular and mature multivariate statistical method first used by Legendre in 1805, but popularized by Pearson in 1903 [17]. It is used to predict the value of a single dependent variable using one (simple regression) or more (multiple regression) independent variables. Multiple regression assumes that the dependent, and usually the independent variables as well, are both metric. Metric variables are measured on interval-ratio scales as opposed to nominal categories. When the data are nonmetric, or involve more than one dependent variable, other multivariate techniques such as canonical correlation, multiple discriminate analysis and conjoint analysis may be more appropriate alternatives.

In multiple regression analysis, each independent variable is weighted to maximize its ability to predict the value of the independent variable. Their respective weights denote the relative contribution and influence of each factor on the value of the outcome variable. The set of weighted independent

variables are known as the regression variate and define the model expressed as equation 1.

$$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ...b_nX_n \quad (1)$$

where Y_1 is the predicted dependent variable, X_1 through X_n are different independent variables and b_1-b_n is the weight or coefficient for each independent variable. The term b_0 is a constant known as the y-intercept.

The key to designing accurate and robust models is selecting independent variables that maximize the predictive ability of the dependent variable, while at the same time ensuring that the independent variables are not highly correlated with each other (a condition known as collinearity). Various analysis and optimization techniques are available for this [18]. Finally, regression is a linear technique with four underlying assumptions that must be met: 1) linearity in the relationship between the dependent and independent variables 2) constant variance of the error terms (residuals) 3) normality of the error term distribution and 4) independence of the error terms. Statistical tests and examination of the standardized residual plots can determine whether the assumptions have been met.

Approaches for estimating the regression variate include confirmatory and sequential searches. Sequential searches include step-wise and forward-backward elimination where various independent variables are added or deleted to the model until some criterion is met. Combinatorial approaches test all possible subsets of variables. For models that have a large number of potential variables, the number of subsets can grow significantly (10 variables = 2 ^10 or 1024 possible combinations). An algorithm known as "Leaps and Bounds" is a compromise between all subsets and forward-backward stepwise regression [19].

The accuracy of the model is based on how well the predicted values fit to or correlate with the observed values. The statistic commonly used is Pearson's correlation coefficient (r) and its related coefficient of determination (r^2). The latter represents the percentage of variance in the dependent variable accounted for by the model. The adjusted-r value takes into account the sample size and number of independent variables in the model and adjusts it accordingly. Mallow's C_p criterion is a statistic particularly useful for all-subsets since it automatically accounts for the number of independent variables and prevents selection of a model that is over-fitted. An acceptable C_p value is equal to or lower than the number of independent variables in the model [20]. A common problem with regression models is that the models are over-fitted and are not very generalizable to other samples. This can happen when the ratio of observations to number of independent variables falls below 5:1. Ideally there should be 15-20 observations for each independent variable. Another common problem occurs with models that have high multicollinearity among 2 or more variables. As the correlation between 2 variables increases above r = 0.3 there is a limit in the ability of each variable to explain and represent the unique effects on the dependent variable. As the correlation between 2 variables approaches r= 0.8 or higher, the sign of the coefficient can become reversed. An extreme case known as a singularity occurs where the correlation between two variables is 1. This prevents the estimate of any coefficients.

The final step in developing a regression model is to validate the results. The results should be generalizable to the population (of speakers) and not specific to the sample used for estimation. The most direct approach to validation is to obtain another sample from the population and determine the correspondence in results between the two samples. In the absence of a new sample, other approaches are possible.

3 MODEL DEVELOPMENT

This section defines the set of independent variables used in our model including the scientific rationale for their selection. We examine the predictive power of each variable by looking at its correlation with the preference ratings from the listening tests reported in Part One [1]. (These test results from Part One shall be referred to hereafter as "Test One"). The multicollinearity or correlation between the independent variables is also examined.

3.1 Premise of Our Model and Selection of Independent Variables

There should be a strong theoretical basis for inclusion of any independent variable in a model so that the variable's effect on the predicted outcome makes sense, and has a rational scientific explanation.

Our model is based on the premise that a loudspeaker's preference rating is related to the mean amplitude deviation in its frequency response measured around its horizontal and vertical radiating orbits. A decrease in the mean measured amplitude deviations should correspond to an increase in the loudspeaker's preference rating. All of the independent variables in our model statistically quantify amplitude deviations in the loudspeaker frequency response. Therefore, our model only considers linear distortions related to frequency response.

Our premise is supported by a substantial body of scientific evidence from previous loudspeaker studies [5]-[16], including the test results reported in Part One [1]. Together these studies show that the frequency response of the loudspeaker is the most important factor related to perceived sound quality.

3.2 Definition of Independent Variables

A total of 30 independent variables were considered as potential candidates for our model. These include four different statistical measures applied to each frequency response curves shown in fig. 1. Two of the variables are specifically related to the low frequency response of the loudspeaker.

The 7 frequency response curves represent the on-axis response (ON), the listening window (LW), the early-reflections (ER), the predicted in-room response (PIR), the sound power (SP), and two different directivity indices that represent the early reflections (ERDI) and sound power (SPDI). The rationale and calculation of the measurements was based on a statistical survey of loudspeaker setups in a large number of domestic listening rooms [21]. All the anechoic measurements have high frequency resolution (2 Hz) from 2 Hz to 20 kHz with a 1/20-octave smoothing filter applied to the raw data. Spatial averaging is used for all 7 curves (except the on-axis curve) to remove interference and diffraction effects from the measurements. This helps separate these effects from resonances, which we believe have more serious audible consequences [12]-[13], [22]-[23].

The 7 spatial averages allow independent assessment and prediction of the qualities of the direct, early-reflected and reverberant sounds at the listener. The relative importance each component has on the overall perceived sound quality of the loudspeaker is represented by its relative contribution to predicting preference in the model. The direct sound is best characterized by the loudspeaker's on-axis (ON) and listening window (LW) curves, while the early reflected and reverberant sounds can be predicted using the ER and SP curves, respectively. The predicted in-room response (PIR) represents a weighted average of the on-axis, early-reflected and sound power measurements. In previous papers (see figs. 18-20 in reference [13]), [21] the predicted in-room response has been shown to accurately correlate with the actual measured in-room response between 300-10kHz. This is reconfirmed with new measurements reported in section 6.

Throughout this paper, the nomenclature such as AAD_{ON} is used to describe the independent variables. The first term describes the statistic or metric, while the subscript term indicates the measurement curve to which the statistic is applied. Table1 summarizes each statistical metric and the frequency response measurements involved in its calculation.

3.2.1 Absolute Average Deviation

The first statistic examined for the model is the absolute average deviation (AAD), expressed in dB as defined in equation 2

$$AAD(dB) = \left(\sum_{Band=16kHz}^{Band=100Hz} \left| (y_{REF@200\text{-}400Hz} - y_{band.n}) \right| \right) \div N \quad (2)$$

where the average absolute deviation in band n is calculated from the reference level y_{REF} based on the mean amplitude between 200-400 Hz. The deviation is calculated in each 1/20-octave band over N bands from 100 Hz-16 kHz. Higher values of AAD indicate larger deviations in amplitude from our reference band. Therefore the variable should be negatively correlated with preference according to our hypothesis.

The use of a reference band of 200-400 Hz is based on an observation made in Part One (see section 4.8 of [1]). When asked to judge the spectral balance of each loudspeaker across 6 frequency bands, listeners referenced or anchored their judgments to the band centered around 200 Hz. One plausible explanation is that many of the fundamentals of instruments, including voice, fall within 200-400 Hz, and the levels of the higher harmonics are referenced to it.

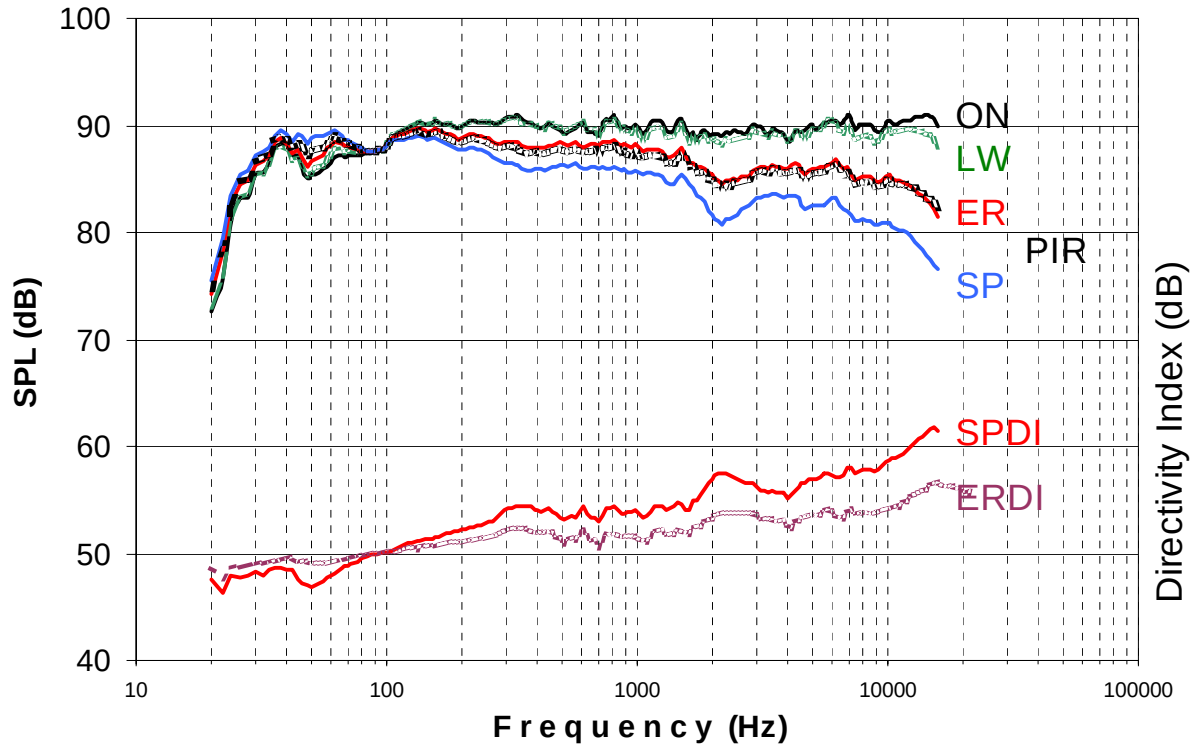

Figure 1 The 7 different frequency curves used in our model from top to bottom are: on-axis response (ON), the listening window (LW), the early reflected curve (ER), the predicted in-room response (PIR), the sound power (SP), and directivity indices (SPDI and ERDI) related to the sound power and early reflections. For details see reference [21].

Table 1- Below is a description of the 6 statistics available to our model and the loudspeaker measurement curves to which they are applied.

Statistic	Description	Measurement Applied to:
AAD	Absolute Average Deviation (dB) relative to mean level between 200-400 Hz	ON, LW, ER, PIR, SP, ERDI, SPDI
NBD	Average Narrow Band Deviation (dB) in each ½-octave band from 100 Hz- 12 kHz	ON, LW, ER, PIR, SP, ERDI, SPDI
SM	Smoothness (r^2) in amplitude response based on a linear regression line through 100 Hz -16 kHz	ON, LW, ER, PIR, SP, ERDI, SPDI
SL	Slope of Best Fit linear regression line above (dB)	ON, LW, ER, PIR, SP, ERDI, SPDI
LFX	Low frequency extension (Hz) based on -6 dB frequency point transformed to $\log_{10}$	SP relative to mean sensitivity in LW from 300 Hz – 10 kHz
LFQ	Absolute average deviation (dB) in bass response from LFX to 300 Hz.	SP relative to mean sensitivity in LW

An iterative method was used to vary both the bandwidth of the reference band as well as the range over which the deviation was calculated (from 100 Hz to 20 kHz). The final parameters used in our model produced the highest correlation with the measured preference ratings.

3.2.2 Narrow Band Deviation

The narrow band deviation is defined by equation 3,

$$NBD(dB) = \left(\sum_{Band=12\,kHz}^{Band=100\,Hz} \left| \bar{y}_{\left(\frac{1}{2}\text{OctaveBand n}\right)} - y_b \right| \right) \div N \qquad (3)$$

where $\bar{y}_{\left(\frac{1}{2}\text{OctaveBand n}\right)}$ is the average amplitude value within the ½-octave band n, y_b is the amplitude value of band b, and N is the total number of ½-octave bands between 100 Hz-12 kHz. The mean absolute deviation within each ½-octave band is based a sample of 10 equally log-spaced data points.

The final resolution and bandwidth parameters used to calculate NBD was determined using an optimization process to best predict the measured preference ratings of 70 loudspeakers. Whereas AAD measures deviations from flatness relative to the average level of the reference band 200-400 Hz, NBD measures deviations within a relatively narrow ½-octave band. NBD might be a better metric for detecting medium and low Q resonances in the loudspeaker.

3.2.3 Smoothness and Slope

For each of the 7 frequency response curves, the overall smoothness (SM) and slope (SL) of the curve was determined by estimating the line that best fits the frequency curve over the range of 100 Hz-16 kHz. This was done using a regression based on least square error. SM is the Pearson correlation coefficient of determination (r^2) that describes the goodness of fit of the regression line defined by equation 4,

$$SM = \left(\frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left(n\sum X^2 - (\sum X)^2\right)\left(n\sum Y^2 - (\sum Y)^2\right)}} \right)^2 \qquad (4)$$

where n is number of data points used to estimate the regression curve and X and Y represent the measured versus estimated amplitude values of the regression line. A natural log transformation is applied to the measured frequency values (Hz) so that they are linearly spaced (see equation 5). Smoothness (SM) values can range from 0 to 1, with larger values representing smoother frequency response curves. Therefore SM is the only predictor variable in our model that should produce positive correlations with preference according to the premise of our model. The other statistic in our model is slope (SL), which is defined as b in equation 5, that mathematically defines the regression line that best fits to the measured frequency curve. Equation 5 is defined as:

$$\hat{Y}_i = b(\ln(x_i)) + a \qquad (5)$$

where $\hat{Y}$ is the predicted value (amplitude) of the regression line at a given frequency x_i, b is the slope, and a is the y-intercept.

The raw slope value can have either negative (tilting downwards) or positive values (tilting upwards). In order to make sense to our model, we define slope (SL) in our model as the absolute difference between target slope, $b_{T\arg et}$ versus the measured slope, $b_{measured}$ as described in equation 6.

$$SL = \left| b_{T\arg et} - b_{measured} \right| \qquad (6)$$

Target slopes were determined separately for Test One and for our larger test sample (70 loudspeakers) used for the generalized model described in section 5. The target values are based on the mean slope values of speakers that fall into the top 90 percentile based on preference ratings. Target slopes are defined for each of the 7 frequency curves (see table 2). The ideal target slope for the on-axis and listening window curves (0 and –0.2) is identical for both test samples, which indicates that the on-axis curve should be flat, while the off-axis curves should tilt gently downwards. The degree of tilt varies among curves for Test One and the larger sample. Test One includes mostly 2-way designs whereas the larger sample includes several 3-way and 4-way designs that tend to have wider dispersion (hence smaller negative target slopes) at mid and high

frequencies. This suggests that the ideal target slope may depend on the loudspeaker's directivity.

Table 2 Target slopes for each frequency curve based on samples from Test One and all 70 samples combined.

Measured Curve	Target Slope Value Test One	All Tests (70 loudspeakers)
ON	0.0	0.0
LW	-0.2	-0.2
ER	-1.2	-1.0
PIR	-2.1	-1.75
SP	-1.2	-1.0
ERDI	1.0	0.8
SPDI	2.0	1.4

3.2.4 Low Frequency Extension and Quality

The low frequency extension (LFX) and quality (LFQ) of the loudspeaker are the final two variables in our model. LFX is defined by equation 7,

$$LFX = \log_{10}(x_{SP-6dB.} re: \bar{y}\_LW(300Hz-10kHz)) \quad (7)$$

where LFX is the $\log_{10}$ of the first frequency $x_{\_SP}$ below 300 Hz in the sound power curve, that is -6 dB relative to the mean level y_LW measured in listening window (LW) between 300 Hz-10 kHz. LFX is log-transformed to produce a linear relationship between the variable LFX and preference rating. The sound power curve (SP) is used for the calculation because it better defines the true bass output of the loudspeaker, particularly speakers that have rear-firing ports.

Low frequency quality (LFQ) is defined by equation 8,

$$LFQ(dB) = \left(\sum_{Band\_SP=300Hz}^{Band\_SP=LFX} \left| (y\_LW - y\_n) \right| \right) \div N \quad (8)$$

where the y is the level within each n band of the sound power curve calculated across N bands, from the lowest frequency defined by LFX up to 300 Hz.

LFQ is intended to quantify deviations in amplitude response over the bass region between the low frequency cut-off and 300 Hz. Speakers with good low bass extension may well have high deviations in amplitude response due to under/over damped alignments or incorrectly set subwoofer levels. The popular use of multiple woofers wired in parallel increases the directivity rapidly above 100 Hz, which also causes amplitude deviations in the sound power response.

3.3 Predictive Power of Variables

The predictive power of each independent variable can be determined by calculating its partial correlation with preference rating for each of the 7 frequency curves. This analysis is summarized in fig. 2 based on the 13 loudspeakers from Test One. Analysis of the larger sample of 70 loudspeakers showed similar trends.

If the premise of our preference model is well founded, all independent variables (except smoothness) should produce negative correlations with preference since larger variable values represent larger deviations from an ideal frequency response. Smoothness (SM), on the other hand, should produce positive correlations since larger values of SM indicate increased smoothness in the frequency response. These assumptions are all true for the variables NBD, LFX and LFQ, where higher values correspond to lower preference ratings. For the other variables (AAD, SL and SM), the expected magnitude and sign of the correlation vary significantly depending on which curve the metric is applied to. AAD shows the expected strong negative correlation when it is applied to the on-axis and listening window curves (i.e. a flat response produces higher preference ratings). But when applied to other measurements (ER, PIR and the two directivity indices), AAD has a weak correlation with preference. When applied to sound power, AAD shows a relatively strong but positive correlation (r = 0.6) telling us that as the sound power response becomes flatter it actually produces lower preference ratings. This finding completely contradicts the basic premise of the CU model, which defines perfection as a speaker with perfectly flat sound power. There is clearly something flawed in this premise. A better metric for assessing the quality of the sound power is smoothness, which has a correlation of 0.7 with preference.

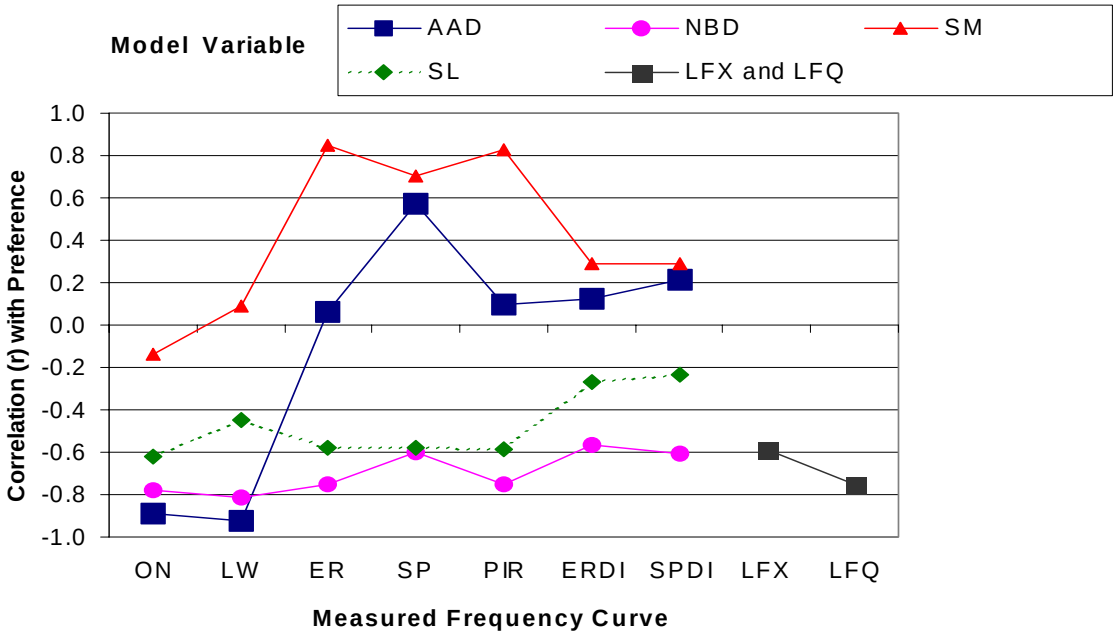

Figure 2 The correlation (r) with preference for each of the 6 independent variables applied to the frequency curves shown in Fig. 1. The data are from Test One [1].

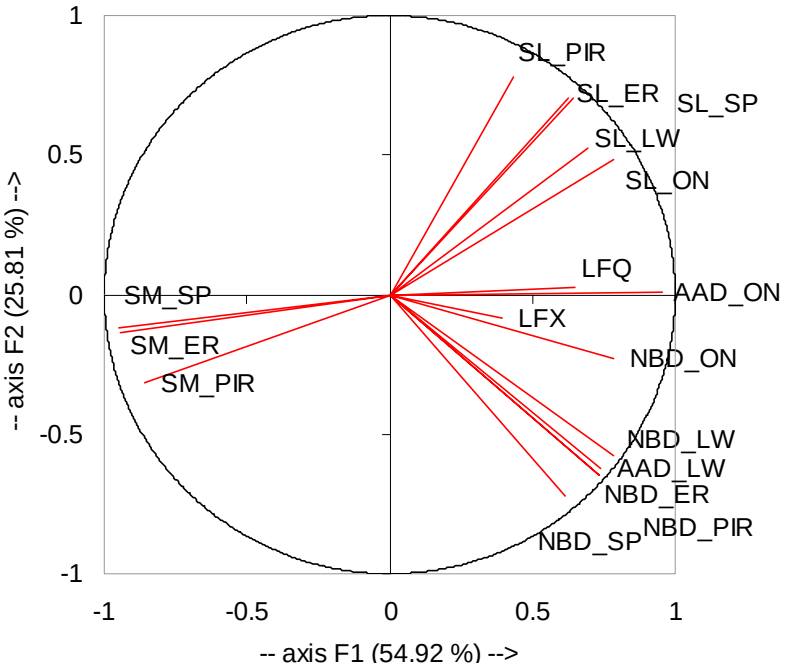

Figure 3 A correlation circle showing the mapping of 23 different independent variables into factor space 1 and 2 based on principal component analysis of the loudspeakers from Test One.

Variables that have small correlations with preference are smoothness (SM) and slope (SL) when applied to the ON and LW curves, and AAD applied to ER and PIR. The two directivity indices generally yield poor correlations regardless of which metric is applied, with the exception of NBD. In fact, the narrow band deviation (NBD) metric yields some of the highest correlations with preference, independent of the frequency curve to which it is applied.

3.4 Multicollinearity among Model Variables

An ideal regression model contains independent variables that are highly correlated to the predicted variable (i.e. preference rating) but are uncorrelated with each other [18]. The degree to which the independent variables show multicollinearity should always be assessed.

We examined the multicollinearity among the 23 independent variables considered in our model using principal component analysis (PCA). For the loudspeakers in Test One, five factors account for 97% of the variance among the 23 independent variables. The interdependence among the independent variables is plotted using a correlation circle in fig. 3.

It shows a projection of the 23 independent variables mapped into 2-dimensional factor space. Factors 1 and 2 account for almost 81% of the variance represented within our model independent variables. Variables strongly associated with factors 1 and 2 are located far from the center along the x-axis and y-axis, respectively. Close proximity between two variables indicates they are highly correlated with each other. Variables opposite to the center have negative correlation with each other. As expected, the metrics smoothness (SM) and narrow band deviation (NBD) are negatively correlated with each other. Slope (SL) and NBD appear also to be negatively correlated with each other and are associated with factor 2. Variables highly associated with Factor 1 include metrics applied to the on-axis sound (AAD_ON, NBD_ON) and to a lesser extent bass extension (LFX) and quality (LFQ).

A certain degree of colinearity and redundancy exists among our 23 variables based on their close proximity to each other. Metrics that are closely related to one another (e.g. AAD and NBD), particularly when applied to the same curve or a related curve (e.g. ER versus SP, SPDI versus ERDI), tend to produce the greatest amount of colinearity.

The variables NBD_ON, AAD_ON, LFX and model metrics applied to the predicted-in room response are all desirable predictor variables because they are strongly correlated with factors 1 and 2, but not overly correlated with each other.

4 ANECHOIC MODEL FOR TEST ONE

In this section we develop a model using anechoic measurements to predict the preference ratings of the 13 loudspeakers in Test One.

Multiple regression analysis of the 23 independent variables was performed using a program that calculates all possible models to determine the best one for a given number of variables (we chose 2-6 variables). The best-fit model uses 5 variables producing a correlation of 1.0 (r = 0.995). The model's equation is:

$$\text{Pref.Rating} = 6.04 - 0.67*\text{AAD\_ON} - 1.28*\text{LFX} - 0.66*\text{LFQ} + 4.02*\text{SM\_ON} + 3.58*\text{SM\_SP} \quad (9)$$

Fig. 4 shows a plot of the measured versus predicted preference ratings showing that the measured values closely fit the predicted values from the model. The model accounts for 99% of the variance in the observed preference ratings. The adjusted-r value (0.96) is also high. The Mallow's C_p value is 4 indicating that the model is not too over-fitted for the number of variables used. The RMS error of the predicted rating is very small, 0.26 preference rating.

An ANOVA test indicated a very small probability that the model's variables could produce the predicted results due to chance ($F = 137.34$, $p < 0.0001$).

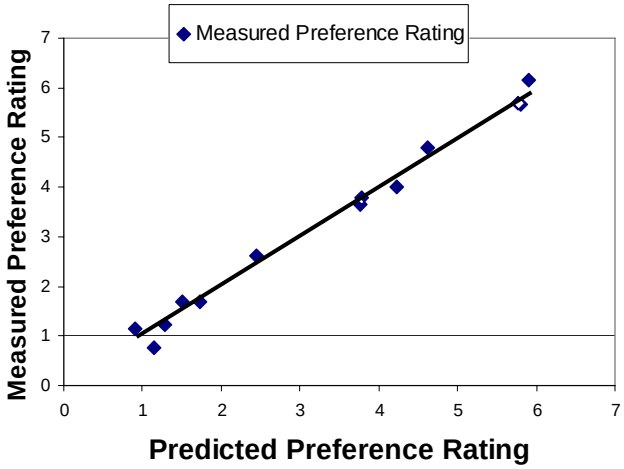

Figure 4 Plotted are the measured versus predicted preference ratings from Test One based on our anechoic model described by equation 9.

4.1 Evaluating the Anechoic Model

The coefficients in our model as described in equation 9 all have the expected sign according the premise of our model. All variables, except smoothness (SM), have negative coefficients indicating that smaller deviations in amplitude response produce an increase in preference ratings. The two variables defined by smoothness both have positive signs, indicating that higher values of smoothness produce large values of preference. All of the underlying assumptions of our model have been met.

We now consider the relative contribution each variable has in predicting loudspeaker preference. Using the standardized coefficients for each variable in the model, we calculated the percentage each variable contributes in predicting the preference rating of the loudspeaker (see table 3).

The variables related to the smoothness (SM) and average absolute deviation (AAD) of the on-axis curve have a combined weighting of 45% in our model. This tells us that the flatness and smoothness of the direct sound is an important factor in predicting sound quality. The next largest contributor is the smoothness of the sound power (SM_SP) weighted at 30%. The remaining two variables related to low frequency deviations contribute a combined 25% to our model (LFQ=17%, LFX=6%).

Finally, we examined the standardized residuals and found them to be normally distributed with constant and independent variance.

Table 3 –The proportional weighting of each variable in the model applied to Test One.

Model Variable	Proportional Contribution in Model (%)
AAD_ON	18.64
LFX	6.27
LFQ	18.64
SM_SP	30.12
SM_ON	26.34
TOTAL	**100.00**

5 TESTING THE ANECHOIC MODEL

To test the generalizability of our model, we applied our model to an additional set of 57 loudspeakers evaluated in 18 different tests. Later, we combine this sample with the 13 speakers from Test One to develop a generalized model based on 70 loudspeakers.

5.1 Selection of Loudspeakers

The selection of 70 loudspeakers was based on the competitive samples purchased for performance benchmarking tests performed for each new JBL, Infinity and Revel model.

The price range of samples varied from $100 to $25,000 per pair and includes models from 22 different brands from 7 different countries: United States, Canada, Great Britain, France, Germany, Denmark and Japan. The loudspeakers included designs that incorporated horns and more traditional designs configured as 1-way to 4-ways. Some used waveguides, while others did not. The sample also included four professional 2-way active models referred to as “near-field” monitors. The vast majority of the speakers were forward-facing driver designs, with one electrostatic dipole sample.

5.2 Listening Tests

The preference ratings for the 70 loudspeakers were based on a total of 19 listening tests conducted over the course of 15 months. All of the tests were performed under identical double-blind listening conditions, as described in Part One (see section 3). Controlled variables common to all 19 tests include listening room, program material, loudspeaker and listener location, playback level, experimental procedure and loudness normalization between speakers.

The preference ratings in one of the tests are based on the mean preferences of 268 listeners (12 trained and 256 untrained) reported in [24]. All other tests were done using trained listeners

There is one important difference between Test One and the other tests that is related to the experimental design. In Test One, we carefully controlled contextual effects by using a relatively large loudspeaker sample (13) and we compared all possible loudspeaker combinations in 13 separate tests [1]. While the procedure is labor-intensive and costly, it assures that any scaling errors related to contextual effects are balanced and minimized (see section 4.6 of [1] for a discussion of context effects].

In the new set of 18 listening tests, each test was performed as an independent evaluation of four different samples, each model in the same competitive product and price class. There were no anchors or references included in the tests to calibrate the preference scale, which is a relative one. We do not know the extent to which the accuracy of our generalized model is limited by a loss of precision from pooling subjective data from 18 unrelated tests. Possible solutions for comparing large samples and calibrating the preference scale are discussed in section 11.

5.3 Generalized Anechoic Model

Our anechoic model described in equation 9 was applied to the new larger loudspeaker sample and produced a correlation of 0.70 between the predicted and measured preference ratings. The lower correlation was likely related to the model being too tightly fitted to the small sample (13 loudspeakers) and/or the loss of precision from combining subjective data from 18 unrelated tests. A more generalized model was necessary to accurately predict the ratings of our 70-loudspeaker sample.

Using 23 independent variables, a model using 4 independent variables was developed that has a correlation of 0.86 for the 70-loudspeaker sample. The model is described by equation 10. A plot of the measured versus predicted preference ratings is shown in fig. 5.

$$\text{Pref. Rating} = 12.69 - 2.49 * \text{NBD\_ON} - 2.99 * \text{NBD\_PIR} - 4.31 * \text{LFX} + 2.32 * \text{SM\_PIR} \quad (10)$$

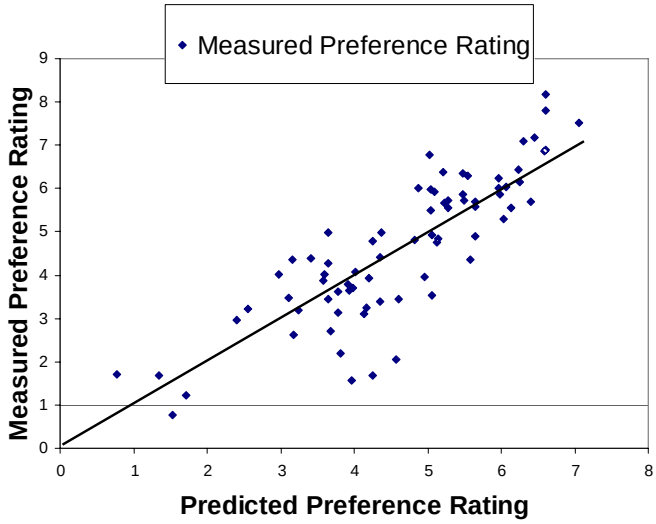

Figure 5 The measured versus predicted preference ratings based on the generalized model described in equation 10.

An ANOVA test indicated a very small probability that the model's variables could predict the ratings due to chance alone (4, 79; $F=54.88$, $p<0.0001$). The residual error from the model is 0.8 preference ratings. Examination of the residuals showed them to be normally distributed with constant and independent variance.

The standardized coefficients were used to determine the proportional contribution of each variable towards predicting preference (see table 4). The mean narrow band deviations in the on-axis curve

contribute a significant amount (31.5%) to the predicted preference rating. The narrow band deviation (NBD) and smoothness (SM) of the predicted in-room response (PIR) contributes a combined 38%, with low frequency extension contributing 30.5%. These findings are consistent with the model developed in Test One, where similar proportional weightings were given to the variables that describe the direct, early reflected, reverberant and low frequency sounds.

Table 4 - The proportional weighting of each variable in the Generalized Model

Model Variable	Proportional Weight in Model (%)
NBD_PIR	20.5
NBD_ON	31.5
LFX	30.5
SM_PIR	17.5
	TOTAL 100.0

6 COMPARISON OF MODEL BASED ON IN-ROOM MEASUREMENTS

Our first two research questions in section 1 have now been answered. We have shown two models based on a set of anechoic measurements that accurately predict loudspeaker preference ratings, and identified those variables that contribute to predicting sound quality.

We now turn our attention to the first part of research question three that asked, what is the relative performance of models based on in-room measurements? Two in-room models are developed that use the same measurements but are post-smoothed using a 1/20-octave and 1/3-octave filter. This is done to answer research question 4, which posits the hypothesis that models based on 1/3-octave measurements are inherently less accurate in predicting sound quality.

In-room measurements were made of the 13 loudspeakers from Test One using the exact physical setup used for the listening tests. A diffuse-field microphone was positioned at the listener's chair, at average ear height, 3 m. away from the loudspeaker. The loudspeaker was placed 1.2 m from the rear wall, slightly off-center from the side walls of the room. A total of 9 measurements were taken at 0º, ± 10º, ± 20º and ± 30º horizontal, and ± 10º vertical. The measurements were averaged to produce a measured in-room response. All measurements were made using 2 Hz frequency-resolution with the 1/20-octave and 1/3-octave smoothing filter applied to the raw data.

Fig. 6 compares the anechoic measured predicted in-room response (PIR) of loudspeaker with its measured in-room response using 1/20 and 1/3-octave smoothing.

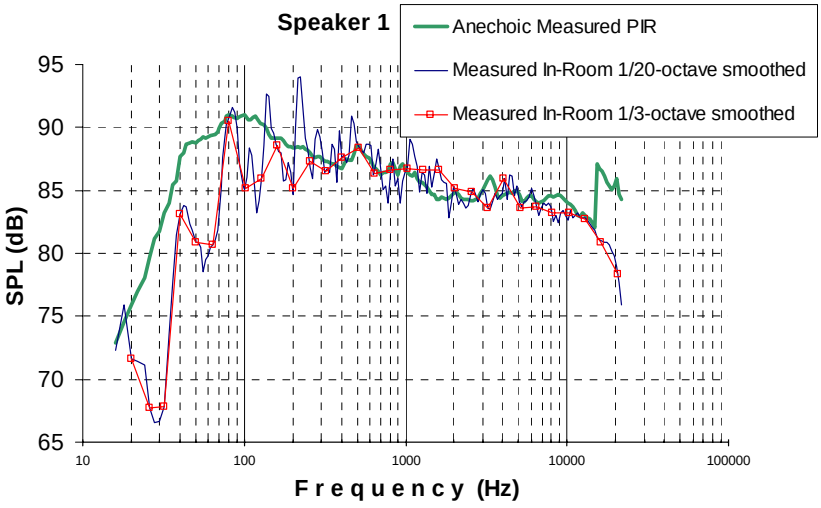

Figure 6 The predicted in-room response (PIR) (1/20-octave smoothing) of a loudspeaker versus its measured in-room response 1/20-octave and 1/3-octave smoothing. All measurements have frequency resolution of 2 Hz.

There is very good agreement between the curves except below 300 Hz and above 10 kHz, where room mode effects and room absorption are not accounted for in the predicted in-room response.

Fig. 7 shows the average difference between the predicted and measured in room response (1/20-octave resolution) based on an average of 13 loudspeakers in Test One. Again we see that the predicted in-room response closely matches the actual measured in-room response. The good agreement between the two suggests that the differences between models using predicted or measured in-room curves should be small as well.

Three predictive models were developed using three different frequency response curves: 1) the anechoic predicted in-room curve 2) the measured in-room curve 1/20-octave smoothed and 3) the same as 2) but using 1/3-octave smoothing. With only one frequency curve available to the model there are just 5 possible independent variables: NBD, SM, SL, LFX and LFQ, all applied to the predicted or measured in-room curves. For the measured in-room curves, LFX and LFQ values were calculated based on the measured in-room curve, rather than using the sound power response referenced to the listening window curve, as in our earlier models.

The correlation coefficients r, r^2 and the adjusted-r^2 values are shown for the best-fitted models in fig 8. The correlation between predicted and measured preference rating is highest for the model based on the anechoic predicted in-room response (r=0.94) followed by the measured in-room response using 1/20-octave resolution (r =0.91). The model using the 1/3-octave smoothed in-room data had the lowest correlation with measured preference (r=0.75).

The explanation for this result can be found by comparing the correlation values for the five independent variables available to the model (see fig. 9). The 1/3-octave data produce consistently lower correlations with preference ratings across four of the five variables. The errors from the coarse smoothing misrepresent the true differences in the loudspeakers' frequency responses defined by smoothness, slope, narrow band deviation and low frequency deviation. The human ear is clearly better at distinguishing these differences than a 1/3-octave measurement.

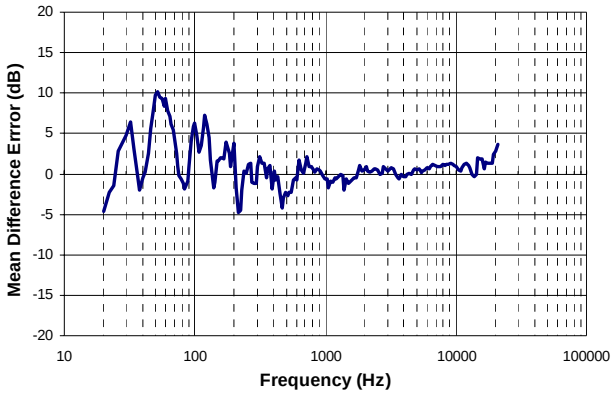

Figure 7 The difference between the predicted and measured in-room response based on 13 loudspeakers in Test One.

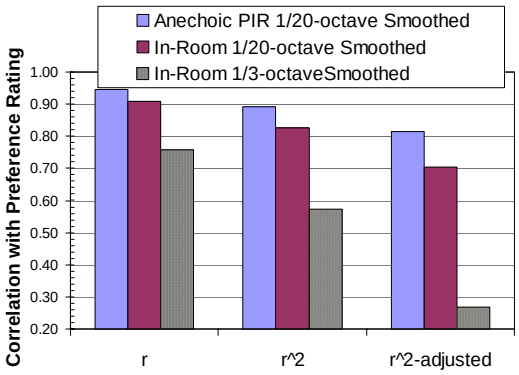

Figure 8 The correlations of 3 models based on the predicted and measured in-room responses with 1/20 and 1/3-octave smoothing.

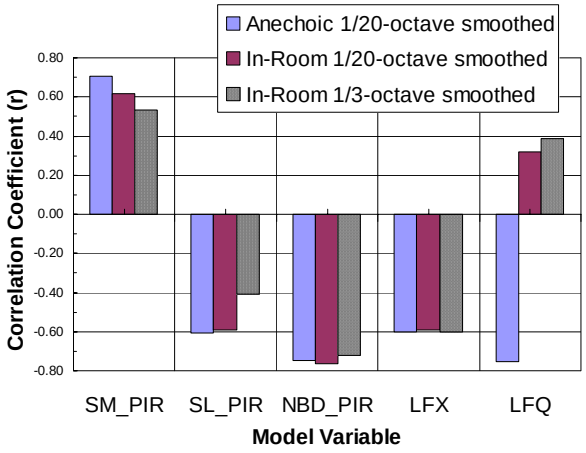

Figure 9 The correlation with preference for the 5 independent variables available to the 3 models in fig. 8. based on predicted and in-room measurements.

An important distinction needs to be made between the 1/3-octave measurements shown here (2 Hz resolution data that is post-smoothed) and measurements made with 1/3-octave analyzers that use filters having fixed center frequencies. These devices will produce measurements less accurate than ours and we expect that they would yield even less accurate models.

7 COMPARISON OF MODELS BASED ON SOUND POWER MEASUREMENTS

The second part to research question three asks whether models based on sound power measurements can accurately predict loudspeaker preference ratings. The CU model tested in Part One is only one example of a model based on sound power. First, we re-visit the CU model to explain why it failed to accurately predict loudspeaker preference. What we learn from this may help to improve the new model based on sound power developed in section 7.3.

7.1 CU Model Revisited

Some evidence was uncovered in section 3.1 suggesting the underlying premise of the CU model may be flawed. The 100-point accuracy score it predicts is based, in large part, on how far the power response deviates from a flat target curve over a specific bandwidth. The deviations are calculated based on loudness errors using a simple loudness model [2]. In section 3.1, one variable related to flatness of sound power, absolute average deviation (AAD) from flat (relative to 200-400 Hz) showed a positive correlation (0.6) with preference rating. This implies that flat sound power response is not a good target for achieving higher preference ratings.

To explore this further the 13 loudspeakers from Test One are plotted in fig. 10 (a). Plotted for each speaker are the smoothness and raw slope values for sound power. The raw slope values are used to remove any presumption about what is an ideal target slope. The raw slope values for the listening window (LW) are plotted as well to study the interrelationship between preference and the qualities of the listening window and sound power curves. The loudspeakers are plotted (from left to right) in descending order of their measured preference rating, which is transformed by a factor of 0.1 to fit the ratings on the same scale as the independent variables.

The plot clearly shows a monotonic relationship between the three independent variables and preference. As the slope of the sound power rises towards 0 (which is the CU model's target), the preference rating falls. The slope of the listening window (LW) more or less tracks the sound power slope. Higher slope values, indicating a rising response on-axis, correspond to lower preference ratings. Lower smoothness values in the sound power also correspond to lower preference ratings. All of these observations nicely conform to the underlying premise of our preference model defined in section 3.1.

Fig. 10(b) shows the same data as fig. 10 (a) except here the data are plotted from left to right in descending order of the predicted accuracy rating according to the CU model. The accuracy rating has been transformed by a factor of 0.01 to fit on the same scale. The first noticeable difference between the graphs is the reversal in the directions of the curves denoting the raw slope values and smoothness values for the sound power curve. As the sound power slope falls from zero (flat), the speaker's predicted accuracy rating tends to drop. This clarifies the basic premise of the CU model where increased flatness in sound power response is associated with higher accuracy ratings.

The criteria on which the CU accuracy and preference models are based seem incompatible. This becomes more obvious in fig. 11 where we plot the correlations with accuracy versus preference for five independent variables applied to the sound power and listening window curves. In all five cases, variables that are positively correlated with preference are negatively correlated with sound accuracy, and vice versa. Speaker characteristics that produce higher preference ratings produce lower accuracy ratings, including smooth sound power response with a downward tilt (i.e. slope < 0) and a smooth, relatively flat frequency response in the listening window. The reason that the CU model penalizes attributes positively associated with preference is more related to its failure to include other measures of loudspeaker sound quality besides flatness of sound power (which is unfortunately an incorrect premise). Some of the blame must be attributed to the use of 1/3-octave measurements, which we have shown produce less accurate predictions of sound quality.

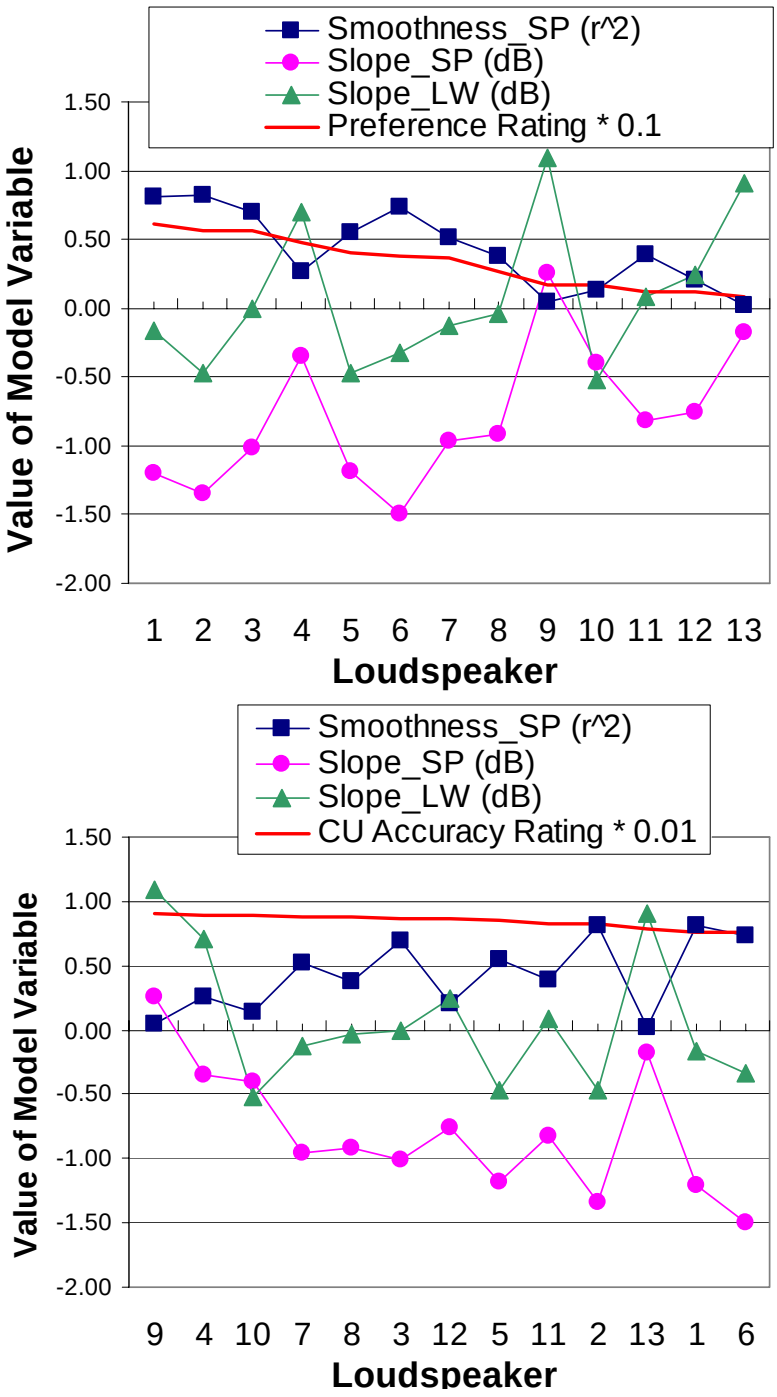

Figure 10 The values for the variables smoothness (SM) and slope (SL) applied to sound power (SP), and the values of the slope of the listening window (LW) for loudspeakers in Test One. The raw slope values are shown. The top graph plots the speakers in descending order of measured preference rating. The bottom graph plots the speakers in descending order of their accuracy rating. Both preference and accuracy ratings have been linearly transformed to fit on the same scale.

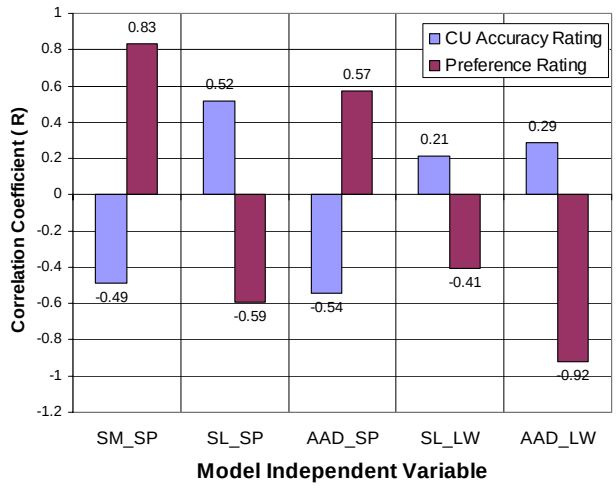

Figure 11 The correlation between measured preference and CU accuracy ratings for 5 independent variables that measure amplitude deviations in the sound power and listening window. The data are based on the loudspeakers in Test One.

7.2 A Modified Accuracy Model Based on Preference Model Variables

Loudspeaker manufacturers may want to determine in advance the accuracy rating of a loudspeaker before it is sold and tested by CU. In most cases, replicating the CU measurements and model is not likely feasible or practical. The author wondered whether he could develop a model to predict CU accuracy ratings using the existing measurements and independent variables from Test One.

Using the 13 loudspeakers from Test One, a model was developed that predicted their measured accuracy ratings with a correlation of 0.99 using 6 variables (see equation 11).

$$\text{Accuracy Rating} = 69.26 - 12.88 * \text{AAD\_ON} + 40.69 * \text{NBD\_PIR} + 19.97 * \text{LFX} + 14.91 * \text{SM\_PIR} - 13.03 * \text{SL\_ON} + 29.9 * \text{SL\_ER} \quad (11)$$

Fig. 12 shows the measured versus predicted accuracy ratings for each of the 13 loudspeakers.

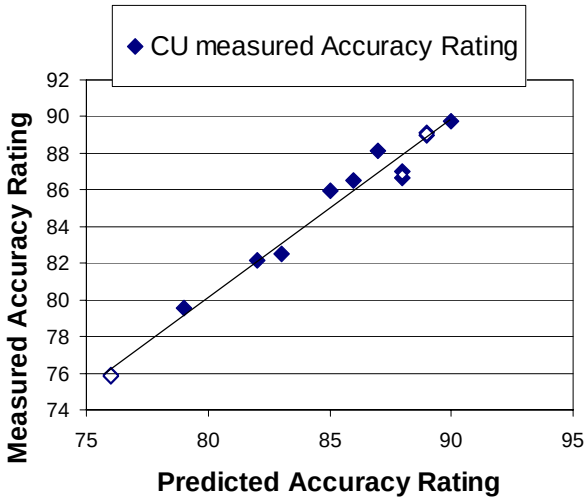

Figure 12 A plot of the measured versus predicted accuracy ratings using our modified CU model described in equation 11. The measured values are based on published CU accuracy ratings for the speakers in Test One.

It is difficult to determine if the signs of the coefficients in equation 11 are correct, given that the only known premise of the model is that the sound power be flat. Using the premise of our preference model, we see that the signs of the coefficients are reversed for 3 of the 6 variables: NBD_PIR, LFX, and SL_ER. According to this version of the accuracy model, higher accuracy ratings will be given to loudspeakers that have less extended bass, more narrow band amplitude deviations in the predicted in-room response and slopes in the early reflected curves that deviate from our target slope of -1.

This is just another way of demonstrating that designing a loudspeaker to achieve high accuracy ratings according to the CU model may result in serious compromises in sound quality that will produce lower preference ratings. The listening test results from Part One is a clear example, where the correlation between preference and accuracy rating was found to be r= –0.22. The speakers with the flattest sound power had rising frequency responses on-axis and/or reduced low frequency extension. Both are necessary compromises to achieve flat sound power for speakers that have a rising

directivity at higher frequencies. Such speakers represent the vast majority of all speakers sold. A speaker with constant, flat directivity could theoretically satisfy the flat sound power criterion and still achieve high preference ratings, so long as it had a smooth on-axis response well-maintained off-axis. However, such speakers are not widely available.

7.3 A Better Sound Power Model

It may be possible to design a better sound power model if we ignore the premise that the loudspeaker's sound power response should be flat and we use higher resolution, 1/20-octave measurements. If we can design such a model, how accurate are its predictions of sound quality compared to our anechoic and in-room measurement models?

Two sound power models were developed to predict the measured preference ratings from Test One and the 70 speakers used in our earlier generalized preference model. The three variables available to our model are: narrow band deviations (NDB), smoothness (SM) and slope (deviation from a target slope) all applied to the sound power curve. The two models are defined by equations 12 and 13.

$$\Pr ef.Rating = 2.63 - 2.86 * NBD\_SP + 5.15 * SM\_SP + 0.417 * SL\_SP \quad (12)$$

$$\Pr ef.Rating = 6.7 - 6.99 * NBD\_SP + 2.83 * SM\_SP + 1.15 * SL\_SP \quad (13)$$

For the first and second models the correlations with the measured preference ratings are 0.87 and 0.79, respectively. The correlations are quite respectable compared to the CU model (r= –0.22) also based on sound power. Our sound power model is also better than the in-room model (r= 0.75) based on 1/3-octave smoothed data but not quite as good as the in-room model based on 1/20-octave smoothed measurements (r= 0.91). Yet, our best sound power model does still not perform as well as our preference model that uses the complete set of anechoic data (r= 1.0).

Figs. 13(a) and (b) show the measured versus predicted preference ratings for the two models, respectively. The coefficients of the model all meet the same underlying assumptions made in our preference model defined by equations 9 and 10. The model states that the predicted preference rating will decrease as the mean narrow band deviations in the sound power increase and with larger deviation from the defined target slope (-1). An increase in the smoothness of the sound power response corresponds to an increase in the predicted preference rating.

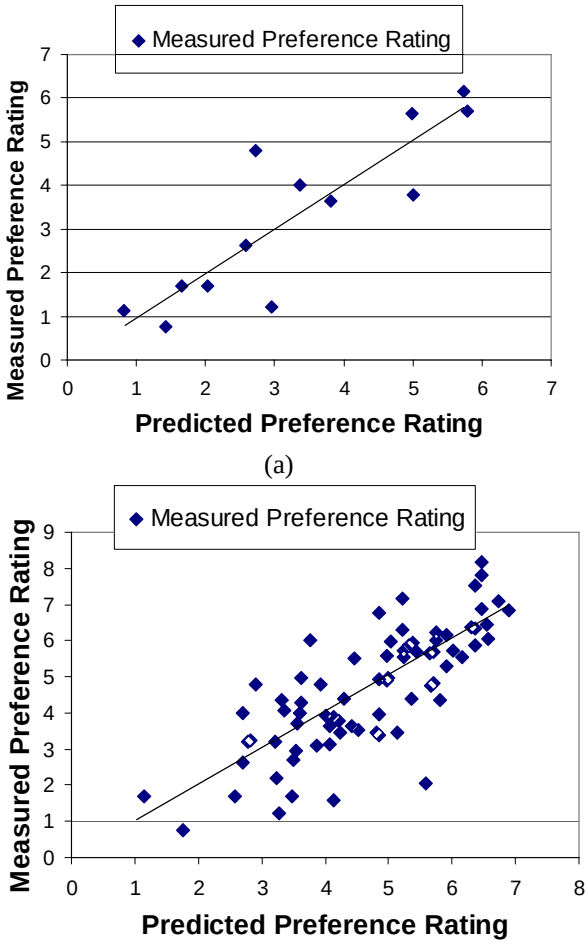

Figure 13 A plot of the measured versus predicted preference ratings based on sound power models defined by equation 12 (top graph, Test One) and equation 13 (generalized model for 70 loudspeakers).

In conclusion, sound power models can give reasonably good predictions of loudspeaker sound quality so long as they use measurement data with adequate frequency resolution, and the models are premised on legitimate performance targets. However, the accuracy of the predictions is still not

as good as those produced with our models that include additional information related to the direct and early-reflected sounds.

8 DISCUSSION

One of the ancillary discoveries in this study is that the quality of the direct sound produced by the loudspeaker is an important contributor to the listeners' overall preference rating, in addition to smooth off-axis frequency response. Both of our models described in equations 9 and 10 give relatively large weighting (45% and 32%) to the quality of the direct sound based on the on-axis curve. If you consider that the predicted in-room response includes a weighted portion of the on-axis sound, the contribution is slightly greater. The important role of the direct sound is well founded in localization and timbre perception theory [25]-[26]. It's responsible for triggering the precedence effect and forward masking. Those models we tested that did not include the direct sound as an independent variable (i.e. the in-room and sound power models) produced less accurate predictions of sound quality. This could explain why they did not perform as well as our model that includes separate measures of both the direct, early-reflected and reverberant sounds.

The relatively good performance of the model (r=0.91) based on in-room loudspeaker measurements (1/20-octave smoothing) is promising. It may provide an alternative approach for assessing loudspeaker performance where anechoic measurements are not available. However, we caution their use until more experiments with additional speakers and rooms are completed. Moving the loudspeakers closer to room boundaries where reflections can cause interference and solid angle gain from the well-known Allison effect [27] may affect the accuracy of the model.

It is clear that the accuracy of all models depends on the frequency resolution of the measurements. Using 1/3-octave smoothing produces less accurate models than those based on 1/20-octave smoothed measurements. Measurements based on 1/3-octave analyzers with fixed center frequency filters will likely produce even worse results. The ITU-R BS1116 recommendation specifies an in-room 1/3-octave loudspeaker response of ± 3 dB between 50 Hz and 2 kHz rising to +3/-6 dB at 16 kHz [16]. Our study clearly shows this cannot adequately distinguish good loudspeakers from mediocre ones. Many of the speakers we measured would meet the ITU performance specification in spite of their very low preference ratings. To borrow a quote from Toole, an ITU compliant loudspeaker guarantees nothing more than that you will be able to tap your foot to the beat and recognize the melody of music played through it.

We have not yet implemented and compared the performance of our model against the Klippel model described in [14]. It is impossible to make meaningful comparisons indirectly due to differences in loudspeaker samples and measurement procedures. Both models are based on similar loudspeaker measurements, except Klippel's model is based on 1/3-octave measurement, which we found produces less accurate models. Klippel's model uses a weighted combination of the free-field on-axis response and the measured or predicted in-room response. This would explain why both models produce good correlations with subjective results. One difference is that Klippel uses a perceptual-based critical band model to calculate loudness errors in a program reproduced through the loudspeaker. Our model may be further improved through the addition of a perceptual model. This question will hopefully be addressed in a future paper.

9 CONCLUSIONS

A new model has been developed that accurately predicts preference ratings of loudspeakers based on their anechoic measured frequency response. Our model produced near-perfect correlation (r = 0.995) with measured preferences based on a sample of 13 loudspeakers reported in Part One. Our generalized model produced a correlation of 0.86 using a sample of 70 loudspeakers evaluated in 19 listening tests. Higher correlations may be possible as we improve the accuracy and resolution of our subjective measurements,which is a current limiting factor.

The independent variables that predict loudspeaker preference rating include measures of the amplitude deviations in the on-axis response, the predicted-in-room response and the low frequency response. Each sound component in the model has equal importance in predicting preference. The predicted in-room response of the loudspeaker captures amplitude deviations in the direct, early-reflected and reverberant sounds and it was found to correlate well

with the measured in-room response of the loudspeaker.

The performance of our model was compared against two sound power models and two models based on in-room measurements. The correlations of the different models from best to worst based on their predicted versus measured correlation were: our anechoic model (1.0), the in-room model based on 1/20-octave smoothed data (0.91), our sound power model (0.87), the in-room model based on 1/3-octave smoothed data (0.75) and the CU model (- 0.22).

Models based on in-room and sound power measurements most likely produce less accurate predictions of sound quality because they are unable include separate quality measures of the direct and early-reflected sounds. These components are shown to be important variables in our model for predicting loudspeaker preference.

Models based on 1/3-octave measurements produce less accurate predictions of sound quality. The CU model is the least accurate model tested because it is based on sound power and uses 1/3-octave data. The model is premised on deviations in amplitude from a flat sound power response, a target we found to be negatively correlated with preference. To achieve flat sound power, traditional speakers with frequency-dependent directivity must sacrifice the quality of the direct and early-reflected sounds that according to our model result in lower preference ratings.

10 LIMITATIONS OF MODEL

The conclusions in this study may only be safely generalized to the conditions in which the tests were performed. Some of the possible limitations are listed below.

1. Up to this point, the model has been tested in one listening room.

2. The model doesn't include variables that account for nonlinear distortion (and to a lesser extent, perceived spatial attributes).

3. The model is limited to the specific types of loudspeakers in our sample of 70.

4. The model's accuracy is limited by the accuracy of the subjective measurements.

The acoustical properties of the listening room used in the development of this model are not unlike those of many professional and domestic listening rooms. It meets current requirements of ITU-R BS 1116 and its reverberation time (RT_{60} = 0.3 s.) falls close to the average value of 0.4 s measured in 603 domestic rooms by Bradley [28]. On this basis, it is more than likely our model can be generalized to many typical rooms.

The effects of nonlinear distortion on preference are not factored into our model. Listeners did not report nonlinear distortion as factoring into their preference ratings, except in the one or two cases reported in Part One. In other large loudspeaker studies conducted by Toole [12]-[13] and Klippel [14] both authors concluded that nearly all of the variance in listener sound quality ratings can be explained by frequency response. Still, nonlinear distortion can be a factor and should not be ignored.

The relationship between loudspeaker measurements and perceived spatial attributes remains elusive. Based on a principal component analysis of listener comments in Part One (see section 4.13), we found that timbre-related attributes accounted for 94% of the variance in comments whereas nonlinear distortion and spatial-related attributes only contributed 3% each. It is our experience that timbre is the dominant factor related to loudspeaker preference, and speakers that accurately reproduce timbre generally have favorable spatial properties. Toole reported high correlation between fidelity ratings (conducted in mono) and spatial ratings (in stereo) and found that most of the spatial effects are strongly related to the recording techniques used in the recording [12]-[13]. Klippel reported that the perceived spaciousness of the loudspeaker is an important dimension that is related to its directivity [14]. In this study, loudspeaker directivity by itself had little predictive power of listener preference. It is unclear what the ideal directivity of the loudspeaker should be, except that it should be smooth. Perhaps the question will become less relevant with multichannel audio where the recording artist has the power to control the balance of the direct, reflected and reverberant sounds within the recording itself.

The loudspeaker sample used in this study may not explain or apply to loudspeakers that have unusual directivities (e.g. omni-directional). Our sample is however representative of the vast majority of

speakers sold today. The range in quality and price ($100-$25,000 per pair) of the sample was quite large excluding only the very least and most expensive models.

Our final limitation is the accuracy of our subjective data. The model from Test One produced a correlation of 1. Test One was very tightly designed with 13 sessions to control contextual effects. Our model's correlation slipped to 0.86 when applied to our listening test data gathered from 18 independent tests. We believe that the lower correlation is, in part, due to differences in how well the listening tests were controlled. The lack of a calibrated and anchored preference scale across the 18 tests most likely resulted in contextual effects and other associated scaling errors in the listener ratings. A solution to this challenging problem is discussed in the following section.

11 FUTURE WORK

Future work will address some of the limitations described in section 10. The model will be tested and verified in different types of listening rooms. A listening room with adjustable acoustics is being designed. Subjective effects related to loudspeaker-room interactions, if necessary, will be included in the model. To facilitate comparison of multiple loudspeakers in different rooms and to control the variable loudspeaker position, a portable index table has been developed.

There is a need to better understand the effects each independent variable has on listener preference by manipulating the variable in isolation while controlling the other independent variables. To do this efficiently, an adjustable loudspeaker has been developed that will allow real-time manipulation of the direct, early-reflected and reverberant sounds. The speaker will allow us to simulate a wide range of different speakers in real-time so that many speakers can be compared in a single listening trial. The ITU MUSHRA method will allow up to 15 speakers to be compared with a hidden reference and anchors [29]. This should help calibrate the preference scale and reduce the contextual effects that occur from comparing a small sample of loudspeakers within a narrow context.

12 ACKNOWLEDGMENT

The author would like to thank Harman International who supported this work. This paper involved literally hundreds of hours of listening, loudspeaker measurements and data analysis that could not have been done without the valuable contribution of many people He is thankful to all of the listeners involved in these tests, and in particular, Sean Barton and John Jackson who conducted all of the subjective and objective measurements that produced the data for these models. The pioneering loudspeaker research of Floyd Toole conducted at the National Research Council of Canada in the 1980's led to the development of the objective measurements used in this model, and he deserves much of the credit. Verification and further refinement of his measurements were provided by Allan Devantier, and they are part of this model. Finally, the author would like to thank his wife, Valerie, and Floyd Toole who proofread this paper.

13 REFERENCES

[1] Olive, Sean, "A Multiple Regression Model For Predicting Loudspeaker Preference Using Objective Measurements: Part 1-Listening Test Results", presented at the 116th AES Convention, Berlin, Germany, preprint 6113, (May 2004).

[2] Consumer's Union, "Loudspeaker Accuracy: CU's Tests," 38, 456-457, (1973August).

[3] "Small Boxes, big sound", Consumer Reports, pp. 33-37, (Aug. 2001).

[4] Rosenberg, U. "Loudspeaker Measurement and Consumer Information," AES 44thConvention (Feb. 1973).

[5] Gabrielsonn, A., Rosenberg U., and Sjogren, H. "Judgments and dimension analyses of perceived sound quality of sound-reproducing systems", J. Acoust. Soc. Am. 55(4), 854-861. (1974)

[6] Staffeldt, H. "Correlation Between subjective and objective data for quality listening tests", J. Audio Eng. Soc., No. 22, pp. 402- 415, (1974 July/Aug.).

[7] Staffeldt, H. and Rasmussen, E. "The Subjectively Perceived Frequency Response in Small and Medium Sized Rooms," Soc. Of Motion Picture and TV Eng., J. SMPTE (1982 July).

[8] Staffeldt, H., "Measurement and Prediction of the Timbre of Sound Reproduction," J. Audio Eng. Soc., Vol. 32, No. 6. pp. 410-414. (June 1984).

[9] Gabrielsonn, A.," Loudspeaker frequency response and perceived sound quality", J. Acoust. Soc. Am., 90(2), Pt. 1, pp. 707-719, (1991).

[10] Gabrielsonn, A, Hagerman, Bjorn, Bech-Kristensen, T., & Lundberg, G., "Perceived sound quality of reproductions with different frequency responses and sound levels", J. Acoust. Soc. Am., 83(3), pp. 1359-1366 (1990).

[11] Gabrielsonn, A. and Lindstrom, B. "Perceived Sound Quality of High-Fidelity Loudspeakers", J. Audio Eng. Soc., 33, pp. 33-53. (1985).

[12] Toole, F.E. "Loudspeaker Measurements and their Relationship to Listener Preferences: Part 1", J. Audio Eng. Soc., 34, pp.227, (1986).

[13] Toole, F.E., "Loudspeaker Measurements and their Relationship to Listener Preferences: Part 2", J. Audio Eng. Soc., 34, pp. 323-348, (1986).

[14] Klippel, W. "Multidimensional Relationship between Subjective Listening Impression and Objective Loudspeaker Parameters," Acustica, Vol. 70, pp. 45-54. (1990).

[15] Klippel, W., "Assessing the Subjectively Perceived Loudspeaker Quality on the Basis of Objective Parameters," presented at the 88th AES Convention, Montreux, preprint 2929 (J5). (1990).

[16] ITU-R BS 1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel audio systems" (1994-1997)

[17] WordIQ, "Linear Regression" see http://www.wordiq.com/definition/Linear_regression

[18] Hair, Anderson, Tatham and Black, "Multivariate Data Analysis", 5th edition, Prentice Hall, New Jersey, (1998).

[19] Mallows C.P. "Choosing a Subset Regression", Joint Statistical Meetings, Los Angeles, CA (1966).

[20] Furnival and Wilson, "Regression by Leaps and Bounds", Technometrics, vol. 16, 4 (November 1974)

[21] Devantier, A., "Characterizing the Amplitude Response of Loudspeaker Systems", presented at the 113th AES Convention, Los Angeles, USA, preprint 5638 (October 2002).

[22] Toole, F. and Olive. S., "Perception and Measurement of Resonances", J. Audio Eng. Soc., 36(3), 122-141, (1989).

[23] Olive, P. Schuck, J. Ryan, S. Sally, M. Bonneville, "The Detection Thresholds of Resonances at Low Frequencies", J. Audio Eng. Soc., 45,116-127. (1997).

[24] Olive S.E., "Differences in the Performance and Preference of Trained versus Untrained Listeners: A Case Study", J. Audio Eng. Soc., vol. 51, No. 9, pp. 806-825. (2003 Sept.).

[25] Moore, B.C.J., *An Introduction to the Psychology of Hearing*, Academic Press, 4th Edition, (1997).

[26] Blauert J., *Spatial Hearing* (1997), MIT.

[27] Allison R.F. and Berkovitz R.,"The Sound Field in Home Listening Rooms," J. Audio Eng. Soc., vol. 20, pp. 459-469 (1972 July-Aug.).

[28] Bradley, J.S.,"Acoustical Measurements in Some Canadian Home", Canadian Acoustics, vol. 14, pp.19-25 (1986 Oct.).

[29] ITU-R BS.1534-1,"Method for the subjective assessment of intermediate quality level of coding systems" (2001-2003).