# Correlating Sound Quality Metrics and Jury Ratings

**David L. Bowen**, Acentech Incorporated, Cambridge, Massachusetts

**An investigation was conducted into how a large group of sound quality metrics might be used to predict user reactions to sounds from a particular type of product, as expressed in terms of product-specific attributes such as ratings of "acceptability" and perceived "quality" of the product. We assume that a jury study has already been conducted for such attributes, producing rating values for various product sounds, and that the objective is to determine which metrics or combinations of metrics can best be used to predict user judgments for the sounds of different versions of the product. The basic methodology employs the use of principal components analysis to group the large number of sound quality metrics into just a few orthogonal (principal) components or factors composed of a weighted sum of the original metrics. A "metrics profile" is then computed for each sound based on the resulting principal components, followed by the creation of a transformation matrix to convert between mean jury ratings and the metrics profile. The expected performance of this transformation in predicting jury ratings from the metrics profile is then assessed. The procedure is illustrated using an example drawn from yard maintenance equipment.**

This article describes an investigation that was conducted into how sound quality (SQ) metrics might be used to predict user reactions to product sounds, where such user reactions are expressed in terms of judgments or ratings on product-specific attributes such as "acceptability" of the sound, or perceived "quality" or overall "effectiveness" of the product itself based on its sound. We assume that at least one jury study has already been conducted on the product class of interest, producing attribute ratings for different versions of the product. The objective then is to determine whether various metrics or weighted combinations of metrics can be used to predict user ratings for the sounds of similar products, avoiding the need to reconvene separate jury studies for each product iteration.

The basic methodology that we have investigated in an attempt to meet this objective involves using principal components analysis (PCA) to group a large number of SQ metrics into just a few orthogonal (principal) components or factors, where such components are composed of a weighted sum of the (standardized) original metrics. A "metrics profile" (MP) is next computed for each sound based on the first few principal components (PCs), followed by the creation of a "transformation matrix" to convert between mean jury ratings and the MPs.

## Principal Components Analysis

PCA and the related method of common-factor analysis (CFA) are often used to determine if a large number of observed variables can be accounted for in terms of a smaller number of inferred "fundamental" factors.[1] In our case, PCA was used to transform a large set of metrics into a smaller set of linear combinations of these metrics based on the values of the metrics calculated over a large set of sounds originating from a particular product class. The resulting combinations are the "principal components" (PCs). This new set of variables accounts for most of the total observed variance, with each combination being orthogonal to the others, meaning that there is no redundant information from one PC to the next. The PCs as a whole form an orthogonal basis for the space of the data, and the first PC is a single axis in this space. When each observation is projected onto this axis, the resulting values form a new variable containing the maximum variance among all possible choices of the first axis. The second PC is another axis in this space, perpendicular to the first, and the variance of this particular variable is the maximum among all possible choices of this second axis, and so on. Usually the first few PCs will account for a large portion of the total variance, and it is this reduction that makes PCA and CFA attractive.

## Example Set of Sound Quality Metrics

A total of 25 different metrics was calculated on a number of product sounds that had been presented to jurors in a previous jury study involving yard maintenance equipment. These metrics are summarized in Table 1. As the characteristics of the product change, we expect some of the metric values to change in a significant way, but not others. Additional metrics could be added to this list, but the ones in Table 1 will be used to illustrate the techniques employed here.

The first 17 metrics in Table 1 are fairly standard ones and were calculated using routines as implemented in a system supplied by LMS. Metrics 18-25 are customized metrics that we have developed and used in the past.[2] These customized metrics relate to "spectral balance" (high vs. low frequency content), tonality, and modulation. A brief description of each of these particular metrics is given below.

Metric 18 (spectral rotation) represents the balance of high frequencies relative to low frequencies, with an A-weighted filter spectrum taken as "neutral." This is done by determining the degree of "pivoting" of the A-weighting curve needed for minimizing the difference between the original A-weighted, 1/3-octave band

*Table 1. Description of metrics calculated on sounds used in consumer jury study.*

| Metric | Name / Description | Units |
|---|---|---|
| 1 | Linear SPL | dB, re 20 µpa |
| 2 | A-weighted SPL | dB, re 20 µpa |
| 3 | B-weighted SPL | dB, re 20 µpa |
| 4 | Zwicker loudness (free field) | Sones |
| 5 | Roughness | Asper |
| 6 | Sharpness (free field) | Acum |
| 7 | Fluctuation strength | Vacils |
| 8 | ANSI speech interference level | dB |
| 9 | Open articulation index | % |
| 10 | Kurtosis | Unitless |
| 11 | Pitch unit | Hz |
| 12 | Pitch value | Pa |
| 13 | Tonality | Ratio |
| 14 | Impulse occurrence rate | Hz |
| 15 | Impulse duration | msec |
| 16 | Impulse peak level | dB, re thres. |
| 17 | Impulse rise rate | dB/msec |
| 18 | "Rotation" of A-weighted 1/3-octave band spectrum about 1000 Hz | dB |
| 19 | Spectral "roughness" )avg. deviation from rotated A-weighted spectrum) | dB |
| 20 | Low-frequency, slow-modulation index | % |
| 21 | Low-frequency, fast-modulation index | % |
| 22 | Mid-frequency, slow-modulation index | % |
| 23 | Mid-frequency, fast-modulation index | % |
| 24 | High-frequency, slow-modulation index | % |
| 25 | High-frequency, fast-modulation index | % |

Based on a paper presented at NOISE-CON 08, Institute of Noise Control Engineering, Dearborn. MI, July 2008.

spectrum and an amplitude-shifted and rotated version of the A-weighting curve itself, while matching the overall A-weighted level. A positive rotation (rotation of the A-weighting curve in the counter-clockwise direction) corresponds to a relative increase in the "treble" end and a reduction in the "bass" end, while a negative rotation corresponds to the reverse. We have used 1000 Hz as the "pivot point" for the rotations, and the frequency analysis range includes the 1/3-octave bands from 400 Hz to 2500 Hz. This metric is given in terms of dB per 1/3-octave band.

Metric 19 (spectral roughness) reflects the deviation of the actual 1/3-octave spectrum from a "smooth" spectrum and is a measure of its spectral irregularity, which is affected primarily by strong tones in the sound. Information for determining this metric arises out of the computations needed for Metric 18 in the form of level differences between the shifted and rotated A-weighting factors for 1/3-octave bands and the original 1/3-octave band spectrum. These differences are then averaged across the frequency bands used in the evaluation to yield an average value for the spectral deviation. A large value can indicate the presence of strong tones that deviate away from the smooth (rotated and shifted) A-weighting curve. This metric is expressed in terms of dB. Note that this metric is different from the traditional time-based roughness described by Metric 5. It also represents another way of estimating the tonality of the sound (other than Metric 13).

The six modulation metrics (Metrics 20-25) are designed to represent different types of modulation that may be present in the measured signals. Modulation of sounds is a characteristic that people readily perceive and can be an undesirable acoustic characteristic of machinery sounds. Slow modulation is perceived as variations in amplitude, while fast modulation can be perceived as a "fluttering" or "buzz-like" characteristic. To distinguish between these different types of modulations, the Hilbert transform is used to compute the amplitude envelopes of the signals, which are then filtered between 0.5 and 8 Hz to detect "slow" modulation, and between 50 and 90 Hz to detect "fast" modulation. In addition, since product sounds can be quite complicated, these fast and slow modulations are evaluated within three different frequency ranges – below 400 Hz, between 400 and 2500 Hz, and above 2500 Hz. A modulation "index" is then formed by taking the ratio of the rms amplitude of the slow or fast envelope signal to the rms amplitude of the original envelope obtained from the filtered sound pressure signal. The modulation metrics are then formed by expressing these indices in terms of percent.

## Generation of PC-Based Metrics Profile

Prior to calculating the principal components "weights," the values of the sound quality metrics computed for each sound are first "standardized" (centered with zero mean and normalized to a standard deviation of 1). This step is needed because the metrics may have completely different units of measure from each other. Figure 1 shows the relative contributions of each of the principal components calculated from the standardized matrix formed from the 25 metrics described in Table 1 and computed on each of 32 different sounds that had been previously presented to a jury of consumers.

These sounds consisted of variations on the sound of a particular product targeted for yard maintenance, most of which were created by altering the sounds of the different sources and mechanisms within the device (the sounds of four "extra" existing products in this class were also included in this set). The information in Figure 1 indicates that the first four PCs explain about 85% of the observed variance in the SQ metric values. These four PCs were retained to represent the metrics and were then rotated (using a varimax rotation) and sorted using a modified form of factor analysis to produce the weighted groupings of metrics shown in Table 2.

The PC weightings provide guidance as to what each of the PCs in the reduced set primarily denotes. For example, referring back to the metric descriptions in Table 1, the weightings for each of the PCs in Table 2 appear to group the metrics into what could roughly be translated as:

- *Loudness* – related metrics such as loudness and the overall SPLs as well as AI and speech interference level.
- *Modulation* – related metrics such as the mid-frequency slow modulation index, fluctuation strength, etc.
- *Tonality* – related metrics such as tonality, pitch, and the "spectral roughness" index.
- *Impulsiveness/peakiness* – related metrics such as impulse peak level, impulse rise rate, Kurtosis, etc.

Using different numbers of PCs will, of course, produce different groupings, and it is often revealing to try out using different

*Table 2. Rotated and sorted PC weighting factors for the first four principal components computed from a set of 25 SQ metrics calculated for 32 sounds.*

| Metric Number | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| 3 | 0.402 | –0.003 | –0.014 | –0.123 |
| 4 | 0.392 | –0.03 | 0.003 | –0.004 |
| 1 | 0.362 | 0.012 | –0.057 | –0.223 |
| 2 | 0.354 | 0.001 | –0.013 | 0.086 |
| 8 | 0.33 | –0.027 | 0.029 | 0.118 |
| 9 | –0.323 | 0.045 | 0.004 | –0.123 |
| 5 | 0.255 | 0.215 | 0.136 | –0.165 |
| 22 | –0.123 | –0.498 | 0.057 | –0.118 |
| 7 | 0.14 | –0.369 | 0.019 | 0.087 |
| 18 | 0.061 | 0.359 | –0.166 | 0.274 |
| 20 | –0.2 | –0.27 | 0.037 | 0.188 |
| 21 | –0.127 | 0.247 | 0.085 | 0.014 |
| 11 | –0.182 | 0.216 | 0.072 | –0.124 |
| 13 | –0.059 | 0.113 | 0.478 | 0.079 |
| 12 | 0.096 | 0.14 | 0.476 | 0.075 |
| 15 | –0.088 | 0.267 | –0.419 | 0.091 |
| 23 | –0.063 | –0.15 | 0.39 | 0.01 |
| 19 | –0.032 | 0.3226 | 0.329 | –0.087 |
| 17 | –0.029 | 0.083 | 0.013 | 0.359 |
| 16 | –0.009 | 0.008 | –0.06 | 0.347 |
| 25 | 0.081 | 0.071 | –0.043 | –0.32 |
| 14 | 0.023 | –0.068 | 0.077 | 0.3 |
| 24 | –0.082 | –0.06 | 0.055 | –0.295 |
| 10 | 0.069 | –0.0-44 | –0.025 | 0.291 |
| 6 | 0.051 | –0.045 | 0.164 | 0.288 |

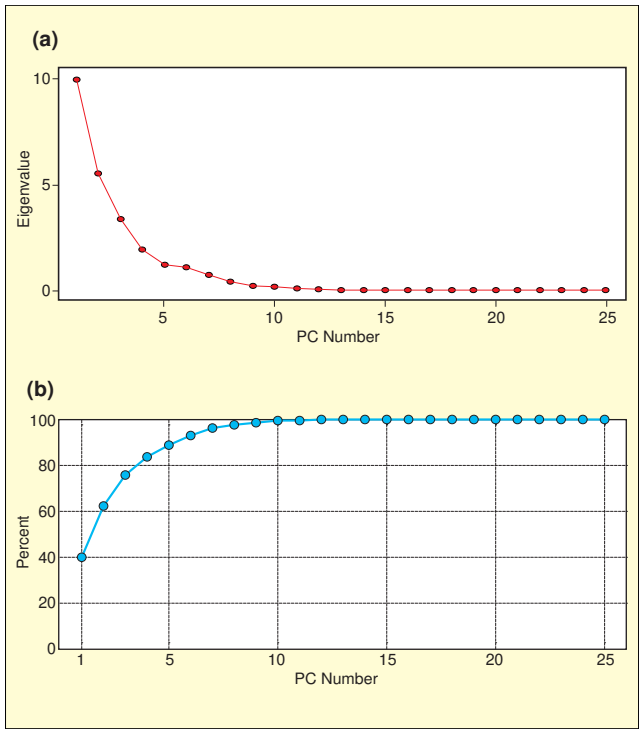


*Figure 1. Judging "importance" of the principal components obtained from a set of 25 sound quality metrics calculated for 32 different sounds: (a) "scree" plot of the eigenvalues, (b) cumulative percentage of observed variance represented by the PCs.*

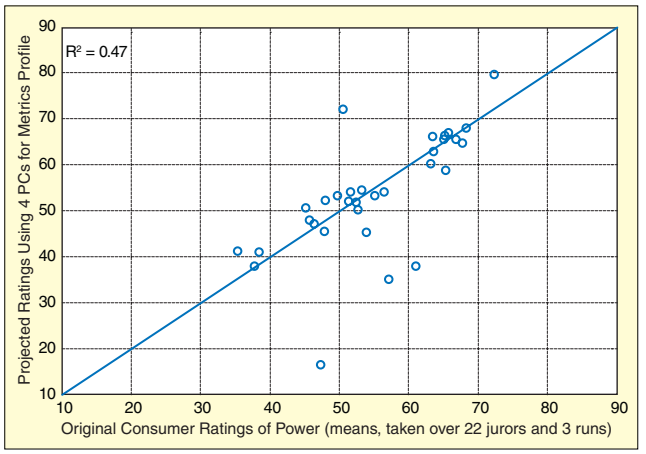


*Figure 2. Jury ratings on the attribute "perceived power" (on a normalized scale of 0-100) for the sounds of 32 different versions of a piece of yard maintenance equipment, vs. ratings predicted by multiplying an SQ "metrics profile" derived from these same sounds by a transformation vector.*

numbers of PCs, while keeping in mind the guiding information such as that provided by Figure 1.

The weightings for the reduced set of PCs are then used to calculate the corresponding "scores" for each of the sounds under consideration using the model for each PC:[1]

$$PCn = w(n)_1 Y_1 + w(n)_2 Y_2 + \ldots + w(n)_p Y_p \qquad (1)$$

where:

$PCn$ = resulting "score" for the $n$th PC
$Y$ = (standardized) metric values for each of the $p$ metrics
$w$ = weights on these variables

For our example, the resulting scores would then consist of four values (from the four PCs) for each of the 32 sounds. We refer to these values as a (PC-based) "metrics profile" (MP) – one MP for each sound. The MPs are then all shifted upward so that all are greater than zero for ease of interpretation.

## Transformation from Metrics Profile to Jury Ratings

The MP scores, along with the mean values of normalized jury ratings previously obtained for these same sounds, were then used to calculate a linear transformation matrix **X** between the MPs and the jury ratings. This was done by solving for **X** in the general system of equations described by:

$$\mathbf{AX} = \mathbf{B} \qquad (2)$$

where:

**A** = $N \times Q$ matrix of metrics profile scores for the sounds ($N$ = number of sounds, $Q$ = number of principal components retained)
**B** = $N \times M$ matrix of mean jury ratings on $M$ attributes for these same sounds
**X** = desired $Q \times M$ transformation matrix

Since $N$ (the number of sounds) will generally be greater than $Q$ (the number of PCs retained), Eq. 2 becomes an over-determined system of equations (32 equations in four unknowns for our example). The transformation matrix (or vector if jury ratings are for a single attribute only) **X** can then be determined in a least-squared error sense using, for example, singular-value decomposition to generate a "pseudo-inverse" of the **A** matrix.[3]

Figure 2 summarizes how well the resulting transformation vector in this example was able to predict the jury ratings for the attribute "perceived power" of the product using the MP values derived from the first four principal components. In this case, the $R^2$ "goodness-of-fit" indicator was about 47%. The four furthest "outliers" in Figure 2 are associated with the four sounds included in the jury study that were not created by altering the sounds of various components in the baseline unit (these "extra" sounds were the sounds of competitor units, alternate models, etc.). If we do not include these four outliers, the resulting $R^2$ value increases to about 88%.

This same transformation could now be evaluated in terms of predicting user reactions to the sounds of other products in this same general class using the same set of weighting factors given by Table 2 to compute the MP scores for these new sounds. Alternatively, PCA could be applied again to form a new set of weighting factors based on the SQ metrics values computed for the new set of sounds. However, this latter approach would not be recommended unless the number of new sounds (the number of "observations" for the PCA) was comparable to or greater than the number of sounds used to form the original PC weighting factors like those in Table 2.

## Conclusions

An approach has been described that attempts to establish a link between a set of objective sound quality metrics and subjective impressions of product sounds. The method makes use of principal components analysis to first reduce a large number of metrics into a weighted combination of smaller groups of metrics. To do this, PCA is applied to a large set of metric values calculated on a large set of sounds, all of which are presumed to originate from a general type of product class (vacuum cleaners or front-loading washing machines or lawn tractors, for example). The first few PCs are then used to develop a set of weighting factors that are applied to the metric values to obtain a (reduced dimension) PC-based "metrics profile."

A transformation matrix between the resulting MP "scores" for these sounds and a set of corresponding jury ratings on particular attributes for these same sounds can then be calculated and evaluated in terms of its ability to predict the jury ratings. A satisfactory transformation can therefore allow physical measurements of sounds from different products or product versions within a product class (made as changes are made to the product) to reasonably predict the effect of these changes on perceived SQ without the need to conduct repeated jury studies.

Applying the technique to a set of 25 metrics calculated on the sounds from 32 different variations of a particular type of yard maintenance equipment resulted in a regression coefficient of 0.47 when used to predict attribute-rating values obtained from a consumer jury that was exposed to these same sounds. The next step would be to assess the accuracy of the ratings predicted if this same transformation were then applied to a new set of sounds obtained from this same product class.

Future directions in this area include investigating the possible use of alternate statistical techniques that are somewhat related to principal components analysis, such as the regression techniques of principal components regression and partial least-squares (PLS) regression. This later technique may offer a more direct and possibly more robust way to generate a reduced-order model for predicting SQ ratings from metric values than the PCA-based metric-profiles approach described here.

PLS can be thought of as a cross between multiple linear regression and PCA, but unlike PCA, PLS directly considers the observed response values (the jury ratings in our case), finding combinations of predictor PCs that have large covariance (a measure of the degree to which two variables change together) with response values.[4] In general, PLS is more of a predictive technique compared to the more interpretive technique of PCA. We hope to report on the results of this and our other on-going work in these areas in the near future.

## References

1. Dillon, W. R., and Goldstein, M., *Multivariate Analysis: Methods and Applications*, John Wiley & Sons, New York, 1984.
2. Bowen, D. L., and Lyon, R. H., "Mapping Perceptual Attributes of Sound to Product Design Choices," *Noise Control Engineering Journal*, Vol. 51, No. 4, pp. 271-279, 2003.
3. Lawson, C., and Hanson, R., *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
4. Rosipal, R., and Kramer, N., "Overview and Recent Advances in Partial Least Squares," in Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS 2005), Revised Selected Papers (Lecture Notes in Computer Science 3940), Springer-Verlag, pp. 34-51, 2006.

The author can be reached at: dbowen@acentech.com.