Can the Perceived Spatial Impression of Classical Recordings Be Predicted by Audio Descriptors?

Paula Klein[1,3*], Markus von Berg[1,3], Dagmar Birwe[2], Jochen Steffens[1,3]
[1] *Institute of Sound and Vibration Engineering (ISAVE), Hochschule Düsseldorf, Germany*
[2] *Institut für Musik und Medien, Robert Schumann Hochschule Düsseldorf, Germany*
[3] *Fachgebiet Audiokommunikation, Technische Universität Berlin, Germany*
[*] *Email: paula.klein@tu-berlin.de*

Abstract
The perceived spatial impression is an important factor in the aesthetic appreciation of classical music recordings. Crucial aspects of spatial perception can be described by the attributes *ensemble width*, *depth*, *envelopment*, and *reverberation duration*. In a listening experiment, a sample of 30 experts rated 24 orchestra recordings by these attributes. To compare these ratings to signal-related measures of spatial impression, an audio descriptor for the interaural cross-correlation coefficient in music recordings was chosen as a measure of perceived *ensemble width*. A second algorithm estimated the reverberation time from decay segments in the recorded music. Linear mixed-effects models were then used to test whether these two descriptors, combined with other audio features related to timbre, are suitable for predicting participants' ratings of the perceptual attributes. Overall, results confirmed that the signal-related measures used can indeed predict perceptual attributes, with 41% of rating variance explained in the case of *reverberation duration*. In contrast, these objective measures turned out to be less accurate predictors of the perceived *ensemble width*, with a variance explanation of 21%. This study can thus be regarded as a starting point for further studies aiming to predict the perceived spatial impression in music recordings.

Introduction
Spatial cues are a crucial aspect in the aesthetic appreciation of recorded music. This is particularly true for the case of classical music recordings, where reproduction of the performance space typically plays a significant role. Measurement methods for describing spatial perceptual qualities of concert halls have been the subject of research in the field of concert hall acoustics for more than 100 years (Sabine, 1906). The question yet arises as to why there are hardly any standardised methods for calculating them from music recordings.[1] The perception of other characteristics related to the musical performance, such as tempo, timbre and dynamics, can already be explained on the basis of audio features (Repp, 1994; Weinzierl & Maempel, 2011). To address this gap, the present study investigated the research question whether and how the perceived spaciousness in classical music recordings can be technically captured. Although spatial perception is characterised by many aspects (e.g., depth, envelopment), this paper will in particular focus on the perception of *ensemble width* and *reverberation duration* due to existing theoretical groundwork.

Among other things, the perception of the extent of a sound source is relevant for describing the spatial qualities of reproduced audio (Rumsey, 2002). To this end, the attribute *ensemble width* was chosen to describe the perceived extent of the ensemble in the horizontal plane. Mason (2002) found a correlation between perceived auditory width of sound sources and signal measures based on interaural cross-correlation coefficient (IACC) calculations. Furthermore, the strength in the low-frequency range has been shown to be particularly relevant for the perception of auditory width (Morimoto & Maekawa, 1988). Therefore, we proposed a model for perceiving *ensemble width* based on IACC calculations and signal measures related to the timbre of a recording.

Furthermore, since the perception of reverberation is a dominant factor in concert hall acoustics, it is reasonable to assume that it could also be applied to music recordings. As the perceived reverberation can be characterised by the *reverberation duration*, we implemented an algorithm for detecting and estimating the duration of reverberation tails in audio signals based on the work of Prego et al. (2012). Thus, the model for perceiving *reverberation duration* will be based on estimations of the length of reverberation tails and signal measures related to timbre.

Methods
Listening experiment
To obtain a valid data base regarding the perceived spatial impression of various classical recordings, we conducted a listening experiment with 30 expert listeners (25 men, 5 women) with an average age of 33 years. Invitations to participate were addressed to students, graduates, and teachers of the Institut für Musik und Medien at the Robert Schumann Hochschule Düsseldorf, as well as to members of the Cologne regional group of the Verband Deutscher Tonmeister. For the experiment, participants reported on the extent to which they have dealt with classical music productions as part of their studies or profession. The listening test took place in a recording studio, that complies with specifications for listening tests given in [ITU-R BS 1116-3] (International Telecommunication Union, 2015). All stimuli were presented in stereo using Klein + Hummel O-500 C loudspeakers.

[1] Another study on capturing room acoustic properties in the audio signal is included in these proceedings (Klouche, 2024).

In total, participants rated three different classical works according to the six attributes of *ensemble width*, *depth*, *envelopment*, *reverberation duration*, *overall spatial impression* and *pleasantness*. The graphical user interface for the presentation and evaluation of the music productions was realized using the experiment-building software PsychoPy (Peirce, 2007). Figure 1 displays the graphical user interface for one classical work and the item *reverberation duration*, where the stimuli (e.g., different productions of the same musical piece) were selected for playback in the numbered fields. The start time of the audio playback could be changed using a slider on the audio timeline. Using the vertically aligned sliders, participants could indicate the perceived magnitude of the attribute on a nine-point Likert scale. The order of presentation of the items as well as the assignment of the audio stimuli to the numbers 1 to 8 were randomized.

Figure 1: Graphical user interface for the listening test.

As part of the experiment, additional demographic data as well as data on the participants' musical expertise was collected using the online survey tool Unipark (Unipark & QuestBack, 2023). Musical expertise was measured with the sub-scales *active engagement*, *perceptual abilities* and *musical training* of the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report inventory by Müllensiefen et al. (2014) in its German translation (Schaal et al., 2014).

Table 1 shows the excerpts from symphonic works used as stimuli in the listening test, which can be assigned to the Classical and Romantic eras. The selected music productions were sampled across several decades (1961 to 2022) and represent a variety of production styles. The excerpts were on average 53 seconds (Beethoven), 40 seconds (Mahler) and 47 seconds (Mozart) long. The respective productions of the musical works were adjusted in their loudness.

Measuring perceived ensemble width

As a measure for perceived *ensemble width*, a calculation of IACCs proposed by Mason (2002) was implemented:

$$IACC = \frac{\sum_{k=0}^{N-1} x(k)y(k+\tau)}{\sqrt{\left(\sum_{k=0}^{N-1} x^2(k)\right) \cdot \left(\sum_{k=0}^{N-1} y^2(k)\right)}}, \qquad (1)$$

Table 1: Music excerpts used in the listening test.

Composition	Excerpt
L. v. Beethoven (1804): Symphony No. 3 in E flat major, Op. 55	1st mvt.: Allegro con brio, bars 1-45
G. Mahler (1894): Symphony No. 2 in c minor	5th mvt.: Im Tempo des Scherzos, bars 454-461
W. A. Mozart (1788): Symphony No. 41 in C major, K. 551	3rd mvt.: Menuetto. Allegretto, bars 1-24

whereby $x(k)$ and $y(k)$ represent the left and right signal. N indicates the block length and τ the offset value, which were set to 50 ms and 1 ms, respectively. The calculation is based on short signal excerpts of about 1 s duration, which represent the overall sound of the recorded ensemble. IACC values were calculated in one-third octave bands, averaged over time, and weighted according to the loudness distribution across the frequency range. This allows perceptually more prominent frequency bands to be taken into account. The resulting coefficients were then inverted to ensure a positive correlation with ratings from the listening experiment. For the statistical analysis, an average value of coefficients in the frequency range between 125 and 1000 Hz was used.

Measuring perceived reverberation duration

To establish an objective measure for the perceived *reverberation duration*, an adaption of the algorithm for blind reverberation time estimation by Prego et al. (2012) was implemented. First, the audio signal was divided into overlapping frames using a Hamming window function, which were then transformed into the frequency domain through the discrete Fourier transform (DFT). Subsequent frames were examined for a continuous decrease in signal energy in order to detect so-called free decay regions (FDR). FDRs are perceptually relevant if they do not fall below a certain minimum duration. Hence, FDRs decreasing in energy over a period of at least 500 ms were specified.

At this point, it should be emphasized that the algorithm by Prego et al. (2012) is optimized for speech analysis. The orchestral recordings to be analyzed in this work, however, exhibit greater irregularities in their decay processes since they are the result of the summation of a large number of sound sources. Therefore, the criterion for identifying FDRs was modified by introducing a parameter that sets a minimum percentage of consecutive frames that must exhibit a decrease in signal energy. This way, non-monotonously decaying reverberation as depicted in Figure 2 could also be recognized.

For detected FDRs, energy values were calculated that were 10, 20, and 40 dB below the -5dB peak value, using

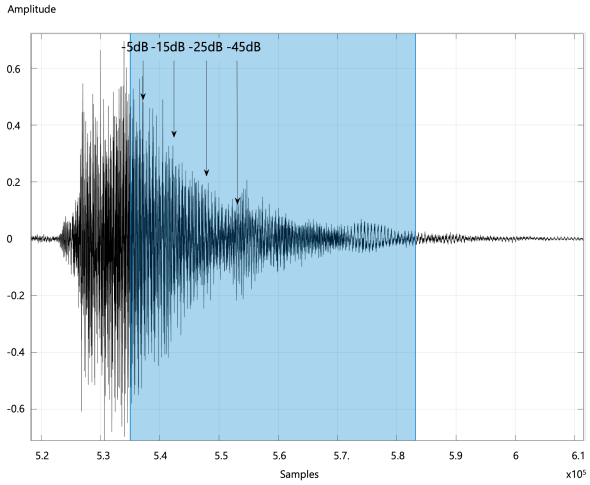

Figure 2: Exemplary representation of an FDR (area marked in blue), in which a level drop of 5, 15, 25 and 45 dB from the peak value is marked.

the frame-based subband energy decay function, which is

$$c(k,l) = 10 \cdot \log_{10} \left(\frac{\sum_{\lambda=n}^{L-1} E(k,\lambda)}{\sum_{\lambda=0}^{L-1} E(k,\lambda)} \right) \text{dB}. \qquad (2)$$

Here, $E(k,\lambda)$ refers to the signal energy curve, with the variables k representing the subband and λ constituting the frame number within a FDR of the total length L. $c(k,l_1) = -5$dB represents the "left" and $c(k,l_2) = -15/-25/-45$dB the "right" limit of an FDR. From the values l_1 and l_2 fulfilling this condition, time values of the FDR can be calculated similar to T_{10}, T_{20}, and T_{40}. For the orchestral recordings used as stimuli in this work, T_{10} extrapolations proved to be the most practicable. Frequency averaging was carried out over the frequency range 1 to 2 kHz.

Feature Extraction

In addition to the aforementioned parameters, further timbre-related audio descriptors, such as the spectral centroid (sc), spectral rolloff (sr), spectral flux (sf) and the Mel Frequency Cepstral Coefficients 0 to 3 (MFCC) were extracted from the audio signal using the MATLAB Audio Toolbox (The MathWorks, 2023). These features were aggregated to single-values using typical measures of descriptive statistics, such as mean, median, standard deviation, minimum, maximum, and interquartile range.

Statistical analysis

Linear mixed-effects (LME) models were utilised to investigate whether the perceptual ratings collected in the listening experiment can be predicted by the proposed signal calculations. LME models allow for the specification of both fixed and random effects, whereby, in the present study, random intercepts were estimated for both participants and musical works. In the first step, LME models were computed with the T_{10}-estimations and the calculated IACC values as the only fixed effects (baseline models). Subsequently, the extracted timbre features were added using a method based on the optimisation of the conditional Akaike information criterion. To assess model performance, marginal R^2 values were used as a measure for the variance explained by fixed effects.

The statistical analysis was performed using R (R Core Team, 2022), and the lme4 package (Bates et al., 2015) was used to calculate the LME models.

Results and Discussion

Listening experiment

Table 2 shows the GOLD-MSI scores of the participants in this study compared to those of a large, representative English-speaking sample from the main study by Müllensiefen et al., 2014, indicating a high level of musical expertise of our sample.

Table 2: Mean values and standard deviations (figures in brackets) in three Gold MSI dimensions of the samples from Müllensiefen et al. (2014) and this study.

Sample	musical training	active engagement	perceptual abilities
Müllensiefen et. al (2014)	0.46 (0.11)	0.60 (0.03)	0.76 (0.02)
This study	0.72 (0.09)	0.79 (0.15)	0.90 (0.08)

Statistical models

First, ICC values were calculated to assess the inter-rater reliability in the listening experiment suggesting a high agreement of judgements across raters with values of .876 (*ensemble width*) and .929 (*reverberation duration*). In a next step, Pearson correlation coefficients between the averaged ratings and the proposed signal calculations were computed, showing a small, but significant correlation r = .119 (p<.001) in the case of the applied IACC calculations and *ensemble width* ratings and a significant moderate correlation r = .380 (p<.001) between the proposed T_{10} estimations and *reverberation duration* ratings.

Table 3: Fixed effects estimates and R^2 values of the *ensemble width* models with different numbers of fixed effects, $^*p < 0.05$.

Fixed effects	baseline model	extended model
IACC	.161*	.458*
MFCC3_Iqr	-	-.392*
Sr_Sd	-	.629*
Sc_Max	-	-.549*
Sf_Sd	-	.352*
R_c^2	.048	.309
R_m^2	.025	.205

Table 3 shows the results of the LME modelling for perceived *ensemble width*, where R_m^2 indicates the variance explained by fixed effects, and R_c^2 represents the variance explained by both fixed and random effects. It becomes obvious that the baseline model including only the IACC calculations as fixed effect achieves a poor variance explanation with $R_m^2 = .025$. However, by gradually adding further timbre-related predictors, the R_m^2 increases to 21% (cf. Table 3).

Table 4: Fixed effects estimates and R^2 values of the *reverberation duration* models with different numbers of fixed effects, $^*p < 0.05$.

Fixed effects	baseline model	extended model
T_{10}	.376*	.431*
MFCC1_Max	-	.507*
MFCC0_Max	-	.368*
R^2_c	.182	.621
R^2_m	.142	.414

Table 4 shows the results of the LME modelling for perceived *reverberation duration*. Here, the baseline model only including the T_{10} estimation values as fixed effect explains 14% of the variance. By adding timbre features as further predictors, the R^2_m increases to 41% which constitutes a decent model fit.

Comparing the R^2_m value of all calculated models, it is evident that the proposed IACC calculations make a minor contribution to the variance explanation of the perceived *ensemble width* compared to FDR calculations for *reverberation duration*. It can thus be concluded that FDR calculations in music recordings are a promising approach to predict perceived *reverberation duration*. However, it must be noted that the algorithmic detection of FDRs only works with audio signals that contain a adequate number of free-standing decay processes.

Conclusion

This paper presents an approach to the explanation of spatial perception in classical music recordings showing significant relationships between expert ratings in a listening experiment and distinct signal calculations based on room-acoustical theory. However, since these relationships were shown to be moderate at best, a combination of theory-based and data-driven approaches (as performed in the course of computing and selecting additional timbre-related audio features) might be a feasible approach to increase predictive power. Since *ensemble width* and *reverberation duration* only reflect partial aspects of spatial perception, future research should furthermore aim on identifying appropriate (phyiscal) audio descriptors to predict other spatial perception parameters, such as depth or envelopment, or the overall spatial impression.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

International Telecommunication Union. (2015). Methods for the subjective assessment of small impairments in audio systems.

Klouche, T. (2024). Empirisch-Systematische Messung von virtualisierten Audiosignalen: Einfluss von Musik und Raum. *This volume*.

Mason, R. (2002). Elicitation and measurement of auditory spatial attributes in reproduced sound [Doctoral Thesis].

Morimoto, M., & Maekawa, Z. (1988). Effects of Low Frequency Components on Auditory Spaciousness. *Acta Acustica united with Acustica*, *66*(4), 190–196.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Prego, T. d. M., de Lima, A. A., Netto, S. L., Lee, B., Said, A., Schafer, R. W., & Kalker, T. (2012). A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. *The Journal of the Acoustical Society of America*, *131*(4), 2811–2816. https://doi.org/10.1121/1.3688503

R Core Team. (2022). R: A language and environment for statistical computing.

Repp, B. H. (1994). On Determining the Basic Tempo of an Expressive Music Performance. *Psychology of Music*, *22*(2), 157–167. https://doi.org/10.1177/0305735694222005

Rumsey, F. (2002). Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *Journal of the Audio Engineering Society*, *50*(9). https://www-aes-org/e-lib/browse.cfm?elib=11067

Sabine, W. C. (1906). The accuracy of musical taste in regard to architectural acoustics. *Proceedings of the American Academy of Arts and Sciences*, *42*(2), 53–58.

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe. *Musicae Scientiae*, *18*(4), 371–504. https://doi.org/10.1177/1029864914541851

The MathWorks, I. (2023). Audio Toolbox Release 2023a.

Unipark & QuestBack. (2023). Unipark. https://ww2.unipark.de/

Weinzierl, S., & Maempel, H.-J. (2011). Zur Erklärbarkeit der Qualitäten musikalischer Interpretationen durch akustische Signalmaße. In von Loesch, Heinz, Weinzierl, Stefan (Ed.), *Gemessene Interpretation*. Schott.