Speech Quality Assessment for Listening-Room Compensation

STEFAN GOETZE,[1] *AES Member*, **EUGEN ALBERTIN**[1], **JAN RENNIES,**[1] *AES Member*,
(s.goetze@idmt.fraunhofer.de) (eugenalbertin@web.de) (jan.rennies@idmt.fraunhofer.de)

EMANUËL A.P. HABETS,[2] *AES Member*, **AND KARL-DIRK KAMMEYER,** *AES Member*
(emanuel.habets@audiolabs-erlangen.de) (kammeyer@ant.uni-bremen.de)

[1] *Fraunhofer Institute for Digital Media Technology, IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany*
[2] *International Audio Laboratories Erlangen, University of Erlangen-Nuremberg, Erlangen, Germany*
[3] *University of Bremen, Dept. of Communications Engineering, Bremen, Germany*

In this contribution objective measures for quality assessment of speech signals are evaluated for listening-room compensation algorithms. Dereverberation of speech signals by means of equalization of the room impulse response and reverberation suppression has been an active research topic within the last years. However, no commonly accepted objective quality measures exist for assessment of the enhancement achieved by those algorithms. This paper discusses several objective quality measures and their applicability for dereverberation of speech signals focusing on algorithms for listening-room compensation.

0 INTRODUCTION

State-of-the-art hands-free communication devices as they are used, e.g., in offices or car environments, use algorithms to reduce ambient noise, acoustic echoes, and reverberation. Reverberation is caused by numerous reflections of the signal on room boundaries (walls, floor, and ceiling) in enclosed spaces. Reverberant speech sounds distant and echoic [1]. Large amounts of reverberation decrease speech intelligibility and perceived quality at the position of the near-end speaker of a communication system [2–4]. In general, two distinct reverberation reduction classes exist, viz. reverberation suppression and reverberation cancellation. Reverberation suppression approaches focus on removing the reverberant part of the speech signal by calculating a spectral weighting rule for each time-frequency coefficient in a way similar to well-known approaches for noise reduction (cf., e.g., [5] and the references therein). Reverberation cancellation approaches remove the influence of the acoustic channel between the sound source and the listener by equalizing the room impulse response (RIR) of the acoustic channel. Furthermore, the equalizer can be applied to the loudspeaker signal or the microphone signal. Listening-room compensation is achieved in the former case, i.e., when the equalizer is applied to the signal that is emitted by the loudspeaker such that the influence of reverberation on the perceived signal is reduced at the position the listener is assumed to be located. In order to compute the equalizer one requires knowledge of the RIR. This knowledge can be obtained either by means of blind [6] or non-blind [7–9] channel identification methods. Non-blind methods identify the acoustic channel based on reference information, e.g., the loudspeaker signal in a hands-free system. Such methods are commonly used for acoustic echo cancellation (AEC) [5, 7] where the loudspeaker signal as received by the microphone is estimated by identifying the acoustic channel between the loudspeaker and the microphone and subtracting the estimated signal from the microphone signal. If such a reference signal is not available, e.g., if the source signal is unobservable, the acoustic channel has to be estimated blindly, i.e., without a reference. While the aim of listening-room compensation (LRC) algorithms is to improve the sound quality of the dereverberated signal, they may also decrease the sound quality if they are not designed properly [7, 10]. Thus, especially during algorithm design periods a reliable objective quality measure is required to evaluate and compare different algorithms and their parameters.

Many signal processing strategies change a signal, e.g., to enhance speech quality, speech intelligibility or to reduce listening difficulty [11] (i.e., the effort related to extracting speech information from a distorted signal; in some cases listening effort can differ markedly between signals although they do not differ with respect to speech intelligibility, see, e.g., [11–13]). For all such signal modifications, the general question arises how to assess the achieved enhancement. Since subjective listening tests [14–17] that involve humans are not applicable in every case because they are time-consuming and costly, objective quality measures that assess the performance of the dereverberation

algorithm based on impulse responses, transfer functions or signals are needed [20]. While several commonly accepted quality measures exist to assess the performance of audio codecs [14, 16, 17, 19, 20] noise reduction algorithms [20, 21] or acoustic echo cancelers [22, 23], the assessment of dereverberation algorithms is still an open issue [1, 10, 24].

This work discusses several measures that can be used for evaluating dereverberation algorithms. An evaluation of the sound quality of the dereverberated signals is conducted by subjective listening tests and compared to the results of the objective measures. As previously shown by the authors [10], most signal-based measures have difficulties to assess the performance of dereverberation algorithms properly, especially if distortions are introduced that are small in amplitude but clearly perceivable by the human listener. However, these measures are of particular interest since, e.g., for non-linear dereverberation suppression approaches, channel-based measures may not be applicable since the impulse response of such an algorithm may be neither linear nor time-invariant. Thus, artifacts that may be introduced by the dereverberation algorithms such as late echoes or spectral distortions and their effect on the quality measures are analyzed and discussed. The algorithms are analyzed regarding their capability to assess the properties *reverberation*, *coloration*, *spectral distortion*, perceived *distance*, and *overall quality* of the signals.

The remainder of this paper is organized as follows. Methods for LRC that were used for generating the test signals are briefly summarized in Sec. 1 and some general remarks on quality assessment for LRC algorithms are given in Sec. 2. Section 3 gives an overview of objective quality measures that principally can be used for quality assessment of LRC algorithms and Sec. 4 describes the experimental setup for the subjective listening tests. Results of the correlation analysis are presented in Sec. 5 and Sec. 6 concludes the paper.

Notation: The following notation is used throughout the paper. Vectors and matrices are printed in boldface while scalars are printed in italic. The discrete time and frequency indices are denoted by n and k, respectively. The superscripts $\times^T$ and $\times^+$ denote the transposition and the Moore-Penrose pseudo inverse, respectively. The operator $\mathrm{E}\{\cdot\}$ is the expectation operator, the operator $\mathrm{convmtx}\{\mathbf{h},L_{\mathrm{EQ}}\}$ generates a convolution matrix of size $(L_{\mathrm{EQ}}+L_h-1)\times L_{\mathrm{EQ}}$ and the operator $\mathrm{diag}\{\cdot\}$ yields a matrix of size $L\times L$ from a vector of size $L\times 1$ that has the vector's elements on its main diagonal and zeros elsewhere.

1 LISTENING-ROOM COMPENSATION

A general setup for listening-room compensation is shown in Fig. 1. For LRC the equalization filter

$$\mathbf{c}_{\mathrm{EQ}} = [\, c_{\mathrm{EQ},0},\, c_{\mathrm{EQ},1},\, \ldots,\, c_{\mathrm{EQ},L_{\mathrm{EQ}}-1}\,]^T \tag{1}$$

of length L_{EQ} precedes the acoustic channel characterized by the RIR

$$\mathbf{h} = [\, h_0,\, h_1,\, \ldots,\, h_{L_h-1}\,]^T \tag{2}$$

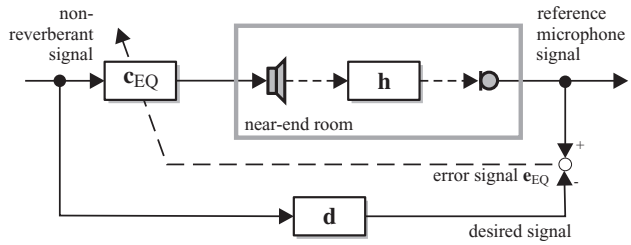

Fig. 1. General setup for listening-room compensation (LRC) using an equalizer filter $\mathbf{c}_{\mathrm{EQ}}$.

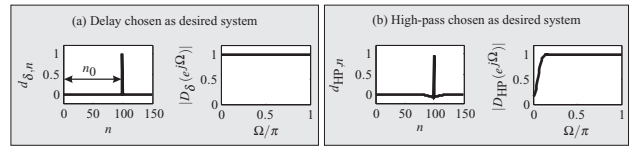

Fig. 2. Two possible desired systems (a) $\mathbf{d}_\delta$ (delay) and (b) $\mathbf{d}_{\mathrm{HP}}$ (delayed high-pass) in time and frequency domain.

of length L_h. The aim of the equalizer is to remove the influence of the RIR at the position of the reference microphone [8, 27] and, by this, to remove reverberation from the signal.

Four different LRC approaches were used to generate sound samples with the goal of covering a large amount of distortions that may occur while using such algorithms. These four approaches are briefly introduced in the following, i.e., (i) the least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}}$, (ii) the weighted least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{WLS}}$, (iii) an impulse-response shaping approach with post-processing $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwPP}}$ according to [25], and (iv) an impulse response shaping with infinity-norm optimization $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwINO}}$ according to [26]. In the following, the least-squares LRC filter and the weighted least-squares LRC filter are briefly derived and the impulse response shaping approaches are briefly introduced. For a deeper discussion of the LRC algorithms we refer the reader to [1, 8, 9, 26–28].

Since an RIR is a mixed-phase system having thousands of zeros close to or even outside the unit-circle in z-domain, a direct inversion by a causal stable filter is not possible in general [28]. Therefore, least-squares approaches focus on minimizing the error vector

$$\mathbf{e}_{\mathrm{EQ}}^{\mathrm{LS}} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}} - \mathbf{d}, \tag{3}$$

where $\mathbf{H} = \mathrm{convmtx}\{\mathbf{h}, L_{\mathrm{EQ}}\}$ is the channel convolution matrix built up by the RIR coefficients and

$$\mathbf{d} = [\underbrace{0,\ldots,0}_{n_0},\, d_0,\, d_1,\ldots,\, d_{L_d-1},\, \underbrace{0,\ldots,0}_{L_h+L_{\mathrm{EQ}}-1-L_d-n_0}\,]^T \tag{4}$$

is the desired response of length $L_h + L_{\mathrm{EQ}} - 1$ that usually is chosen as a delayed delta impulse, a delayed high pass or a delayed band pass as exemplarily depicted in Fig. 2 for a delayed impulse $\mathbf{d}_\delta$ (left panel) and a delayed high pass $\mathbf{d}_{\mathrm{HP}}$ (right panel).

The delay introduced by the equalizer is denoted by n_0 (cf., [29] for a discussion of n_0).

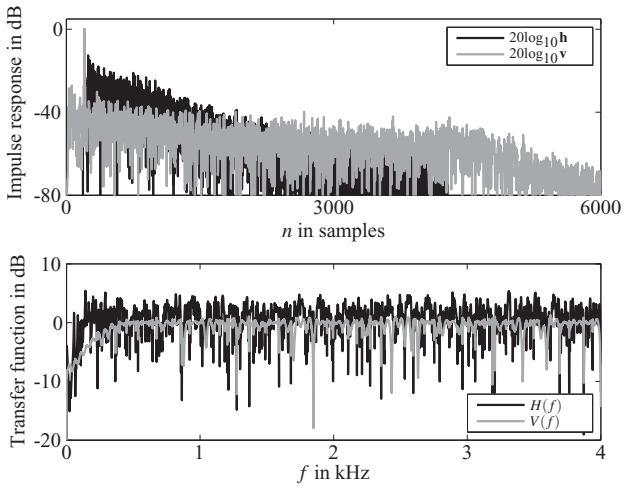

Fig. 3. RIR $\mathbf{h}$ and equalized IR $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}}$ in time-domain in dB (upper panel) and the corresponding squared-magnitude spectra in dB (lower panel).

In theory, perfect equalization is achieved when $\mathbf{d} = \mathbf{d}_\delta$ because the error vector $\mathbf{e}_{\mathrm{EQ}}^{\mathrm{LS}}$ in Eq. (3) becomes $\mathbf{0}$ if the concatenated system of LRC filter $\mathbf{c}_{\mathrm{EQ}}$ and RIR $\mathbf{h}$ equals $\mathbf{d}_\delta$, thus no reflections cause reverberation in time domain and the spectrum is absolutely flat. To account for the frequency responses of imperfect transfer characteristics of loudspeakers and microphones usually a delayed high pass as in Fig. 2 or a delayed band pass is chosen.

Minimizing the norm of the error vector $\mathbf{e}_{\mathrm{EQ}}^{\mathrm{LS}}$ given by Eq. (3) leads to the well-known least-squares equalizer

$$\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}} = \mathbf{H}^{+}\mathbf{d}. \tag{5}$$

An RIR $\mathbf{h}$ and the respective impulse response $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}}$ after application of the least-squares LRC filter $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}}$ are exemplarily shown in Fig. 3 in time-domain (upper panel) and frequency-domain (lower panel).

The room reverberation time of the RIR $\mathbf{h}$ is $\tau_{60} = 0.5$ s and the respective filter length of the equalizer is $L_{\mathrm{EQ}} = 4096$ at a sampling rate of $f_s = 8$ kHz. Given that limited number of LRC filter coefficients the LS-EQ approach seems to show good results in the time-domain (reflections are 30 to 40 dB suppressed compared to the main peak) as well as in the frequency-domain (approximation of the desired high pass is clearly visible). However, the resulting equalized system looks slightly different from a usual room impulse response, i.e., it does not decay linearly in logarithmic time domain. The human auditory system is used to this linear decay [30], thus although the desired system $\mathbf{d}$ that was chosen as a delayed high-pass is closely approximated a large amount of late reverberation exceeding the original decay can be observed, e.g., around sample $n = 4000$. Although small in amplitude this late reverberation is clearly perceivable and disturbing since it is no longer masked by the natural decay of common RIRs [26, 30]. Furthermore, pre-echoes that occur before the main peak of the equalized channel's impulse response $\mathbf{v}$ further disturb a natural sound perception.

The previously described problem of the least-squares LRC filter can partly be avoided by the so-called weighted least-squares equalizer that will be derived in the following. Rather than minimizing the norm of the error vector $\mathbf{e}_{\mathrm{EQ}}^{\mathrm{LS}}$, one can minimize the norm of a weighted error vector

$$\mathbf{e}_{\mathrm{EQ}}^{\mathrm{WLS}} = \mathbf{W}\mathbf{e}_{\mathrm{EQ}}^{\mathrm{LS}} \tag{6}$$

with

$$\mathbf{W} = \mathrm{diag}\,\{\mathbf{w}\} \tag{7}$$

$$\mathbf{w} = [\underbrace{1, 1, \ldots, 1}_{N_1}, \underbrace{w_0, w_1, \ldots, w_{N_2-1}}_{N_2}]^T \tag{8}$$

$$w_i = 10^{\frac{3\alpha}{\log_{10}(N_0/N_1)}\log_{10}(i/N_1)+0.5}. \tag{9}$$

Here, $\mathbf{W}$ is a diagonal matrix containing a window weighting vector $\mathbf{w}$ on its main diagonal. By a proper choice of the weighting vector $\mathbf{w}$, RIR shortening or RIR shaping can be achieved. Preferably, the weighting is based on the psychoacoustic property of masking observed in the human auditory system in order to alleviate perceptually disturbing late echoes [26, 30]. In Eqs. (8) and (9), the constants N_0, N_1, and N_2 are given as follows: $N_0 = (t_0 + 0.2)f_s$, $N_1 = (t_0 + 0.004)f_s$ and $N_2 = L_h + L_{\mathrm{EQ}} - 1 - N_1$. The time of the direct sound is denoted by t_0. The given window function emphasizes suppression of later parts of the RIR to avoid the previously described problem of late echoes. $\alpha \leq 1$ is a factor that influences the steepness of the window. For $\alpha = 1$ the window corresponds to the masking found in human subjects [26, 30].

Minimizing the ℓ_2-norm of the weighted error vector $\|\mathbf{e}_{\mathrm{EQ}}^{\mathrm{WLS}}\|_2^2$ leads to a weighted least-squares equalizer

$$\mathbf{c}_{\mathrm{EQ}}^{\mathrm{WLS}} = (\mathbf{W}\mathbf{H})^{+}\mathbf{W}\mathbf{d}. \tag{10}$$

Please note that, for $\mathbf{w} = \mathbf{w}^{\mathrm{LS}} = [1, 1, \ldots, 1]^T$, the weighted least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{WLS}}$ reduces to the conventional least-squares equalizer as defined in Eq. (5).

Fig. 4 shows the performance of the weighted least-squares equalizer for the same parameters and the same RIR as in Fig. 3. By applying the window as defined in Eq. (9) disturbing late echoes are reduced. The weighted least-squares LRC filter *squeezes* the RIR to result in a quicker decay of the equalized IR $\mathbf{v}$ than the original RIR $\mathbf{h}$ in time-domain (upper panel). The problem of clearly perceivable late echoes above the original decay of the RIR can be reduced. However, the performance in frequency-domain is decreased as it can be seen comparing lower panels of Figs. 3 and 4.

Please note, that all time-domain impulse responses have been time-aligned and normalized to have their main peak at the same position and at same level.

Another approach for RIR shaping was discussed in [25] and is based on the solution of a generalized eigenvalue problem

$$\mathbf{A}\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwPP}} = \lambda_{\max}\mathbf{B}\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwPP}}, \tag{11}$$

$$\mathbf{A} = \mathbf{H}^T\mathbf{W}_{\mathrm{u}}^T\mathbf{W}_{\mathrm{u}}\mathbf{H}, \tag{12}$$

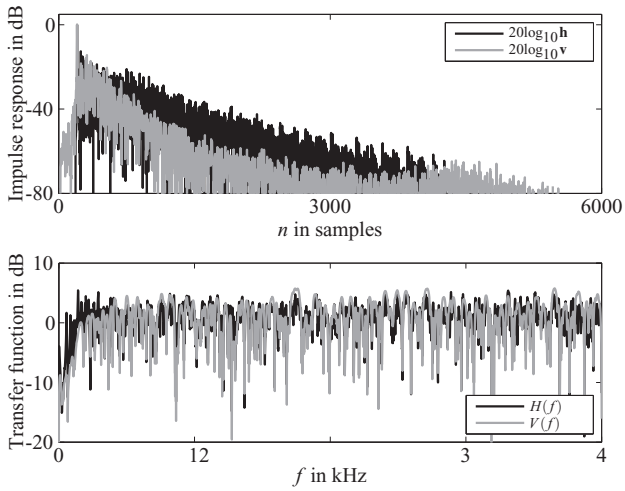

Fig. 4. RIR $\mathbf{h}$ and equalized IR $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{WLS}}$ in time-domain in dB (upper panel) and the corresponding squared-magnitude spectra in dB (lower panel).

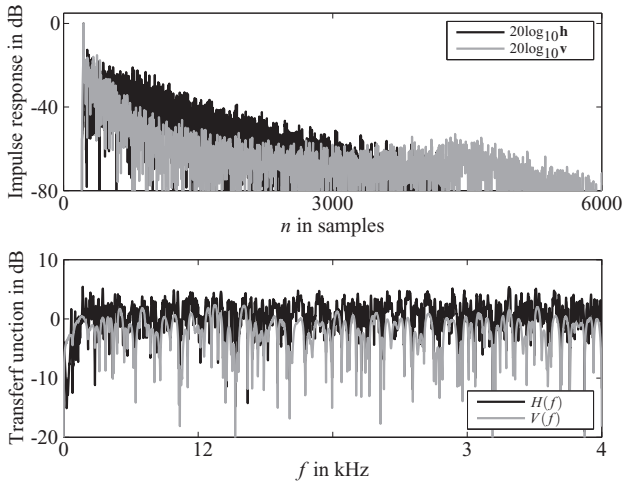

Fig. 5. RIR $\mathbf{h}$ and equalized IR $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwPP}}$ in time-domain in dB (upper panel) and the corresponding squared-magnitude spectra in dB (lower panel).

$$\mathbf{B} = \mathbf{H}^T \mathbf{W}_{\mathrm{d}}^T \mathbf{W}_{\mathrm{d}} \mathbf{H}. \tag{13}$$

Similar to Eq. (10), $\mathbf{W}_{\mathrm{u}}$ and $\mathbf{W}_{\mathrm{d}}$ are diagonal matrices with window functions defining a desired part of the RIR and an undesired part of the RIR, respectively. The greatest eigenvalue is denoted by $\lambda_{\max}$ in Eq. (11). To avoid spectral distortion a post-processor based on linear prediction [25] is used after applying Eq. (11). For a more detailed discussion the reader is referred to [25, 26]. An equalized system $\mathbf{v}$ after application of an LRC filter designed according to Eq. (11) is shown in Fig. 5 again for the same parameters and the same RIR $\mathbf{h}$. Results are similar to those depicted in Fig. 4.

An approach that jointly shapes the impulse response (IR) of the equalized acoustic channel and minimizes spectral distortions is described in [26]. Additionally, the psychoacoustic property of masking is explicitly exploited in the filter design approach described in [26]. Furthermore, this approach is based on a gradient update strategy that avoids computationally complex matrix operations that are needed for the other approaches, e.g., for the inverse of the matrix $\mathbf{H}$ in Eq. (5), the inverse of $\mathbf{W}\,\mathbf{H}$ in Eq. (10), both of size $(L_{\mathrm{EQ}} + L_h - 1) \times L_{\mathrm{EQ}}$, or the solution of the generalized eigenvalue problem in Eq. (11).

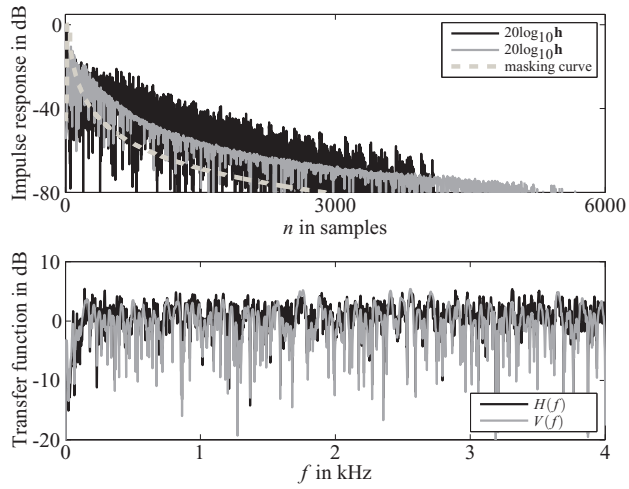

Fig. 6. RIR $\mathbf{h}$ and equalized IR $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwINO}}$ in time-domain in dB (upper panel) and the corresponding squared-magnitude spectra in dB (lower panel).

Table 1. Different LRC approaches and the respective acronyms.

Acronym	Description of method
1. LS-EQ	Least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{LS}}$ according to Eq. (5) without weighting of error signal ($\mathbf{w} = \mathbf{1}$)
2. WLS-EQ	Least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{WLS}}$ according to Eq. (10) with window function according to Eq. (9)
3. ISwPP	Impulse response shaping (IS) according to Eq. (11) with post-processing (PP) $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwPP}}$ [25]
4. ISwINO	Impulse response shaping (IS) with infinity-norm optimization (INO) $\mathbf{c}_{\mathrm{EQ}}^{\mathrm{ISwINO}}$ according to [26]

As visible in Fig. 6 the equalized system $\mathbf{v}$ directly follows the masking curve found in the human auditory system (although due to the limited LRC filter order not reaching it) and a smooth decay can be observed for the whole length of the equalized system $\mathbf{v}$.

Table 1 summarizes the four approaches and the respective acronyms used for LRC and for generating dereverberated signals that were used for the subjective tests described in Sec. 4.

2 QUALITY ASSESSMENT FOR LRC ALGORITHMS

Within this contribution, quality assessment involving human *subjects* is called *subjective* quality assessment while quality assessment based on technical measures is denoted by the term *objective*. If humans are asked for their opinion about the quality of a specific sound sample they

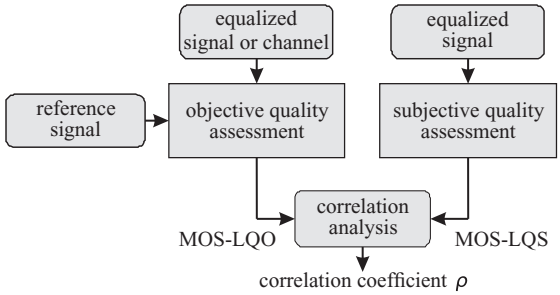

Fig. 7. Quality assessment by means of subjective and objective testing.

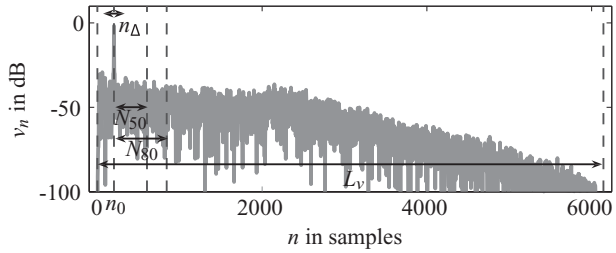

Fig. 8. Impulse response of an equalized acoustic channel $\mathbf{v} = \mathbf{H}\,\mathbf{c}_{\mathrm{EQ}}$ in dB and the corresponding definitions of the position of the main peak n_0, and the discrete samples following 50 ms and 80 ms after this main peak N_{50} and N_{80}. Sampling frequency is $f_s = 8$ kHz.

are able to assess the quality based on an internal reference. This reference is created throughout their lives while listening to various sounds and allows the subject to determine the perceived quality of a sound sample. However, if subjects are asked to assess the quality of a certain sound sample on a categorical scale as for the listening tests conducted for this study, the variance between different subjects may be quite high since each subject may have a different internal reference, i.e., perception of *good, medium* or *bad sound quality*. Variance in the results of listening tests can be decreased by choosing expert listeners instead of naive listeners. The intended target group for hands-free communication systems will be predominantly non-expert listeners. Therefore, we chose mostly non-expert listeners while some subjects had experience with subjective quality assessments.

Unfortunately, subjective quality assessment is time consuming and costly. Thus, especially during algorithm design and test periods reliable objective quality measures are needed that show high correlation with subjective ratings. Since no commonly accepted measure for LRC quality assessment has been identified yet, we analyzed the correlation between subjective quality ratings and various objective measures that are assumed to be applicable for LRC quality assessment as depicted in Fig. 7. Here, the reverberant signal is processed by the LRC algorithm under test that produces a processed signal and a corresponding equalized impulse response. This signal is assessed by human subjects. The objective measures described in Sec. 3 either take the equalized impulse response (channel-based measures) or the processed signal (signal-based measures) as an input. A mean opinion score (MOS) for the subjective listening quality (MOS-LQS) can be calculated as well as for the listening quality obtained by objective measures (MOS-LQO). The correlation between the subjective and objective ratings can be determined by the Pearson product-moment correlation coefficient (PPMCC)

$$\rho = \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2 \sum_i (b_i - \bar{b})^2}}, \tag{14}$$

with a_i and b_i being the subjective and objective ratings of a specific sound sample, respectively, and $\bar{a}$ and $\bar{b}$ the respective mean values.

3 OBJECTIVE QUALITY ASSESSMENT

This section focuses on the description of several objective quality measures that are assumed to be capable of assessing quality of signals processed by LRC algorithms. Two classes of objective quality measures for LRC can be defined: (i) measures that are based on the impulse response or the transfer function of a system (channel-based measures) and (ii) measures that are based on signals only. For LRC algorithms, both the filter impulse response $\mathbf{c}_{\mathrm{EQ}}$ and the RIR $\mathbf{h}$ are available during simulations. However, if gradient algorithms [27] are used to avoid computational complex matrix inversions, e.g., as in Eq. (10), or to track time-varying environments or if the effect of the dereverberation algorithm cannot be characterized in terms of a linear time invariant (LTI) impulse response, e.g., as in [5, 31, 32], the necessary impulse responses of the room or the filter may not be accessible or it may be inappropriate to apply those measures [33]. Such situations restrict the number of applicable measures to those based on signals as described in Sec. 3.2.

It should be noted that besides the *Speech-to-Reverberation Modulation Energy Ratio* measure all objective measures used in this contribution belong to the class of *intrusive* measures, which means that they explicitly need a reference signal or channel while human subjects rely on their internal reference.

3.1 Channel-Based Measures

Objective measures to characterize room impulse responses are mostly based on the energy ratio of early and late part of the RIR, see, e.g., [34]. Since the IR of an equalized acoustic channel $\mathbf{v}$ may look slightly different compared to a normal RIR (e.g., pre-echoes before the main peak) some objective measures were adapted from their original definitions to account for this. Fig. 8 shows such an equalized acoustic channel and illustrates the definitions of the lags n_0, which is the position of the main peak of the impulse response, $N_{50} = \lfloor 0.05\ \mathrm{s} \cdot f_s \rfloor$ and $N_{80} = \lfloor 0.08\ \mathrm{s} \cdot f_s \rfloor$, which are the samples 50 ms and 80 ms later than the main peak, respectively.

The definitions of six measures that are widely used to characterize RIRs are given in the following for the

equalized acoustic channel **v** and are also applicable for an RIR **h**.

The ratio between the energy of the first 50 or 80 ms of the IR after the main peak to the overall energy of the IR is called *Definition* and is denoted by D_{50} or D_{80}, respectively [34]:

$$D_{\{50,80\}} = \frac{\sum_{n=n_0}^{n_0+N_{\{50,80\}}-1} v_n^2}{\sum_{n=0}^{L_v-1} v_n^2}. \tag{15}$$

The *Clarity* [34], denoted here by C_{50} or C_{80}, is the logarithmic ratio of the energy within 50 or 80 ms after the main peak to remaining energy of the impulse response:

$$C_{\{50,80\}} = 10\log_{10}\frac{\sum_{n=n_0}^{n_0+N_{\{50,80\}}-1} v_n^2}{\sum_{n=0}^{n_0-1} v_n^2 + \sum_{n=n_0+N_{\{50,80\}}}^{L_v-1} v_n^2}. \tag{16}$$

Different from the usual definition of the $C_{\{50,\,80\}}$ measure, which is often defined as the ratio of energy of the first 50 ms of a room impulse response **h** to the energy of the remaining part [34], the lags of the equalized impulse response **v** preceding the main peak at position n_0 (cf., Fig. 8) contribute to the denominator in the first summation term in Eq. (16). These lags can be neglected for common room impulse responses since their initial peak usually is the main peak or at least the preceding energy can be neglected. However, for equalized impulse responses **v**, energy before the main peak may be perceived as disturbance (pre-ringing) and, thus, should contribute to the distortion part in the denominator of Eq. (16).

The *Direct-to-Reverberation-Ratio* DRR [35] is defined as the logarithmic ratio between the energy of the direct path of the impulse response and the energy of all reflections. However, since the direct path, in general, does not match the sampling grid, a small range n_Δ around the main peak is considered as the direct path energy [5, 35]:

$$\mathrm{DRR} = 10\log_{10}\frac{\sum_{n=n_0-n_\Delta}^{n_0+n_\Delta-1} v_n^2}{\sum_{n=0}^{n_0-n_\Delta-1} v_n^2 + \sum_{n=n_0+n_\Delta}^{L_v} v_n^2}. \tag{17}$$

In Eq. (17), we chose $n_\Delta = 4\ \mathrm{ms} \cdot f_s$.

The *Center Time* CT [34] is not defined as a ratio but as the center of gravity in terms of the energy of the RIR:

$$\mathrm{CT} = \frac{\sum_{n=0}^{L_v} n \cdot v_n^2}{\sum_{n=0}^{L_v} v_n^2}. \tag{18}$$

Additionally to the time-domain measures described above, we evaluated two common spectral channel-based measures to account for the coloration effect [2, 24]. Since equalization often aims at a flat spectrum, it was proposed in [9, 36] to use the *variance* (VAR) of the logarithmic overall transfer function $V_k = H_k\, C_{\mathrm{EQ},k}$ as an objective quality measure to evaluate LRC algorithms:

$$\mathrm{VAR} = \frac{1}{K_{\max} - K_{\min} + 1} \sum_{k=K_{\min}}^{K_{\max}} \left(20\log_{10}|V_k| - \bar{V}_{\mathrm{dB}}\right)^2 \tag{19}$$

with

$$\bar{V}_{\mathrm{dB}} = \frac{1}{K_{\max} - K_{\min} + 1} \sum_{k=K_{\min}}^{K_{\max}} 20\log_{10}|V_k|. \tag{20}$$

In Eq. (19), $\bar{V}_{\mathrm{dB}}$ is the mean logarithmic spectrum and $K_{\min}$ and $K_{\max}$ are the frequency indices that limit the considered frequency range in which the equalized transfer function is desired to be flat. We chose $K_{\min}$ and $K_{\max}$ corresponding to 200 Hz and 3700 Hz to account for a high-pass or band-pass characteristic of the desired system vector in Eq. (4).

A second measure for the quality of equalization in frequency-domain is the *spectral flatness measure* (SFM) that is the ratio of geometric mean and the arithmetic mean of V_k [37]:

$$\mathrm{SFM} = \frac{\sqrt[K]{\prod_{k=0}^{K-1} |V_k|^2}}{\frac{1}{K}\sum_{k=0}^{K-1} |V_k|^2}. \tag{21}$$

In Eq. (21), K denotes the number of frequency bins.

3.2 Signal-Based Measures

For non-linear dereverberation suppression approaches as in [5], impulse responses or transfer functions are not obtainable or applicable for objective testing. Thus, such algorithms have to be evaluated based on the signals only. Several signal-based measures that exist for assessment of LRC approaches and dereverberation suppression approaches are briefly summarized in the following. Due to the large extent of this topic, the interested reader is referred to the respective references for more details and further reading. Simple measures like the *Segmental Signal-to-Reverberation Ratio* (SSRR) [1] are defined similarly to SNR-based measures known from noise-reduction quality assessment. As already known from speech quality assessment for noise reduction, quality measures incorporating models of the human auditory system show higher correlation with subjective rating [21]. The *Frequency-Weighted SSRR* (FWSSRR) [38] and the *Weighted Spectral Slope* (WSS) [38] represent a first step toward consideration of the human auditory system by analyzing the SSRR in critical bands. To account for logarithmic loudness perception within the human auditory system the *Log-Spectral Distortion* (LSD) compares logarithmically weighted spectra. Since dereverberation of speech is the aim in most scenarios, we also tested measures based on the LPC models such as the *Log-Area Ratio* (LAR) [39], the *Log-Likelihood Ratio* (LLR) [38], the *Itakura-Saito Distance* (ISD) [38], and the *Cepstral Distance* (CD) [38]. As a further extension toward modeling of the human auditory system the *Bark Spectral Distortion* measure (BSD) [40] compares perceived loudness incorporating spectral masking effects. Recently, objective measures have been proposed especially designed for assessment of dereverberation algorithms. For this contribution we tested the *Reverberation Decay Tail* (RDT) measure [41], the *Speech-to-Reverberation Modulation Energy Ratio* (SRMR) [42], and the *Objective Measure for Coloration in Reverberation* (OMCR) [43].

Table 2. Properties of sound samples used for the subjective listening test.

Sample no.	τ_{60} of RIR	LRC filter type	LRC filter length L_{EQ}	gender of speaker
1	1000 ms	WLS-EQ	2048	male
2	1000 ms	ISwPP	4096	female
3	500 ms	LS-EQ	2048	male
4	1000 ms	WLS-EQ	8192	male
5	500 ms	ISwPP	1024	male
6	500 ms	WLS-EQ	4096	male
7	1000 ms	WLS-EQ	4096	female
8	500 ms	ISwPP	8192	female
9	1000 ms	LS-EQ	8192	female
10	500 ms	ISwINO	4000	male
11	500 ms	WLS-EQ	1024	male
12	500 ms	LS-EQ	1024	female
13	1000 ms	LS-EQ	1024	female
14	500 ms	ISwPP	4096	male
15	500 ms	WLS-EQ	8192	male
16	1000 ms	LS-EQ	4096	male
17	1000 ms	LS-EQ	2048	male
18	500 ms	ISwPP	2048	female
19	500 ms	LS-EQ	4096	male
20	500 ms	LS-EQ	8192	male
21	1000 ms	ISwPP	1024	male

From quality assessment in the fields of audio coding and noise reduction it is known that measures that are based on more exact models of the human auditory system show high correlation with subjective data [21]. Thus, we also tested the *Perceptual Evaluation of Speech Quality* (PESQ) measure [38, 44] and the *Perceptual Similarity Measure* (PSM, PSM_t) from PEMO-Q [45] that compares internal representations according to the auditory model of [46].

4 SUBJECTIVE QUALITY ASSESSMENT

For the subjective listening tests, reverberant speech samples were calculated by first convolving RIRs generated by the image method [47] for a room having a size of 6 m × 4 m × 2.6 m (length × width × height) with male and female utterances of about 7 seconds in length (consisting of about 20 words). Pilot listening tests using measured RIRs have shown results similar to those measured with simulated RIRs, thus we restricted the following listening test to the use of simulated RIRs where we adjusted the reverberation time by changing the wall reflection coefficients in the room model [47]. The distance between sound source and microphone was approximately 0.8 m. Room reverberation times were approximately $\tau_{60} = \{500, 1000\}$ ms corresponding to normal and somewhat larger office environments. These reverberant speech samples were then processed by the four LRC approaches discussed in Sec. 1 and presented to the subjects. Filter lengths of the equalizers were $L_{EQ} = \{1024, 2048, 4096, 8196\}$ at a sampling rate of 8 kHz. The parameter α in Eq. (9) was set to 0.8.

From all 64 possible speech samples (2 room reverberation times × 4 LRC approaches × 4 LRC filter lengths × 2 genders), 21 audio samples that represent a wide variety of acoustic conditions and possible distortions were chosen. These audio samples had a length of 8 s and were scaled to have the same level (root-mean-square).

The properties of the chosen sound samples are summarized in Table 2 and an audiovisual presentation of the samples and the corresponding channels can be found in [48]. They were presented diotically, i.e., the same signal was played back for left and right ear, to 24 normal-hearing listeners via headphones (Sennheiser HD650) in quiet (in a sound proof booth) after a training period by example audio samples. The training samples consisted of all signals used in the later test to give the listeners the possibility to get familiar with the sound samples and their respective quality and distortions. Training and listening could be repeated as often as desired, however, none of the subjects repeated listening to the training samples during the actual listening tests although the possibility was provided. The initial training period before the actual listening test was mandatory and, thus, done by all listeners. A graphical user interface was programmed for the listening test as depicted in Fig. 9 based on the suggestions of [14] (with slight differences) asking to assess the attributes *reverberant, colored (distorted), distant*, and *overall quality* on a 5-point *Mean Opinion Score* (MOS) scale for subjective listening quality (MOS-LQS).

As stated in ITU recommendation P.835 [14] for noise-reduction schemes for hands-free systems, the perceived quality after signal enhancement algorithms should be assessed in different dimensions, i.e., overall quality, signal distortion, and reduction of the disturbance. These categories were adopted for our test. It is known that reverberation influences the signal in terms of the coloration effect, and the reverberation decay tail effect [34, 41, 43]. For our subjective test, the attribute *distant* was added since the authors expected in the beginning that the attribute *reverberant* is more difficult to assess for non-expert listeners. Thus it was expected that the attributes *reverberant* and *distant* would lead to similar results. Since for LRC algorithms frequency distortion is perceptually much more prominent

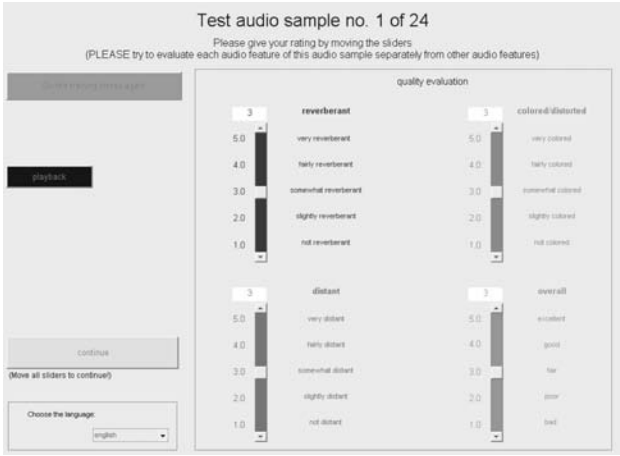

Fig. 9. Subjective speech quality evaluation of the audio samples for the attributes *reverberant, colored/distorted, distant*, and *overall quality*.

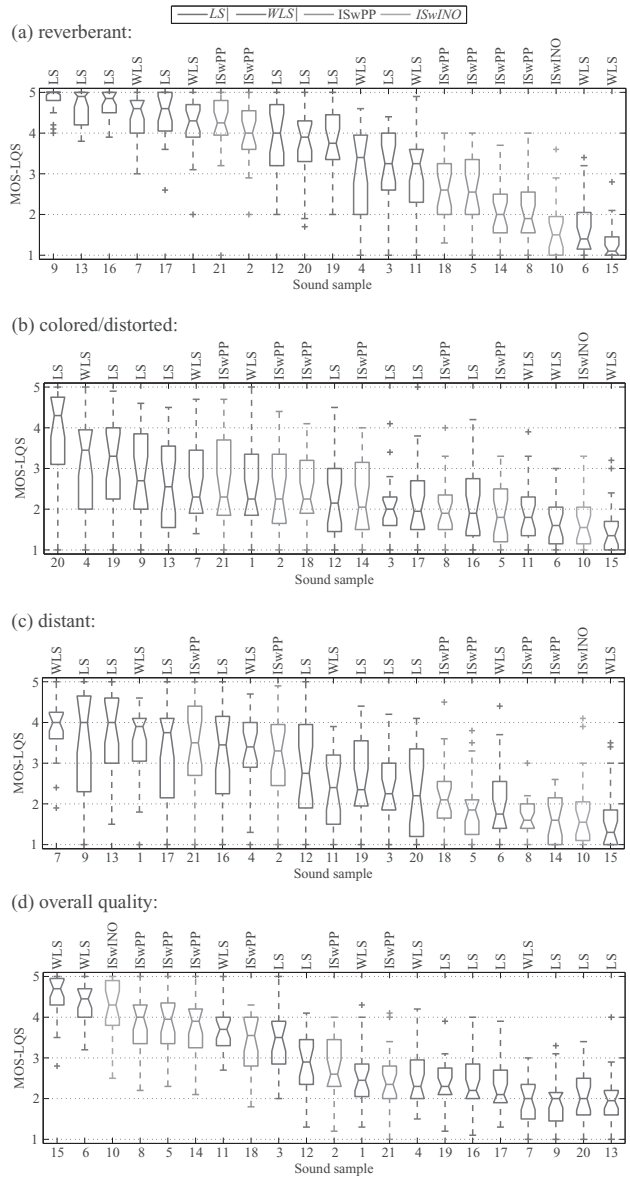

Fig. 10. Subjective rating of sound samples for attribute (a) *reverberant*, (b) *colored/distorted*, (c) *distant*, and (d) *overall quality*

than what usually is understood as coloration, we asked to assess *coloration/distortion* as one spectral attribute. This leads to the fact that common measures that were designed to assess coloration may not correlate well to the subjective data. However, these distortions dominate the spectral perception of subjective quality. Quality assessment was possible in steps of 0.1 between 1.0 and 5.0. A more detailed overview of the training and listening test as well as the GUI can also be obtained from [48].

5 RESULTS

5.1 Rating of the Sound Samples

The subjective ratings of the sound samples [48] for the four attributes *reverberant*, *colored/distorted*, *distant*, and *overall quality* are shown in Fig. 10 by means of box-plots.

The sound samples are ordered according to their median value for the respective attribute. Consequently, the order is different for the different sub-figures.

The subjective ratings were normally distributed (verified by Kolmogorov-Smirnov test) that allowed for conduction of an analysis of variance (ANOVA). A two-way ANOVA revealed significant main effects of attribute type $\{F(3, 2112) = 18.8, p < 0.001\}$ and LRC approach $\{F(3, 2112) = 97.4, p < 0.001\}$. Post-hoc comparisons (Bonferroni tests with level of significance set at 5%) for the factor LRC approach showed statistical differences between all algorithms used with the highest quality for the ISwINO approach and the lowest for the LS approach. Generally, the shaping approaches (i.e., ISwPP and ISwINO) resulted in better rating scores than the least-squares approaches (i.e., LS and WLS).

Increasing the filter length of the LS approach does not necessarily improve the subjective results considerably due to the fact that despite a "good equalization" perceptually relevant late echoes and pre-echoes are clearly perceived as disturbing by the listeners (see, e.g., sound samples no. 9 ($L_{EQ} = 8192$) and no. 13 ($L_{EQ} = 1024$) both for an RIR with $\tau_{60} = 800$ ms).

The differences in the subjective scores between all used attributes were also statistically significant. Therefore, a separate one-way ANOVA was conducted for each attribute to test the quality of the different LRC approaches. For the attribute *reverberant*, the best ratings (indicated by the lowest rating scores) were obtained for the ISwINO algorithm with a mean value of 1.6. The ratings achieved by the ISwINO were significantly better than all remaining algorithms. The scores for the ISwPP and the WLS approach were 1.3 and 1.4 points higher than for the ISwINO approach, respectively (meaning that signals processed by the ISwINO or WLS approach were assessed as being more reverberant than these processed by the ISwINO). No statistically significant differences in rating were found between the ISwPP and WLS approach ($p = 1.0$). The lowest

Table 3. Inter-attribute correlations.

Attribute	Colored/distorted	Distant	Overall
Reverberant	0.44	**0.91**	**0.94**
Colored/distorted	-	0.29	0.66
Distant	-	-	0.86

quality for the attribute *reverberant* was found for the the LS approach with the mean rating score of 4.1. Exactly the same trends were observed for the attribute *overall quality*. Slightly different trends regarding the statistical dependencies of the LRC approaches were observed for the attribute *distant*. The best quality scores were again obtained for the shaping approaches, however, with no significant differences between the ISwINO and ISwPP algorithm ($p = 0.164$). Both least-square approaches were again assessed worse than the shaping approaches and resulted in on average 0.8 points higher rating scores. A different trend between the attributes might be related to the fact that for the assessment of the attribute *distant* the differences between the four different approaches were smaller than for the attribute *reverberant* or *overall quality*. Although it seems from panels (a) and (c) of Fig. 10 that the variance for the attribute *distant* is higher, results show similar standard errors for attributes *reverberation* and *distant*. However, for the attribute *reverberant* subjects more often decided for the maximum score of an MOS of 5 (very reverberant) that may be due to the fact that a clearer anchor for high reverberation was given in the training samples than for "very distant." The post-hoc comparisons for the attribute *colored* revealed again the significantly highest quality for the ISwINO approach. No significant differences were found between the ISwPP, WLS, and LS algorithm; however, from Fig. 10 it can be seen that the LS approach usually performs worse than the other approaches, which may be due to the fact that late echoes typical for the LS approaches sometimes sound like distortions.

Table 3 shows the inter-attribute correlations for the given set of speech samples. As expected, the attributes *reverberant* and *distant* show high inter-attribute correlation (0.91) although the attribute *distant* leads to a higher interquartile range (IQR) as it can be seen comparing panels (a) and (c) in Fig. 10. Furthermore, the correlation between the attributes *overall quality* and the attributes *distant* as well as *reverberant* is high. Thus, the perceived audio quality is strongly influenced by reverberation (including late reverberation). The attribute *reverberant* seems to be suitable to assess the *overall quality* since it has the highest correlation (0.94) for the given sound samples and LRC approaches.

5.2 Correlation Analysis

The correlations of subjective rating for the four attributes and the channel-based objective measures are shown in Table 4 while correlations with signal-based objective measures are shown in Table 5.

For each objective measure correlations with the subjective ratings are given for the case that all LRC approaches of Sec. 1 are considered (Method: All EQs) and for the case that only one LRC approach is used. For the latter case no correlation was calculated for the impulse-response shaping approach based on infinity-norm optimization because the number of sound samples was too low for a reliable correlation analysis. The highest correlation for each attribute and approach is highlighted in boldface in the tables. Each column of Tables 4 and 5 contains four indicated maxima, one for the overall correlations ("all EQs") and one for each individual LRC approach ("LS-EQ," "WLS-EQ," and "ISwPP"). The reason for additionally calculating correlations for each LRC approach separately is exemplarily illustrated in Fig. 11 for the SFM.

Table 4. Correlations $|\rho|$ of MOS values of subjective ratings and channel-based objective measures (maxima are indicated in boldface).

Measure	Method	Reverberant	Col./dist.	Distant	Overall
D_{50}	All EQs	0.860	0.629	0.937	0.910
	LS-EQ	0.711	0.329	0.795	0.794
	WLS-EQ	0.942	0.735	**0.993**	**0.982**
	ISwPP	0.943	0.611	0.940	0.934
D_{80}	All EQs	0.905	0.504	0.911	0.904
	LS-EQ	0.733	0.311	0.815	0.817
	WLS-EQ	0.941	0.585	0.976	0.931
	ISwPP	0.850	0.546	0.844	0.844
C_{80}	All EQs	**0.930**	0.607	0.888	0.907
	LS-EQ	0.804	0.305	0.865	0.877
	WLS-EQ	0.982	0.690	0.987	0.963
	ISwPP	0.916	0.543	0.899	0.882
C_{50}	All EQs	0.926	**0.665**	**0.944**	**0.935**
	LS-EQ	0.783	0.320	0.846	0.857
	WLS-EQ	**0.965**	0.755	0.981	0.971
	ISwPP	**0.976**	0.580	0.958	0.933
CT	All EQs	0.845	0.607	0.927	0.911
	LS-EQ	**0.909**	0.288	**0.938**	**0.949**
	WLS-EQ	0.857	0.785	0.958	0.966
	ISwPP	0.973	0.667	**0.979**	**0.974**
DRR	All EQs	0.238	0.101	0.179	0.131
	LS-EQ	0.769	0.335	0.835	0.843
	WLS-EQ	0.399	**0.858**	0.597	0.696
	ISwPP	0.249	**0.692**	0.273	0.360
VAR	All EQs	0.028	0.374	0.231	0.156
	LS-EQ	0.618	**0.416**	0.708	0.694
	WLS-EQ	0.687	0.809	0.841	0.883
	ISwPP	0.599	0.462	0.608	0.647
SFM	All EQs	0.132	0.267	0.126	0.048
	LS-EQ	0.686	0.376	0.769	0.765
	WLS-EQ	0.709	0.821	0.861	0.899
	ISwPP	0.876	0.658	0.885	0.905

As it can be seen from Fig. 11, the SFM shows much higher correlation when a single rather than all LRC approaches are considered. However, the time-domain channel-based measures show consistent correlations for all LRC approaches. The interested reader is referred to [48] for an overview of all correlation patterns.

It can be seen from Table 4 that the time-domain channel-based objective measures show high correlation with the subjective data for the attributes *reverberation, distance,* and *overall quality* (with the exception of the DRR measure). The frequency-domain channel-based measures VAR and SFM show much lower correlation. However, as stated before, they may show somewhat higher correlation for

Table 5. Correlations $|\rho|$ of MOS values of subjective ratings and signal-based objective measures (maxima are indicated in boldface).

Measure	Method	Reverberant	Col./dist.	Distant	Overall
SSRR	All EQs	0.332	0.290	0.432	0.403
	LS-EQ	0.596	0.152	0.648	0.673
	WLS-EQ	0.802	0.737	0.827	0.798
	ISwPP	0.703	0.338	0.652	0.641
FWSSRR	All EQs	0.440	0.404	0.568	0.551
	LS-EQ	0.792	0.037	0.821	0.852
	WLS-EQ	0.943	0.778	0.989	**0.984**
	ISwPP	0.807	0.458	0.763	0.752
WSS	All EQs	0.603	0.580	0.762	0.713
	LS-EQ	0.788	0.441	0.866	0.847
	WLS-EQ	0.892	0.760	0.959	0.981
	ISwPP	0.909	0.580	0.874	0.860
ISD	All EQs	0.639	0.347	0.693	0.684
	LS-EQ	0.352	0.444	0.364	0.408
	WLS-EQ	**0.964**	0.709	**0.999**	0.980
	ISwPP	0.701	0.374	0.672	0.677
CD	All EQs	0.627	0.414	0.702	0.674
	LS-EQ	0.445	0.371	0.478	0.523
	WLS-EQ	0.893	0.811	0.942	0.933
	ISwPP	0.797	0.416	0.749	0.731
LAR	All EQs	0.517	0.384	0.612	0.588
	LS-EQ	0.332	0.504	0.356	0.419
	WLS-EQ	0.934	0.779	0.985	0.976
	ISwPP	0.749	0.386	0.700	0.686
LLR	All EQs	0.663	0.432	0.753	0.713
	LS-EQ	0.469	0.365	0.495	0.544
	WLS-EQ	0.893	0.845	0.956	0.962
	ISwPP	0.836	0.450	0.795	0.778
LSD	All EQs	0.735	0.480	0.814	0.780
	LS-EQ	0.753	0.065	0.809	0.832
	WLS-EQ	0.867	0.834	0.923	0.921
	ISwPP	0.865	0.500	0.833	0.823
BSD	All EQs	0.043	0.303	0.237	0.195
	LS-EQ	0.526	0.470	0.634	0.602
	WLS-EQ	0.848	0.644	0.938	0.937
	ISwPP	0.907	0.635	0.926	0.937
OMCR	All EQs	0.051	0.134	0.028	0.052
	LS-EQ	0.519	**0.827**	0.620	0.538
	WLS-EQ	0.631	0.233	0.640	0.649
	ISwPP	0.163	0.453	0.239	0.257
RDT	All EQs	0.670	0.505	0.790	0.746
	LS-EQ	0.690	0.430	0.776	0.767
	WLS-EQ	0.810	0.745	0.883	0.933
	ISwPP	0.943	0.574	0.922	0.901
SRMR	All EQs	0.526	0.242	0.593	0.511
	LS-EQ	0.437	0.154	0.509	0.538
	WLS-EQ	0.747	**0.885**	0.734	0.803
	ISwPP	0.785	0.451	0.722	0.695
PSM	All EQs	0.803	**0.627**	0.902	0.866
	LS-EQ	0.844	0.642	0.905	0.877
	WLS-EQ	0.843	0.832	0.922	0.971
	ISwPP	**0.982**	0.653	0.963	0.945
PSM_t	All EQs	**0.915**	0.611	**0.950**	**0.942**
	LS-EQ	**0.895**	0.558	**0.958**	**0.920**
	WLS-EQ	0.896	0.761	0.960	**0.984**
	ISwPP	0.979	**0.787**	**0.970**	**0.964**
PESQ	All EQs	0.596	0.349	0.691	0.628
	LS-EQ	0.465	0.354	0.503	0.552
	WLS-EQ	0.842	0.772	0.898	0.874
	ISwPP	0.893	0.458	0.847	0.816

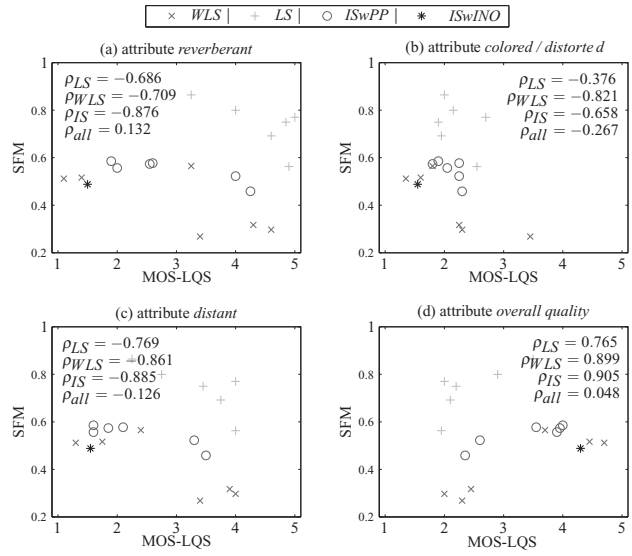

Fig. 11. Correlations of subjective ratings and SFM measure for all four attributes.

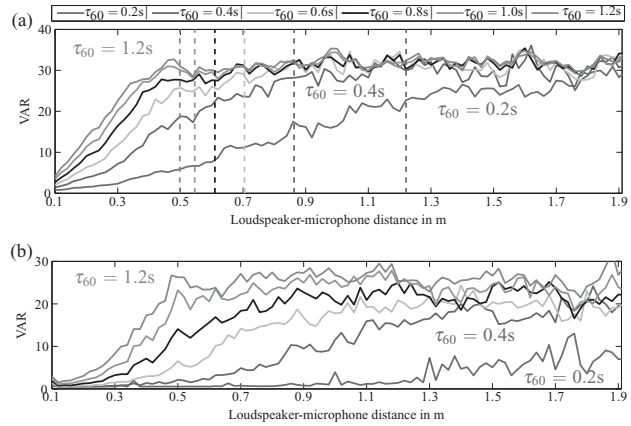

Fig. 12. VAR measure of (a) RIR H_k and (b) equalized acoustic channel V_k over loudspeaker-microphone distance for different room reverberation times (critical distances are indicated as dashed vertical lines). Sub-figure (b) shows the VAR measure for an equalized acoustic channel using an LS-EQ with $L_{\mathrm{EQ}} = 2048$ at $f_s = 8$ kHz.

single LRC approaches such as SFM for the WLS-EQ. In general, and this is also true for the signal-based measures (cf., Table 5), only low correlation was obtained with the attribute *colored/distorted* for all measures. This can be attributed to the fact that the source-receiver distance for our experiment (0.8 m) is larger than the critical distance.

To clarify this fact, the dependence of the frequency-domain quality measure *variance* on the distance between source and microphone is visualized in Fig. 12 for a common RIR (upper panel) and an equalized impulse response (lower panel) for different room reverberation times τ_{60}. The critical distance for each reverberation time is additionally indicated in the upper panel of Fig. 12 by a vertical dashed line. It can be seen that the variance does not further increase once it reaches a maximum value. This observation is in consilience with the findings in [5, 36]. The maximum

value was calculated to be at about 31 dB in [36] for RIRs. This point is approximately reached at the critical distance as it is shown in Fig. 12. However, another reason for lower correlations for the spectral measure VAR and SFM may be that they equally assess spectral peaks that are perceived as being very annoying [25] and spectral dips that do not decrease the perceived quality to a great extent.

Table 5 shows the correlations of subjective ratings with signal-based objective measures. It can be seen that the signal-based measures generally show lower correlation to subjective data than the channel-based measures. The LPC-based measures outperform purely signal-based measures like the SSRR. By far, the highest correlations are obtained by the measures PSM and PSM_t that rely on auditory models. PSM_t, in addition to PSM, evaluates short-time behavior of the correlations of internal signal representations and focuses on low correlations as it is done by human listeners [45]. The auditory-model based measures show even higher correlation than RDT, SRMR, and OMCR although the latter were designed to explicitly assess reverberation. The performance of RDT and OMCR measures can be adjusted by changing internal parameters. By this, higher correlation to the specific set of samples can be obtained. However, we used standard values for these parameters given in [41, 43]. Furthermore, it has to be emphasized that the attribute *coloration/distortion* is most difficult to assess by objective measures at least for the discussed LRC algorithms, since distortions are perceptually relevant and measures like OMCR try to assess coloration effects only (the same holds for the variance measure). They succeed in doing so, but coloration alone is not well correlated to our subjective data due to distortions like late echoes and pre-echoes that are much more prominent than the coloration effect [48]. As the tested measures are incapable of explicitly assessing those influences further development of objective measures is required.

6 CONCLUSION

Objective quality measures were compared to data from subjective listening tests to identify objective measures that can be used to evaluate the performance of listening-room compensation algorithms. Channel-based measures showed higher correlations between objective and subjective data than most of the tested signal-based measures. However, especially if impulse responses are not properly accessible, e.g., as for dereverberation suppression algorithms, measures that incorporate sophisticated auditory models should be used for quality assessment. The Perceptual Similarity Measure (PSM) showed highest correlations to subjective data. A detailed assessment of coloration effects and distortions that may be introduced by LRC algorithms is a topic for future research.

7 ACKNOWLEDGMENT

This work was supported in parts by the German Research Foundation DFG under Grant Ka841-17 and the EU ITN Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS, project no. 316969). The authors would like to thank Anna Warzybok for help with the statistical analysis.

8 REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, "Speech Dereverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands (Sept. 2005).

[2] J. B. Allen, "Effects of Small Room Reverberation on Subjective Preference," *J. Acous. Soc. Am. (JASA)*, vol. 71, no. 1, p. S5 (1982).

[3] D. A. Berkley, "Normal Listeners in Typical Rooms—Reverberation Perception, Simulation, and Reduction," in *Acoustical Factors Affecting Hearing Aid Performance*, pp. 3–24 (University Park Press, Baltimore, 1980).

[4] IEC 1998, "Sound System Equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index" (1998).

[5] E. A. P. Habets, *Single and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*, Ph.D. thesis, University of Eindhoven, Eindhoven, The Netherlands (June 2007).

[6] J. Benesty, Y. Huang, and J. Chen, "A Blind Channel Identification-Based Two-Stage Approach to Separation and Dereverberation of Speech Signals in a Reverberant Environment," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895 (Sept. 2005).

[7] S. Goetze, M. Kallinger, A. Mertins, and K.-D. Kammeyer, "System Identification for Multi-Channel Listening-Room Compensation Using an Acoustic Echo Canceller," in *Proc. Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, pp. 224–227 (May 2008).

[8] S. J. Elliott and P. A. Nelson, "Multiple-Point Equalization in a Room Using Adaptive Digital Filters," *J. Audio Eng. Soc.*, vol. 37, pp. 899–907 (1989 Nov.).

[9] J. N. Mourjopoulos, "Digital Equalization of Room Acoustics," *J. Audio Eng. Soc.*, vol. 42, pp. 884–900 (1994 Nov.).

[10] S. Goetze, E. Albertin, M. Kallinger, A. Mertins, and K.-D. Kammeyer, "Quality Assessment for Listening-Room Compensation Algorithms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA (Mar. 2010).

[11] M. Morimoto, H. Sato, and M. Kobayashi, "Listening Difficulty as a Subjective Mmeasure for Evaluation of Speech Transmission Performance in Public Spaces," *J. Acous. Soc. Am.*, vol. 116, no. 3, pp. 1607–1613 (2005).

[12] D. U. Ebem, J. G. Beerends, J. Van Vugt, C. Schmidmer, R. E. Kooij, J. O. Uguru, "The Impact of Tone Language and Non-Native Language Listening on Measuring Speech Quality," *J. Audio Eng. Soc.*, vol. 59, pp. 647–655 (2011 Sep.).

[13] M. Huckvale and G. Hilkhuysen, "Performance-Based Measurement of Speech Quality with an Audio

Proof-Reading Task," J. Audio Eng. Soc., vol. 60, pp. 444–451 (2012 June).

[14] ITU-T P.835, "Subjective Test Methology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm, ITU-T Recommendation P.835" (Nov. 2003).

[15] ITU-R BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunication Union, Geneva, Switzerland (2003).

[16] ITU-T P.800, "Method for Subjective Determination of Transmission Quality," *International Telecommunication Union*, Geneva, Switzerland (1996).

[17] A. Raake, M. Wältermann, U. Wüstenhagen, B. Feiten, "How to Talk about Speech and Audio Quality with Speech and Audio People," *J. Audio Eng. Soc.*, vol. 60, pp. 147–155 (2012 Mar.).

[18] M. Wältermann, A. Raake, and S. Möller, "Direct Quantification of Latent Speech Quality Dimensions," *J. Audio Eng. Soc.*, vol. 60, pp. 246–254 (2012 Apr.).

[19] N. Côté, V. Koehl, S. Möller, A. Raake, M. Wältermann, and V. Gautier-Turbin, "Diagnostic Instrumental Speech Quality Assessment in a Super-Wideband Context," *J. Audio Eng. Soc.*, vol. 60, pp. 156–164 (2012 Mar.).

[20] R. Huber, *Objective Assessment of Audio Quality Using an Auditory Processing Model, Ph.D. thesis*, University of Oldenburg, Germany (2003).

[21] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Measures for the Evaluation of Noise Reduction Schemes," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)* (2005).

[22] E. Hänsler and G. Schmidt, "Acoustic Echo and Noise Control: a Practical Approach" *(Wiley, Hoboken*, NJ, 2004).

[23] S. Möller, F. Kettler, H.-W. Gierlich, S. Poschen, N. Côté, A. Raake, and M. Wältermann, "Extending the E-Model for Capturing Noise Reduction and Echo Canceller Impairments," *J. Audio Eng. Soc.*, vol. 60, pp. 165–175 (2012 Mar.).

[24] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of Speech Dereverberation Algorithms Using the MARDY Database," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France (Sept. 2006).

[25] M. Kallinger and A. Mertins, "Room Impulse Response Shaping—A Study," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. V101–V104 (2006).

[26] A. Mertins, T. Mei, and M. Kallinger, "Room Impulse Response Shortening/Reshaping with Infinity- and *p*-Norm Optimization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 249–259 (Feb. 2010).

[27] S. Goetze, M. Kallinger, A. Mertins, and K.-D. Kammeyer, "Multi-Channel Listening-Room Compensation Using a Decoupled Filtered-X LMS Algorithm," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, USA, pp. 811–815 (Oct. 2008).

[28] S. T. Neely and J. B. Allen, "Invertibility of a Room Impulse Response," *J. Acous. Soc. Am. (JASA)*, vol. 66, pp. 165–169 (July 1979).

[29] S. Goetze, M. Kallinger, A. Mertins, and K.-D. Kammeyer, "Estimation of the Optimum System Delay for Speech Dereverberation by Inverse Filtering," in *Int. Conf. on Acoustics (NAG/DAGA 2009)*, Rotterdam, The Netherlands, pp. 976–979 (Mar. 2009).

[30] L. D. Fielder, "Practical Limits for Room Equalization," presented at the 111th *Convention of the Audio Engineering Society* (Nov. 2001), convention paper 5481.

[31] S. M. Griebel and M. S. Brandstein, "Wavelet Transform Extrema Clustering for Multi-Channel Speech Dereverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Pocono Manor, PA, USA (Sept. 1999).

[32] B. Yegnanarayana and P. S. Murthy, "Enhancement of Reverberant Speech Using LP Residual Signal," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–280 (May 2000).

[33] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-Based Performance Evaluation of Dereverberation Algorithms," *J. Electrical & Computer Eng., Article ID 127513* (2010).

[34] H. Kuttruff, *Room Acoustics*, 4th Edition (Spoon Press, London, 2000).

[35] M. Triki and D. T. M. Slock, "Iterated Delay and Predict Equalization for Blind Speech Dereverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France (Sept. 2006).

[36] J. J. Jetzt, "Critical Distance Measurement of Rooms from the Sound Energy Spectral Response," *J. Acous. Soc. Am. (JASA)*, vol. 65, no. 5, pp. 1204–1211 (May 1979).

[37] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Communication*, vol. 6, no. 2, pp. 314–232 (Feb. 1988).

[38] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press Inc., Boca Raton, LA, USA, 2007).

[39] J. H. L. Hansen and B. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, vol. 7, pp. 2819–2822 (Dec. 1998).

[40] W. Yang, Enhanced Modified Bark Sectral Distortion (EMBSD): A Objective Speech Quality Measure Based on Audible Distortion and Cognition Model, Ph.D. thesis, Temple University, Philadelphia, USA (May 1999).

[41] J. Y. C. Wen and P. A. Naylor, "An Evaluation Measure for Reverberant Speech Using Decay Tail Modeling," in *Proc. EURASIP European Signal Processing Conference (EUSIPCO)*, Florence, Ital (Sept. 2006).

[42] T. H. Falk and W.-Y. Chan, "A Non-Intrusive Quality Measure of Dereverberated Speech," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA (Sept. 2008).

[43] J. Y. C. Wen and P. A. Naylor, "Objective Measurement of Colouration in Reverberation," in *Proc. EURASIP*

European Signal Processing Conference (EUSIPCO), Poznan, Poland (Sept. 2007), pp. 1615–1619.

[44] ITU-T P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, ITU-T Recommendation P.862" (Feb. 2001).

[45] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. on Audio, Speech and Language Processing - Special Issue on Objective Quality Assessment of Speech and Audio*, vol. 14, no. 6 (2006).

[46] T. Dau, D. Püschel, and A Kohlrausch, "A Quantitative Model of the Effective Signal Processing in the Auditory System: I. Model Structure," *J. Acous. Soc. Am. (JASA)*, vol. 99, no. 6, pp. 3615–3622 (June 1996).

[47] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acous. Soc. Amer.*, vol. 65, pp. 943–950 (1979).

[48] Sound samples, correlation patterns, and MATLAB code for quality assessment available online at http://www.ant.uni-bremen.de/∼goetze/aes2010/.

NOMENCLATURE

ANOVA = Analysis of variance
BSD = Bark Spectral Distortion, [40]
C50, C80 = Clarity, [34]
CD = Cepstral Distance, [38]
CT = Center Time, [34]
D50, D80 = Definition, [34]
DRR = Direct-to-Reverberation-Ratio, [35]
FWSSRR = Frequency-Weighted SSRR, [38]
ISD = Itakura-Saito Distance, [38]
ISwPP = impulse response shaping with post processing
ISwINO = impulse response shaping with infinity-norm optimization [26]
LAR = Log-Area Ratio, [39]
LLR = Log-Likelihood Ratio, [38]
LRC = listening-room compensation
LS = least-squares
LSD = Log-Spectral Distortion, [38]
MOS-LQS = mean opinion score for listening quality (subjective)
MOS-LQO = mean opinion score for listening quality (objective)
OMCR = Objective Measure for Coloration in Reverberation, [43]
PESQ = Perceptual Evaluation of Speech Quality, [38]
PPMCC = Pearson product-moment correlation coefficient
PSM, PSM_t = Perceptual Similarity Measure, [45]
RDT = Reverberation Decay Tail, [41]
SFM = Spectral Flatness Measure, [37]
SRMR = Speech-to-Reverberation Modulation Energy Ratio, [42]
SSRR = Segmental Signal-to-Reverberation Ratio, [1]
VAR = Variance of logarithmic transfer function, [9]
WLS = weighted least-squares
WSS = Weighted Spectral Slope, [38]

THE AUTHORS


Stefan Goetze


Eugen Albertin


Jan Rennies


Emanuël A.P. Habet


Karl-Dirk Kammeyer

Stefan Goetze is head of Audio System Technology for Assistive Systems at the Fraunhofer Institute for Digital Media Technology (IDMT), project group Hearing, Speech and Audio (HSA) in Oldenburg, Germany. He received his Dipl.-Ing. and Dr.-Ing. in 2004 and 2013, respectively, at the University of Bremen, Germany, where he worked as a research engineer from 2004 to 2008. His research interests are assistive technologies, sound pick/up and enhancement, such as noise reduction, acoustic echo cancellation and dereverberation, as well as detection and classification of acoustic events and automatic speech recognition. He is lecturer at the University of Bremen and project leader of national and international projects in the field of ambient assisted living (AAL). He is member of IEEE and AES.

•

Eugen Albertin received his Diploma degree (Dipl.-Ing.) in 2010 at University of Bremen, Germany. Since 2010 he is with DSI GmbH, Bremen, Germany, where he works as testing and verification engineer for aerospace electronic systems. From 2008 to 2010 he was with Fraunhofer Institute for Digital Media Technology (IDMT), project group Hearing, Speech and Audio (HSA) in Oldenburg, Germany, where he worked on speech quality assessment for dereverberation algorithms.

•

Jan Rennies is head of the groups Audio Quality and Auditory Modeling as well as Personalized Hearing Systems at the Fraunhofer Institute for Digital Media Technology (IDMT), project group Hearing, Speech and Audio (HSA) in Oldenburg, Germany. He received his B.Eng. (2006) and M.Sc. (2008) in engineering physics and his Dr.rer.nat (2013) at the University of Oldenburg, Germany. During his studies at the University of Oldenburg, Denmark's Technical University, and the Technical University of Munich, he specialized in psychoacoustic perception and auditory modeling. His current research interests are subjective methodologies and psychoacoustic modeling of loudness, speech intelligibility, listening effort, sound quality, and personalized hearing support with applications in automotive, communication systems, room acoustics, signal enhancement, sound design, and quality control.

•

Emanuël A. P. Habets received his B.Sc degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and his M.Sc and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2002 and 2007, respectively. From March 2007 until February 2009, he was a Postdoctoral Fellow at the Technion - Israel Institute of Technology and at the Bar-Ilan University in Ramat-Gan, Israel. From February 2009 until November 2010, he was a Research Fellow in the Communication and Signal Processing group at Imperial College London, United Kingdom. Since November 2010, he is an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS) and a Chief Scientist at Fraunhofer IIS, Germany. His research interests center around audio and acoustic signal processing, and he has worked in particular on dereverberation, noise estimation and reduction, echo reduction, system identification and equalization, source localization and tracking, and crosstalk cancellation. Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, and a general co-chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York. He is a member of the Audio Engineering Society, a Senior Member of the IEEE, and a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. Since 2013 he is an Associate Editor of the IEEE Signal Processing Letters.

•

Karl-Dirk Kammeyer studied electrical engineering at the Technical University of Berlin. He graduated from the University of Erlangen, Erlangen, Germany, in the field of digital signal processing in 1977. Postdoctoral lecturing qualification was done for the subject communications technology at the University of Paderborn in 1985. He accepted an offer of professorship from the Technical University of Hamburg-Harburg, Germany, in 1984, and in 1995 he accepted an offer of professorship at the University of Bremen, Germany, to hold a chair for communications engineering. His main research interests include mobile communications, channel coding, adaptive receiver structures, signal processing, blind channel estimation, audio and speech processing, hands-free telephones, and videoconferencing systems. He has written three course books and more than 200 technical papers. Since 2011, he has been an emeritus professor—but he still supervises several Ph.D. students at the University of Bremen.