Forum Acusticum 2011 - Aalborg - Denmark

Obtaining objective, content-specific room acoustical parameters using auditory modeling

Jasper van Dorp Schuitman
Philips Research, Digital Signal Processing Group, Eindhoven, The Netherlands.

Diemer de Vries
Laboratory of Acoustical Imaging and Sound Control, Delft University of Technology, Delft, The Netherlands.

Summary
The acoustical properties of rooms are generally described using objective parameters as determined from measured - or simulated - room impulse responses. However, this method has some disadvantages. For example, room impulse responses are generally measured in empty rooms, while the acoustical properties will usually depend on the occupancy of the room. Furthermore, the source type (speech, music, etc.) is hardly taken into account. Finally, in some cases the objective parameters do not correlate with the human perception of the acoustical properties of a particular room. In previous papers, we proposed a new method for assessing objective parameters related to four aspects of room acoustics: reverberance, clarity, apparent source width and listener envelopment. The method uses a nonlinear, binaural auditory model, which is able to extract these parameters from arbitrary binaural recordings. This means, that these parameters can be obtained during a performance, for example. Another advantage is, that the source type is automatically taken into account. So far, the new method has proven its predictive values; from various listening tests it was found that in most cases the new objective parameters correlate better with the perceptual results than the conventional parameters. In this paper, an overview of the method will be presented, including some practical aspects. For example, the minimum required signal-to-noise ratio for getting accurate results will be discussed, as well as the minimum signal length.

PACS no. 43.66.Ba, 43.55.Mc

1. Introduction

When acousticians assess the acoustical qualities of a room, they typically start with measuring the room impulse response for one or multiple source / receiver combinations. From these impulse responses, objective parameters are determined which are believed to be related to perceptual attributes like reverberance, clarity and apparent source width. Some of these objective parameters are well-established by now and described in the ISO standard 3382-1 [1].

However, there are some issues with this method. First of all, the measurement of the impulse response(s) usually takes place in empty rooms because of the equipment needed and the type of measurement signals (sweeps, noise, etc.). The occupancy of a room can have a significant effect on the acoustics, and corrections on the parameter values are needed to predict the acoustical qualities for fully occupied rooms. These corrections, however, are only effective for concert halls with well-upholstered seats [2].

Furthermore, the impulse response-based method does not take into account the fact that the perception of room acoustics can be highly content-specific. The temporal and spectral properties of a signal can result in masking effects, for example, which may affect the amount of reverberation which is perceived [3].

Finally, besides masking effects, the human auditory system has more features which can have an influence on the perception of acoustics. For example, the auditory system is non-linear. The masking effects described above are dependent on the absolute sound pressure level of the signal [4] and therefore the perceptual attributes will also depend on the SPL.

As a result of the mentioned shortcomings, the resulting objective parameters do not always correlate very well with perception, see for example [5, 6, 7]. In this paper, a new method is proposed where new objective room acoustical parameters can be obtained, directly from arbitrary binaural recordings, using an auditory model.

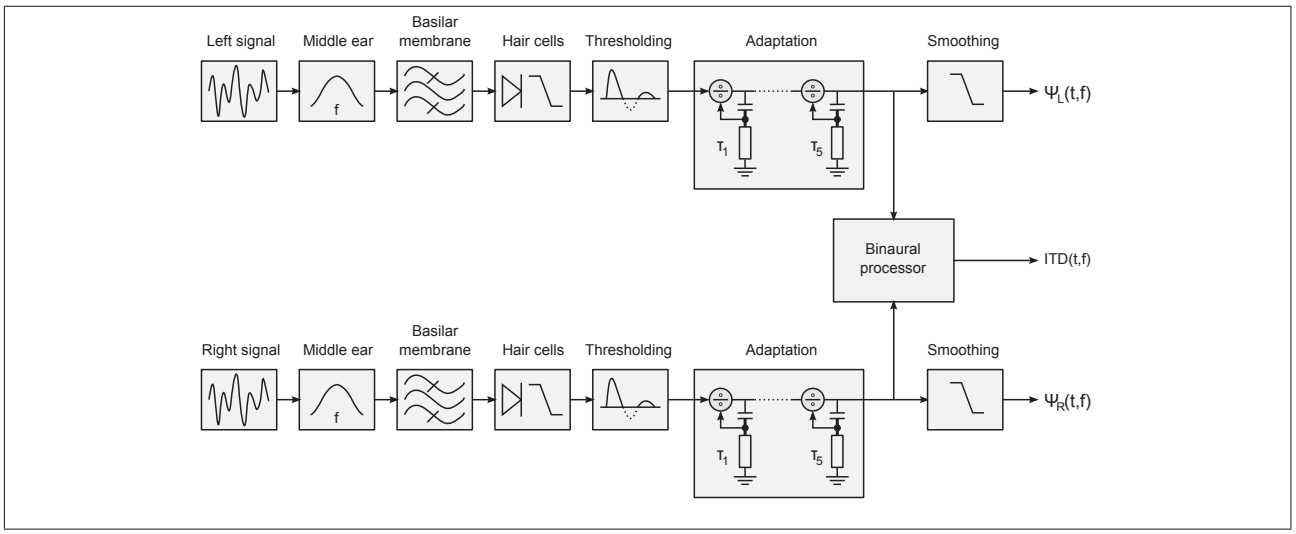

Figure 1. A schematic version of the binaural, nonlinear auditory model.

2. The auditory model

The method which is proposed in this paper makes use of a nonlinear, binaural auditory model. The model is shown schematically in Fig. 1. The auditory model is based on the binaural model as proposed by Breebaart [8]. The various stages of the model will be briefly discussed below.

1. The transfer function of the **ear canal** is modelled using a band-pass filter.
2. The **basilar membrane** inside the cochlea is modelled by a filterbank consisting of Gammatone filters.
3. Furthermore, the behaviour of the **hair cells** is simulated by applying half-wave rectification.
4. **Neural adaptation** is simulated using a chain of five adaptation loops with specific time constants. This part models the effect that the neurons need some time to 'relax' after sudden on- and offsets. For stationary input signals, the adaptation stage approximately acts as a logarithmic compressor.
5. To incorporate an **absolute threshold of hearing** (ATH), frequency-dependent thresholding is applied. Note that this method differs from the one as proposed by Breebaart, where a Gaussian noise signal with a frequency-independent amplitude was added to the signal [8].
6. The binaural processor calculates the Interaural Time Difference (ITD) as a function of time using the monaural outputs for each frequency band.
7. Finally, the monaural outputs are filtered with a low-pass filter with a cutoff frequency of 20 Hz. The resulting signals Ψ basically reflect the signal's envelopes [4]. The Ψ outputs, as well as the ITD as a function of time and frequency, are processed by the **central processor** to extract the objective acoustical parameters which will be discussed in the next section.

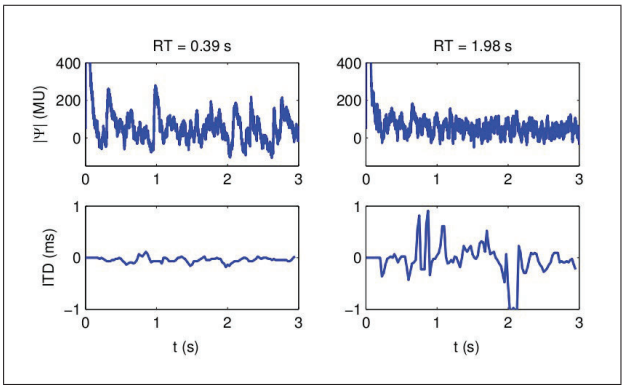

Figure 2. Example output of the binaural model. A speech signal was used which was convolved with simulated binaural room impulse responses (BRIRs). The BRIRs were simulated for two virtual halls with reverberation times RT = 0.39 s (left) and RT = 1.98 s (right). The top figures show monaural outputs in Model Units (MU), the bottom figures the interaural time difference ITD. All plots are for the 256 Hz band.

Figure 2 shows example outputs of the binaural model. A dry, male speech signal was convolved with two different simulated binaural room impulse responses; one for a room with a short reverberation time (0.39 s, left) and one for a room with a longer reverberation time (1.98 s, right). Both input signals were normalized to the same RMS level. Figure 2 shows the monaural outputs Ψ for the left channel as well as the ITD values as a function of time.

Some important observations can be made from the outputs. First, the more reverberant signal results in an overall lower *output* from the model, even though the RMS levels of the *input* signals were identical. Furthermore, in the less reverberant case the signal components (phonemes) of the speech signal are more distinct compared with the more reverberant case. They show up as clear peaks in the model output.

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767

Finally, ITD values fluctuate more as a function of time in the more reverberant case, due to reflections arriving from lateral directions. This will lead to a broadening of the source and a subjective impression of envelopment, as found by Blauert and Lindemann [9, 10]. In the next section this effect will be used to derive objective parameters related to these attributes. Authors who recently published research on this subject include Mason [11] and Rumsey *et al.* [12].

3. Objective parameters

As discussed in the previous section, the model outputs reflect the acoustical environment and this can be used to obtain objective parameters related to the room acoustics. It was chosen to focus on four attributes which are believed to be important to the perception of acoustics (see [2]): reverberance, clarity, apparent source width (ASW) and listener envelopment (LEV). In the next sections, each attribute and its corresponding objective parameter will be discussed.

3.1. Reverberance

Reverberance is related to the amount of reverberation which is perceived in a room. Its corresponding objective parameter is currently the reverberation time (the time for the sound to decay by 60 dB after the sound source stops), which is defined in ISO 3382-1 [1]. It is calculated from the decay of the impulse response. Another parameter related to reverberance is the early decay time EDT, which is calculated from the first 10 dB of decay in the impulse response. The EDT has been said to be a better predictor for perceptual reverberance than the reverberation time [13].

In order to determine the amount of reverberation from the model outputs, an algorithm was developed which separates the monaural model outputs into two streams: one belonging to the source ('direct sound') and one for the reverberant field. The splitting procedure is based on a peak detection algorithm, where parts of the monaural streams which have a level above a threshold for a minimum amount of time are labeled as belonging to the source. An example result of this algorithm is shown in Fig. 3. For more information about the threshold and minimum peak width, the reader is referred to [14].

Next, the avarage level of the part of the output which is assigned to the reverberant stream is used as an objective parameter (P_{REV}) for reverberance:

$$P_{\mathrm{REV}} = L_{\mathrm{rev}} = \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{m=m_0}^{m_1} \left| \Psi_{\mathrm{rev}}[n,m] \right|, \quad (1)$$

where N is the number of time frames, M is the number of frequency bands over which L_{rev} is calculated, n is the sample index and m the frequency band index. Ψ_{rev} is defined as follows:

$$\Psi_{\mathrm{rev}}[n,m] = \sqrt{\Psi_{L,\mathrm{rev}}[n,m]^2 + \Psi_{R,\mathrm{rev}}[n,m]^2}, \quad (2)$$

with $\Psi_{L,\mathrm{rev}}$ and $\Psi_{R,\mathrm{rev}}$ the detected reverberant streams for the left and right channels.

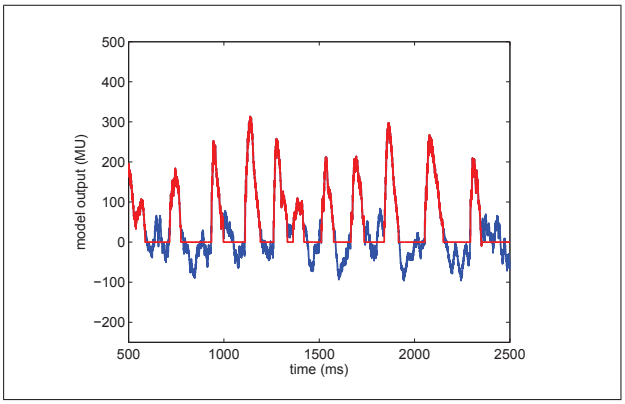

Figure 3. Separation of the model output into streams belonging to the source (red) and reverberant sound (blue).

3.2. Clarity

Clarity is the degree to which discrete sounds in a signal stand apart in time from one another subjectively. For music, clarity is usually estimated using the clarity index C_{80} which is the ratio between early (< 80 ms) and late (> 80 ms) energy in the impulse response [15]. For speech signals usually C_{50} is used, where the 80 ms time limit is changed to 50 ms ([15]).

Here, clarity will be estimated as the ratio between the average level of the direct sound stream over that of the reverberant stream ($P_{\mathrm{CLA}} = L_{\mathrm{dir}}/L_{\mathrm{rev}}$). L_{dir} is defined in a way similar to Eq. 1.

$$L_{\mathrm{dir}} = \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{m=m_0}^{m_1} \left| \Psi_{\mathrm{dir}}[n,m] \right|, \quad (3)$$

with:

$$\Psi_{\mathrm{dir}}[n,m] = \sqrt{\Psi_{L,\mathrm{dir}}[n,m]^2 + \Psi_{R,\mathrm{dir}}[n,m]^2}. \quad (4)$$

3.3. Apparent source width

In a room apparent broadening of a sound source can occur as a result of early lateral reflections, resulting in a certain apparent source width (ASW). ASW is most often assessed using the early interaural cross-correlation ($1 - \mathrm{IACC}_{E3}$), or the early lateral fraction LF, which are calculated from binaural impulse reponses as measured using an artificial head [1].

Basically, lateral reflections interfering with the direct sound cause the ITD to fluctuate over time, as discussed in the section describing the binaural model.

Therefore the amount of ITD fluctuation is related to ASW, as was also proposed by Blauert and Lindemann [9], Griesinger [16], Mason [11] and Becker [17]. Furthermore, Okano showed that the perceived source width is not only related to the interaural decorrelation above 500 Hz, but also depends on the absolute sound pressure level at frequencies below 500 Hz [18]. Therefore, in this research the output of the binaural processor and the level in the lower bands are used to estimate ASW using the model:

$$P_{\mathrm{ASW}} = \alpha_1 L_{\mathrm{low}} + \log_{10}\left(1 + \beta_1 \sigma_{\tau,\mathrm{dir}}\right), \qquad (5)$$

where L_{low} is the average monaural output level for low frequencies, and $\sigma_{\tau,\mathrm{dir}}$ is the average standard deviation of ITD for the direct sound stream. α_1 and β_1 are constants which will be discussed later on.

3.4. Listener envelopment

Listener envelopment (LEV) is the second important subjective parameter related to spaciousness. It is related to the environment instead of the source. A sound field is called enveloping when a perception of being surrounded by the sound occurs, since the sound seems to originate from all directions.

Currently, the objective parameter for envelopment is LEV which can be determined from impulse responses (note that in literature both the subjective and objective parameters for listener envelopment are referred to as LEV). This parameter was proposed by [19] and has recently been turned into a more practical form by Beranek [20]:

$$\begin{aligned}\mathrm{LEV} &= 0.5 G_{\mathrm{late,mid}} \\ &+ 10 \log\left(1 - \mathrm{IACC}_{\mathrm{late,mid}}\right),\end{aligned} \qquad (6)$$

where $G_{late,mid}$ is the late (> 80 ms) sound strength, averaged over mid frequencies (500 and 1000 Hz octave bands). $\mathrm{IACC}_{\mathrm{late,mid}}$ is the late (> 80 ms) interaural cross-correlation averaged over those frequency bands.

Following this line of reasoning LEV consists of two elements: the absolute late sound pressure level and a spacious aspect (interaural cross-correlation). This concept is used to obtain an objective parameter predicting LEV from the auditory model outputs:

$$P_{\mathrm{LEV}} = \alpha_2 L_{\mathrm{rev}} + \log_{10}\left(1 + \beta_2 \sigma_{\tau,\mathrm{rev}}\right), \qquad (7)$$

where L_{rev} is the mean level of the reverberant stream and $\sigma_{\tau,\mathrm{rev}}$ is the mean standard deviation for the ITD values in that stream.

3.5. Optimization and validation of the method

In order to validate the new method, double-blind listening tests were conducted. In these tests, subjects had to rate the various room acoustical attributes (reverberance, clarity, ASW and LEV) for different acoustical environments while listening to the samples through headphones. In two tests (I and II), virtual environments were used which were simulated using a software package which was developed especially for this purpose. This simulator is able to simulate binaural room impulse responses for shoebox-shaped rooms. In test I, "realistic" virtual rooms were used. In test II, more "unrealistic" room designs were used, for example with highly absorbing side walls and fully reflective front- and rear walls, making the various attributes more independent from each other.

In the other two tests (III and IV), a variaty of real rooms was used, of which binaural room impulse responses were measured to construct the test samples. In all four listening tests, two different stimuli were used to test the influence of the type of source signal on the attributes: a male speech signal and solo cello music.

The number of subjects participating in the tests was 5 (tests I and II) and 15 (tests III and IV); the number of different acoustical environments was 9 (test I), 8 (test II) or 10 (tests III and IV).

In three of the four tests, the samples were normalized to the same loudness level using the Replaygain algorithm (www.replaygain.org). This algorithm estimates the perceived loudness of a sample by evaluating the RMS level in windows of 50 ms length, where frequency weighting is applied according to an approximation of the equal loudness curve. The 95% highest value of all windows is considered to be the perceived loudness value for the complete sample. In test III, the original loudness differences were maintained in order to see if this had an influence on the subjects' ratings (however, an in-depth discussion of this is out of the scope of this paper).

The results of one of the four listening tests were used to optimize the model ("training data"), while the results for the other three tests were used as "test data". It was chosen to optimize all the free parameters in the model, like frequency ranges and the α and β constants in Eqs. 5 and 7, using a Genetic Algorithm (GA). Genetic algorithms form a class of algorithms which are capable of optimizing nonlinear, multi-modal problems with a lot of free parameters, like the auditory model presented in this paper. These algorithms search the solution space by simulating evolution (survival of the fittest). A complete explanation of the algorithm is out of the scope of this paper; for more information the reader is referred to [21].

The model - including the set of optimized parameters - was validated by evaluating the correlation co-

Table I. The correlation coefficients between the listening test results and the objective parameters for the attribute **reverberance**. Two conventional objective parameters were evaluated: T_{20} and EDT. P_{REV} is the objective parameter related to reverberance as resulting from the new method.

Test (stimulus)	T_{20}	EDT	P_{REV}
I (cello)	0.84	0.85	**0.98**
I (speech)	0.79	0.80	**0.96**
II (cello)	**0.96**	0.83	0.87
II (speech)	-0.30$^{(*)}$	-0.55$^{(*)}$	**0.88**
III (cello)	0.86	**0.88**	0.76
III (speech)	0.74	0.76	**0.81**
IV (cello)	0.79	0.78	**0.95**
IV (speech)	0.73	0.73	**0.96**

Table II. The correlation coefficients between the listening test results and the objective parameters for the attribute **clarity**. Two conventional objective parameters were evaluated: C_{50} and C_{80}. P_{CLA} is the objective parameter related to clarity as resulting from the new method.

Test (stimulus)	C_{50}	C_{80}	P_{CLA}
I (cello)	0.79	0.77	**0.94**
I (speech)	0.87	0.86	**0.90**
II (cello)	0.82	0.88	**0.83**
II (speech)	0.03$^{(*)}$	0.04$^{(*)}$	**0.82**
III (cello)	0.79	**0.82**	0.79
III (speech)	0.91	**0.94**	0.82
IV (cello)	0.91	0.93	**0.96**
IV (speech)	0.94	**0.96**	0.87

Table III. The correlation coefficients between the listening test results and the objective parameters for the attribute **ASW**. Two conventional objective parameters were evaluated: $1 - \mathrm{IACC_E}$ and LF. P_{ASW} is the objective parameter related to ASW as resulting from the new method.

Test (stimulus)	$1 - \mathrm{IACC_E}$	LF	P_{ASW}
I (cello)	0.71	0.33$^{(*)}$	**0.83**
I (speech)	0.74	0.42$^{(*)}$	**0.90**
II (cello)	0.40$^{(*)}$	0.09$^{(*)}$	**0.64$^{(*)}$**
II (speech)	0.66$^{(*)}$	0.21$^{(*)}$	**0.83**
III (cello)	0.86	N/A	**0.92**
III (speech)	0.86	N/A	**0.94**
IV (cello)	0.67	N/A	**0.86**
IV (speech)	**0.82**	N/A	0.75

efficients between the listening test results (i.e., the average rating for the acoustical attributes) and the objective parameters as determined using the model. The results are shown in Tables I, II, III and IV. For each row (test/stimulus combination), the highest correlation coefficient is shown in bold. Values marked with $^{(*)}$ are not significant at the $p < 0.05$ level.

4. Practical aspects

Various practical aspects of the new method were examined. Two of these aspects will be discussed in this paper. First, it is important to know the minimum signal length which is necessary for obtaining accurate results with the method. By truncating the inputs signals at different time lengths, it was found that after approximately 10 seconds, the objective parameters become more or less stable; meaning that it is not really necessary to use longer signals. However, even though this can be used as a rule of thumb, longer measurement times might be needed if signals contain lots of variation. Or, different parameters could be specified for the different parts of an orchestral piece, for example.

A second practical aspect which was examined was the accuracy of the objective parameters as a function of signal-to-noise ratio (SNR). This was tested by simulating 100 rooms with random properties using the simulator for shoebox-shaped rooms. For each room, the objective parameters were determined using a male speech signal as stimulus. It was found that, if the parameters are obtained *directly* (i.e., by recording the sound source in a room directly), then an SNR of at least 30 dB is necessary for obtaining parameters with an absolute error smaller than 0.1. The parameters can also be determined using an indirect method, by first measuring the room impulse responses using a logarithmic sweep signal with a length of 10 seconds. Next, these responses were convolved with the anechoic stimulus and these convolved versions were used as input signals for the method. Using the indirect method, the absolute error in the parameters was smaller than 0.01 for SNRs of 30 dB and higher. So, the indirect method is much more robust to noise compared with the direct method.

Table IV. The correlation coefficients between the listening test results and the objective parameters for the attribute **LEV**. Two conventional objective parameters were evaluated: $1 - \mathrm{IACC_L}$ and $\mathrm{LEV_{calc}}$. P_{LEV} is the objective parameter related to LEV as resulting from the new method.

Test (stimulus)	$1 - \mathrm{IACC_L}$	$\mathrm{LEV_{calc}}$	P_{LEV}
I (cello)	0.78	0.61$^{(*)}$	**0.85**
I (speech)	0.78	0.61$^{(*)}$	**0.93**
II (cello)	0.44$^{(*)}$	**0.81**	0.69$^{(*)}$
II (speech)	0.54$^{(*)}$	0.17$^{(*)}$	**0.70$^{(*)}$**
III (cello)	0.67	0.64	**0.90**
III (speech)	0.76	0.75	**0.96**
IV (cello)	0.84	0.82	**0.94**
IV (speech)	0.86	0.84	**0.96**

5. Conclusions and discussion

In this paper, a method was proposed for obtaining new objective parameters which are related to four room acoustical attributes: reverberance, clarity, ASW and LEV. In this method, arbitrary binaural recordings are processed by a non-linear, binaural auditory model. Using a splitting procedure, the monaural model outputs are separated into two streams: one

belonging to the source (direct sound) and one to the environment (reverberant sound). Together with the ITD (fluctuations) as a function of time and frequency, the new parameters are calculated from these streams.

The model was validated using the results of four listening tests, in which subjects had to rate the acoustical attributes while listening through various samples through headphones. The correlation coefficients between the average test results and the objective parameters from the model we evaluated. From these correlation coefficients, it can be seen that the new method performs quite well. For almost all test / stimulus combinations, the parameters from the model yield higher correlation coefficients than those for the conventional parameters. In fifteen cases, the conventional parameters show an insignificant result (at the $p < 0.05$ level), whereas the new parameters yield insignificant results in only three cases. In all of those three cases, the conventional parameters also failed to give significant results.

Another advantage of the model is that it accepts arbitrary binaural input, meaning that there is no need to measure sweeps or other artificial signals in empty rooms. Instead, the acoustical features of a room can be assessed in a concert situation, for example. Furthermore, the model takes the spectral and temporal features of the stimulus automatically into account.

Besides the practical aspects mentioned in this paper (minimum signal length and SNR), more aspects were examined. These include the influence of signal type (other than the speech/cello signals used in this paper), the influence of dummy head position/orientation, and more. For more information, the reader is referred to [14].

Acknowledgement

This project has been funded by the Dutch technology foundation STW.

References

[1] ISO. ISO 3382-1:2009: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces. International Organization for Standardization, 2009.

[2] M. Barron. Using the standard on objective measures for concert auditoria, ISO 3382, to give reliable results. *Acoustical Science and Technology*, 26(2):162–169, 2005.

[3] D. Griesinger. The psychoacoustics of apparent source width, spaciousness & envelopment in performance spaces. *Acta Acustica united with Acustica*, 83(4):721–731, July 1997.

[4] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622, June 1996.

[5] M. Barron. Subjective study of British symphony concert halls. *Acustica*, 66(1):1–14, June 1988.

[6] A. Farina. Acoustic quality of theatres: correlations between experimental measures and subjective evaluations. *Applied Acoustics*, 62(8):889–916, August 2001.

[7] T. Lokki, H. Vertanen, A. Kuusinen, J. Pätynen, and S. Tervo. Auditorium acoustics assessment with sensory evaluation methods. In *Proceedings of the International Symposium on Room Acoustics (ISRA 2010), Melbourne*, 2010.

[8] D. J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.*, 110(2):1074–1088, August 2001.

[9] J. Blauert and W. Lindemann. Auditory spaciousness: Some further psychoacoustic analyses. *J. Acoust. Soc. Am.*, 80(2):533–542, August 1986.

[10] J. Blauert and W. Lindemann. Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *J. Acoust. Soc. Am.*, 79(3):806–813, March 1986.

[11] R. Mason. *Elicitation and measurement of auditory spatial attributes in reproduced sound.* PhD thesis, University of Surrey, 2002.

[12] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares. QESTRAL (Part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. In *Proceedings of the 125th AES Convention, San Fransisco*, October 2008.

[13] G. A. Soulodre and J. S. Bradley. Subjective evaluation of new room acoustic measures. *J. Acoust. Soc. Am.*, 98(1):294–301, July 1995.

[14] J. Van Dorp Schuitman. *Auditory Modeling for Assessing Room Acoustics (preliminary title).* PhD thesis, Delft University of Technology, 2011 (to be published).

[15] O. Abdel Alim. *Abhängigkeit der Zeit- und Registerdurchsichtigkeit von raumakustischen Parametern bei Musikdarbietungen (Dependence of time and register definition of room acoustical parameters with musical performances).* PhD thesis, TU Dresden, 1973.

[16] D. Griesinger. Objective measures of spaciousness and envelopment. In *Proceedings of the 16th International AES Conference*, 1999.

[17] J. Becker. Spectral and temporal contribution of different signals to ASW analysed with binaural hearing models. In *Proceedings of the Forum Acusticum 2002, Sevilla*, 2002.

[18] T. Okano, L. L. Beranek, and T. Hidaka. Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LF_E), and apparent source width (ASW) in concert halls. *J. Acoust. Soc. Am.*, 104(1):255–265, July 1998.

[19] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross. Objective measures of listener envelopment in multichannel surround systems. *J. Audio Eng. Soc.*, 51(9):826–840, September 2003.

[20] L. L. Beranek. Concert hall acoustics - 2008. *J. Audio Eng. Soc.*, 56(7/8):532–544, July 2008.

[21] J. Holland. *Adaptation in natural and artificial systems.* The University of Michigan Press, 1975.