Model-Based Dereverberation Preserving Binaural Cues

Marco Jeub, *Student Member, IEEE*, Magnus Schäfer, Thomas Esch, *Student Member, IEEE*, and Peter Vary, *Fellow, IEEE*

***Abstract*—The ability of the human auditory system for sound localization mainly depends on the binaural cues, especially interaural time and level differences (ITD and ILD). In the context of digital hearing aids and binaural audio transmission systems, these cues can be severely degraded by independent bilateral signal processing such as dereverberation or noise reduction. This contribution presents a novel two-stage binaural dereverberation algorithm which explicitly preserves the binaural cues. The first stage is based on a statistical model of the room impulse responses (RIR) and comprises a spectral subtraction rule which reduces late reverberation only. It includes a smoothing process of the spectral gains to reduce musical tones. In a second stage, the residual reverberation is attenuated by a dual-channel Wiener filter. This is derived from a coherence model of the reverberant sound field taking into account shadowing effects of the head. The overall binaural-input binaural-output structure efficiently reduces both early and late reverberation. In experiments as well as informal listening tests using measured binaural room impulse responses, the proposed algorithm significantly improves speech quality according to objective and subjective measures.**

***Index Terms*—Binaural cue preservation, dereverberation, head shadowing, speech enhancement, spectral subtraction.**

I. Introduction

In speech communication systems, room reverberation often leads to a degradation of speech quality and intelligibility. This especially applies for hands-free devices, binaural telephone headsets, and digital hearing aids. The effects of room reverberation can be categorized into two distinct perceptual components: overlap-masking and coloration. Late reverberation causes mainly overlap-masking effects, whereas the early reflections are known to cause a coloration of the anechoic speech signal.

Many contributions have been made in the past to reduce the effects due to reverberation, cf. [1]–[4]. Since a joint suppression of both early *and* late reverberation is quite challenging, several (single- and multichannel) two-stage algorithms are proposed in the literature. The authors in [5] present an inverse filtering algorithm which maximizes the kurtosis of the linear prediction (LP) residual signal for the reduction of early reverberation, followed by a spectral subtraction rule that reduces long-term reverberation. A similar approach is described in [6] where spatio-temporal averaging is combined with a spectral subtraction algorithm.

The major drawback is that most of these techniques were developed for systems with a single output channel given one or possibly multiple input channels. Therefore, they are only suitable for bilateral processing, which means that each side of the device is independently performing monaural enhancement without taking spatial information into account.

Several studies have shown that unsynchronized processing degrades the ability for sound localization and that hearing impaired persons localize sounds better without their independent bilateral hearing aids than with them; see [7] and the references therein. This can be explained by the fact that the binaural cues, which are the basis for human sound localization, are not preserved. This comprises mostly the interaural level difference (ILD) and interaural time difference (ITD), cf. [8].

Therefore, it is advantageous to perform binaural instead of bilateral processing, especially as an appropriate data link between both sides of the headset and in future between both sides of the hearing aid can be assumed, cf. [9]. An extension of monaural algorithms to a binaural output is not always trivial. A discussion of fixed and adaptive beamforming with binaural output can be found, e.g., in [10] and [11]. The authors in [12] propose a binaural noise reduction system consisting of a binaural superdirective beamformer and a postfilter. A comprehensive study how binaural noise reduction algorithms can preserve binaural cues can be found in [13]. A binaural blind source separation (BSS) strategy is proposed, e.g., in [14]. The problem in terms of binaural dereverberation is addressed, e.g., in [15]–[17]. However, no dereverberation system exists so far, which allows for a reduction of both early and late reverberation and explicitly preserves the binaural cues.

This contribution consists of two major parts. First, in Section II, the necessity of binaural instead of bilateral processing is studied. It will be shown, exemplarily with three known dereverberation algorithms, how bilateral signal processing affects the source localization. This part also comprises an improvement for ILD and ITD estimation in reverberant environments. In the second part of this paper, a novel two-stage binaural dereverberation system is proposed in Section III that does not alter the binaural cues. In the first stage, a spectral subtraction rule is applied which is based on an estimate of the late reverberant energy including a smoothing process of the spectral gains [18], [19]. This is derived by a statistical model

Manuscript received November 13, 2009; revised May 11, 2010. Date of current version August 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

The authors are with the Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, 52056 Aachen, Germany (e-mail: jeub@ind.rwth-aachen.de; schaefer@ind.rwth-aachen.de; esch@ind.rwth-aachen.de; vary@ind.rwth-aachen.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2052156

of the room impulse response (RIR). In the second stage, the residual reverberation is further attenuated by a dual-channel Wiener filter which is based on a new coherence model taking shadowing effects of the head into account [20]. The basic idea of such combination is that Stage I of the algorithm mainly reduces the late reverberant components, while the subsequent Wiener filter in Stage II attenuates all non-coherent signal components. This results in an efficient reduction of both early and late reverberation. Due to the algorithmic structure, the binaural cues are not affected.

The remainder of this paper is organized as follows. In Section IV, the shadowing effects of the human head are discussed and a model for the binaural coherence is given. Experiments and results are presented in Section V and finally, in Section VI we draw some conclusions.

II. Binaural Cues

The human auditory system has a very sophisticated mechanism to analyze the spatial impression of an acoustic environment by exploiting the binaural cues [8]. This comprises the ability for distance and direction estimation. Numerous experiments have shown that the localization in the azimuth plane is mostly based on the interaural time and level differences of the sound event. The localization in terms of elevation is carried out with the help of the spectral coloring of the input signals due to the shape of the outer ear. The distance perception is based on the direct-to-reverberant energy ratio, cf. [21].

Since they are the most important binaural cues for source localization in the azimuth plane, the main focus of this paper will be on ITD and ILD. The estimation of these cues is important for both the development and the evaluation of binaural algorithms. In reverberant sound fields, the reliability of these estimates is not always guaranteed. The binaural cues are degraded by not only taking values different from those of the non-reverberant signals, but also having larger variance, which makes the localization of the source ambiguous. Hence, an improved system will be derived which significantly decreases the variance of the estimation.

In this paper, the correct binaural cues are determined as ones that can be estimated from the reverberant signals by focusing only on the time–frequency portions that are interaurally coherent on the cues.

Throughout the remainder of this paper, source direction θ in the azimuth plane, discrete time source signal $s(k)$ and ear signals $x_r(k)$, $x_l(k)$ will be used according to Fig. 1.

A. Estimation of Binaural Cues

The interaural time difference is defined as the time delay of arrival between the left and right ear. Assuming a simple model of the head as a spherical torso and a source in the far field, the ITD can be expressed by [22]

$$\Delta t = \frac{3r}{c} \sin \theta \tag{1}$$

where r is the radius of the head and c the speed of sound. For an approximate radius of $r = 8.5$ cm and $c = 340$ m/s, the ITD lies in the range $-750\ \mu\text{s} \leq \Delta t \leq +750\ \mu\text{s}$. The interaural level difference in dB is given by the level differences of the signals arriving at the left and right ear

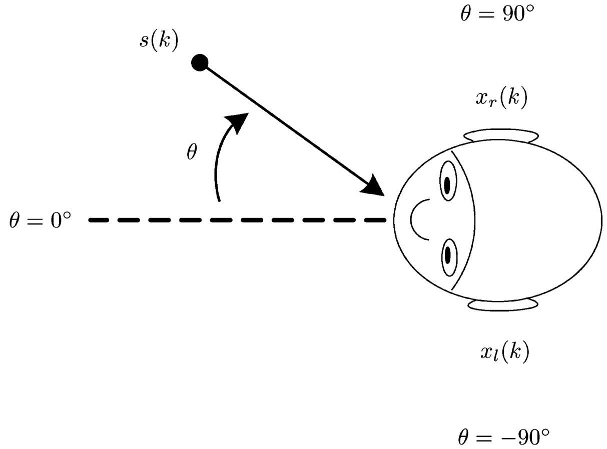


Fig. 1. Coordinate system and signal naming.

$$\Delta E = 10 \cdot \log_{10} \left(\frac{E_l}{E_r} \right) \tag{2}$$

with $E_{l\,|\,r}$ being the energy of the right and left signal $x_{l\,|\,r}(k)$, respectively. As a simple rule-of-thumb, the ITD is relevant for frequencies below and the ILD for frequencies above 1.5 kHz, cf. [8].

For situations with only one dominant source, a straightforward method for estimating the ITD is to calculate the cross-correlation and to measure the time lag of the maximum [23]; an overview on such time delay estimation techniques can be found in [24]. The ILD can simply be calculated by the energy ratio as in (2). However, for multiple sources or reverberant environments, both measures become unreliable, cf. [25].

A promising procedure to improve the estimation robustness for both, ITD and ILD, has been published in [25], where only cues are selected where the interaural coherence (IC) is above a certain threshold. By this procedure, the algorithm tries to estimate only the cues of the direct path (which correspond to the clean speech or free-field cues). This can be seen as a replication of the precedence effect in the human auditory system which mainly relies on the binaural cues of the first wave front for azimuth localization, cf. [8]. The estimation procedure for ILD cues will be described shortly in the following. The extension to ITD estimation is straightforward, the only change is a replacement of the frame-wise ILD by a frame-wise ITD.

The input signals of both channels are first divided into frames of 20 ms (320 samples at a sampling rate of 16 kHz) with an overlap of 319 samples to allow for a detailed and precise analysis. These frames are then decomposed into 24 critical bands using a Gammatone cochlear filterbank [26], [27]. The center frequencies are chosen according to the Glasberg and Moore model [28].

For each band with subband index μ ($\mu = 1, 2, \ldots, 24$) and corresponding center frequency f_c, the estimation is performed by means of recursive averaging. The ILD for each frame is calculated as

$$\Delta E(\lambda,\mu) = 10 \cdot \log_{10}\left(\frac{E_l(\lambda,\mu)}{E_r(\lambda,\mu)}\right) \tag{3}$$

where λ marks the frame index. The energies of left and right channel are calculated by the recursive averages

$$E_l(\lambda,\mu) = \alpha_1 \cdot \sum_{k=1}^{K} x_l^2(k) + (1-\alpha_1) \cdot E_l(\lambda-1,\mu) \tag{4a}$$

$$E_r(\lambda,\mu) = \alpha_1 \cdot \sum_{k=1}^{K} x_r^2(k) + (1-\alpha_1) \cdot E_r(\lambda-1,\mu) \tag{4b}$$

with K being the number of samples in each frame of 20-ms duration. The smoothing factor α_1 is determined from the time constant $T = 10$ ms and sampling frequency f_s in Hz as in [25]

$$\alpha_1 = \frac{1}{T \cdot f_s}. \tag{5}$$

Since the per-frame ILD estimate $\Delta E(\lambda,\mu)$ gives unreliable results, especially in reverberant environments, the variance of the estimate has to be decreased. One very attractive possibility is to select only cues with an interaural coherence (IC) above a certain threshold for further evaluation. The IC is estimated by the normalized cross-correlation given by

$$\gamma(\lambda,\mu) = \frac{E_{lr}(\lambda,\mu)}{\sqrt{E_l(\lambda,\mu) \cdot E_r(\lambda,\mu)}} \tag{6}$$

where $E_{lr}(\lambda,\mu)$ is calculated by

$$E_{lr}(\lambda,\mu) = \alpha_1 \cdot \sum_{k=1}^{K} (x_l(k) \cdot x_r(k)) + (1-\alpha_1) \cdot E_{lr}(\lambda-1,\mu). \tag{7}$$

In the following, only cues with an IC above the threshold $\gamma_{\mathrm{thr}}(\mu)$ are used:

$$\Delta E_{\mathrm{sel}}(\lambda,\mu) = \{\Delta E(\lambda,\mu) \,|\, \gamma(\lambda,\mu) > \gamma_{\mathrm{thr}}(\mu)\}. \tag{8}$$

The choice of $\gamma_{\mathrm{thr}}(\mu)$ has a strong influence on the estimation. If $\gamma_{\mathrm{thr}}(\mu)$ is chosen too low, the cue selection process will be rendered inefficient as no significant reduction in variance can be achieved. On the other hand, if $\gamma_{\mathrm{thr}}(\mu)$ is chosen too high, the reliability of the selection will be decreased as just very few signal frames will be considered for the determination of $\Delta E_{\mathrm{sel}}(\lambda,\mu)$. In terms of reverberant signals, the necessity for different thresholds per frequency band is motivated by the high frequency-dependence of the reverberation tail, cf. [29]. For the sake of brevity, we omit the index μ for the threshold in the following. In [25], a fixed threshold was given depending on the center frequency f_c of the frequency band:

- $f_c = 500$ Hz ⇒ $\gamma_{\mathrm{thr}} = 0.95$;
- $f_c = 2000$ Hz ⇒ $\gamma_{\mathrm{thr}} = 0.99$.

However, the optimum threshold is not only depending on the center frequency but also on the azimuth angle of the source signal. In Fig. 2, the normalized histograms of the cue selection according to (8) are depicted for a center frequency of $f_c = 2584$ Hz (corresponds to subband $\mu = 17$) and for the two azimuth angles $\theta = 0°$ and $\theta = 30°$. The reverberant speech is generated using eight speech files from the NTT database [30] convolved with binaural room impulse responses (BRIR) measured with a dummy head in a stairway hall (see the Appendix). The vertical lines mark the correct values for this frequency band in terms of the anechoic cues. These have been derived by convolving the same source signal with the (manually segmented) direct path of the corresponding BRIR.

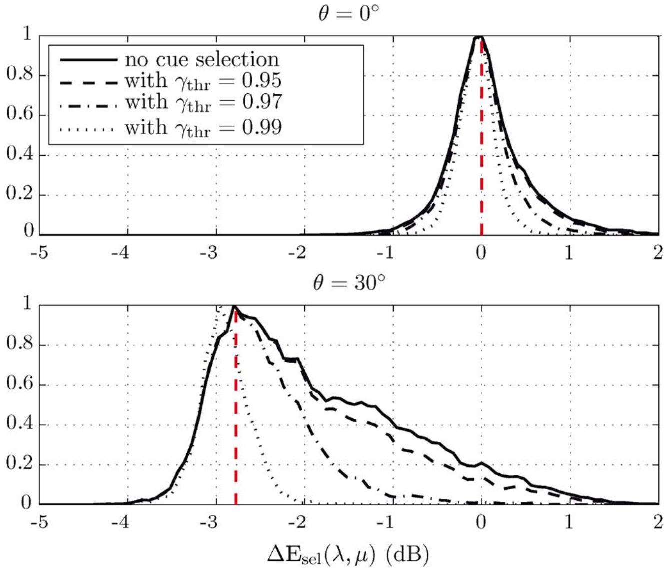


Fig. 2. ILD estimation: normalized histograms of different cue selection procedures for a speech source from two different azimuth angles (without and with cue selection and three different thresholds γ_{thr} at center frequency $f_c = 2584$ Hz). The signals are generated using BRIRs of a stairway hall having an average reverberation time of $\mathrm{T}_{60} = 0.69$ s.

For an azimuth angle of $\theta = 0°$, the optimum result, calculated from the direct speech signals, for the ILD estimation would be 0 dB while it would be -2.78 dB for $\theta = 30°$. The estimated ILD is shown for four different cue selection conditions:

- without any selection of cues ($\gamma_{\mathrm{thr}} = 0$);
- with $\gamma_{\mathrm{thr}} = 0.95$;
- with $\gamma_{\mathrm{thr}} = 0.97$;
- with $\gamma_{\mathrm{thr}} = 0.99$.

It can be seen that for an angle of $\theta = 0°$, the results do not differ significantly between the cue selection strategies. Since $\gamma_{\mathrm{thr}} = 0.99$ leads to the smallest variance in the estimation, it would be the threshold of choice for this case. However, the situation changes for an angle of $\theta = 30°$ where the variance without any selection procedure is quite large and even a threshold of $\gamma_{\mathrm{thr}} = 0.95$ does not lead to a substantial decrease in variance. A threshold of $\gamma_{\mathrm{thr}} = 0.99$ on the other hand leads to another issue: the mean of the histogram deviates from the correct value. The reason for this direction-dependent behavior lies in the strong variation of the direct-to-reverberant energy ratio (DRR) for different azimuth angles as depicted in Fig. 3. A decrease in DRR leads to a stronger impact of diffuse components on the estimation of ILD and ITD.

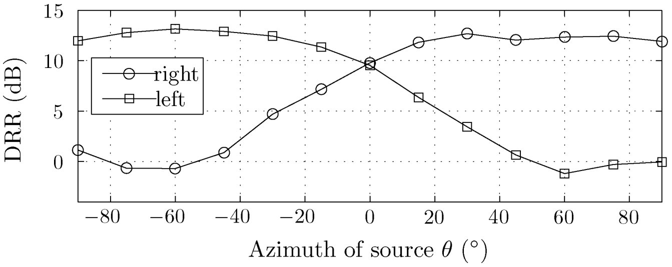


Fig. 3. Direct-to-reverberant energy ratio (DRR) measured in a stairway hall ($\mathrm{T}_{60} = 0.69$ s) at different azimuth angles in the presence of a dummy head.

To allow for a threshold that is better suited to all frequency bands and azimuth angles, we propose an adaptive procedure that makes use of the signal statistics for the determination of $\gamma_{\mathrm{thr}}(\mu)$ from $\gamma(\lambda,\mu)$. The threshold is calculated as the 90th percentile of all individual values $\gamma(\lambda,\mu)$. This procedure guarantees that the ILD is always estimated from the most reliable values (i.e., the 10% of all individual values with the highest IC) for $\Delta E_{\mathrm{sel}}(\lambda)$ while ensuring that single outliers never get too much weight in the calculation. For the cases that are depicted in Fig. 2, this procedure leads to a threshold $\gamma_{\mathrm{thr}} = 0.99$ for $\theta = 0°$ and $\gamma_{\mathrm{thr}} = 0.98$ for $\theta = 30°$. Similar properties can be observed for the other subbands μ.

The improved estimator ensures a significant reduction in variance for the ILD estimate leading to more reliable results in accordance with the precedence effect in the human auditory system. Hence, it will now be used to investigate the influence of a bilateral dereverberation on the binaural cues.

B. Influence of Bilateral Dereverberation

In this subsection, we investigate the influence of bilateral, i.e., independent dereverberation on the binaural cues. The objective of any enhancement algorithm with respect to the binaural cues should be to preserve the absolute mean values while ensuring very low fluctuations from frame to frame. The minimum audible changes for ILD (0.5 dB) and ITD (10 μs) should not be exceeded, cf. [22]. Several studies have shown that the human auditory system is capable of relearning the spatial information when receiving altered binaural cues. However, since the adaptation typically takes a week or more, an adjustment to rapid changes is impossible [13], [31]. Exemplarily, we restrict the analysis to changes of ILD. An extension to ITD can be done by switching to an ITD estimator instead of the ILD estimator.

The basis for this investigation lies in the observation that the DRR is highly dependent on the azimuth angle as shown in Fig. 3. To quantify the impact of independent dereverberation of both channels on the binaural cues, three dereverberation algorithms will be used for an independent single-channel enhancement. The first dereverberation algorithm (Spatiotemporal avg.) averages the linear prediction residual signal between consecutive cycles of opening and closing of the glottis (larynx cycle) while excluding the segments around the glottal closure instances [32]. This reduces unwanted peaks in the residual signal which are caused by reverberation. The second algorithm is our previous proposal of the postfilter known from code excited linear prediction (CELP) speech coding for dereverberation [33]. The third algorithm is based on a spectral subtraction technique to remove the late reverberant components of the degraded input signal [18].

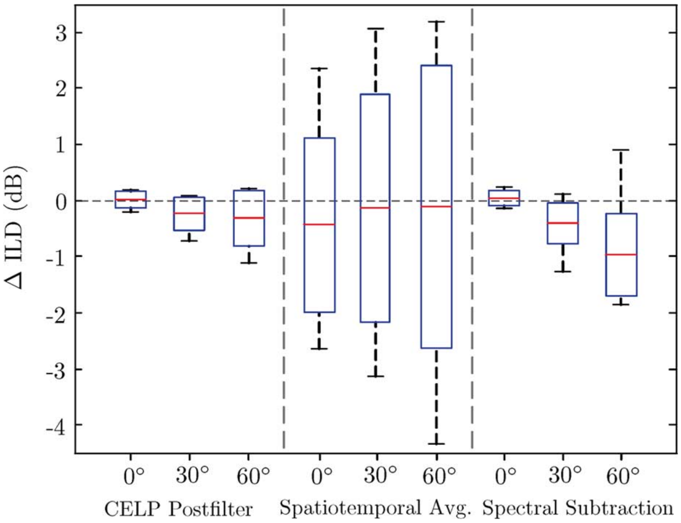


Fig. 4. Differences in ILD estimation compared to reverberant speech (stairway hall, $\mathrm{T}_{60} = 0.69$ s) in dB. The results are equally weighted for all frequency bands above 1.5 kHz.

The influence on the binaural cues will be investigated as follows. From the dual-channel input signals, the binaural cues are estimated before processing. Afterwards, a bilateral dereverberation (independently without any data-link between left and right processing module or synchronization) is performed with the described algorithms. Finally, the binaural cues are estimated again and compared to the cues before processing. For all binaural cue estimation tasks, the cue selection procedure of (8) is used with an adaptive threshold. The evaluation is carried out with speech files from the NTT database that are convolved with BRIRs measured in a stairway hall at different azimuth angles, all in the presence of a dummy head.

The investigation focuses on frame-by-frame fluctuations of the ILD cues, which is an important issue since most algorithms perform enhancement of short speech frames which causes a different degree of enhancement per frame. Therefore, we calculate the ILD framewise for each of the 24 frequency-bands over all frames to measure the variance in ILD estimation. Finally, the results are averaged over all bands above 1.5 kHz.

The results for three different azimuth angles are depicted in Fig. 4. The boxes represent the variance from the mean value (horizontal line inside the box) and the end of the whiskers represent minimum and maximum of the ILD difference compared to the reverberant speech. The corresponding dereverberation performance is listed in Table I (see the experiment Section V for a detailed description of the SRMR measure). It can be seen from Fig. 4 that all tested algorithms cause high variations in the binaural cues as shown exemplarily here for the ILD and hence, influence the source localization.

All algorithms show the lowest influence for the frontal direction (0°) and distort the cues especially for sources from aside. The most significant increase in ILD fluctuations occurs for the spatiotemporal avg. algorithm. Even though the CELP

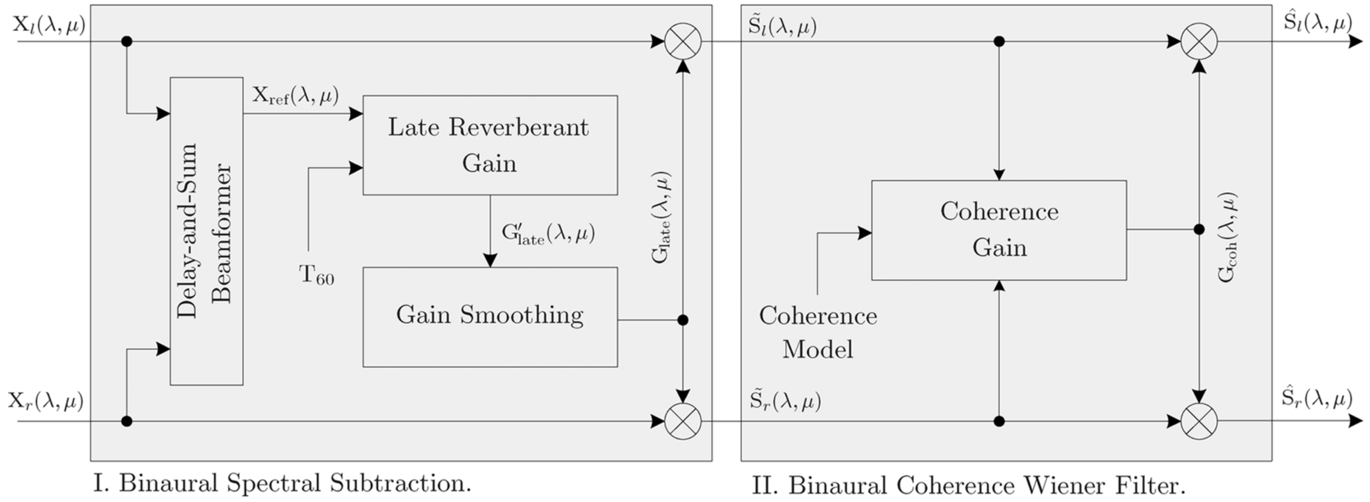

$X_l(\lambda,\mu)$
$X_{\mathrm{ref}}(\lambda,\mu)$
Late Reverberant Gain
Delay-and-Sum Beamformer
$G'_{\mathrm{late}}(\lambda,\mu)$
T_{60}
Gain Smoothing
$G_{\mathrm{late}}(\lambda,\mu)$
$X_r(\lambda,\mu)$
$\tilde{S}_l(\lambda,\mu)$
$\hat{S}_l(\lambda,\mu)$
Coherence Gain
$G_{\mathrm{coh}}(\lambda,\mu)$
Coherence Model
$\tilde{S}_r(\lambda,\mu)$
$\hat{S}_r(\lambda,\mu)$
I. Binaural Spectral Subtraction.
II. Binaural Coherence Wiener Filter.

Fig. 5. Schematic diagram of the proposed two-stage binaural cue preserving dereverberation algorithm.

TABLE I
DEREVERBERATION PERFORMANCE IN TERMS OF SRMR IMPROVEMENT, AVERAGED OVER LEFT AND RIGHT CHANNEL. PLUS INDICATES IMPROVEMENT

	0°	30°	60°
CELP Postfilter [33]	$+0.38$	$+0.32$	$+0.37$
Spatiotemporal Avg. [32]	$+0.69$	$+0.63$	$+0.49$
Spectral Subtraction [18]	$+1.38$	$+1.39$	$+1.60$

postfilter leads to the smallest variations among the tested approaches, moderate changes in ILD are still audible, even for frontal sources.

In terms of the dereverberation performance (see Table I), the spectral subtraction technique shows the highest amount of reverberation reduction and is further used in our two-stage dereverberation system in a binaural configuration (see Section III-B).

Since all algorithms exhibit changes in binaural ILD cues that are mostly above the minimum audible difference, we can conclude that bilateral dereverberation has a clearly perceivable influence on the source localization. A listening test with 17 participants confirmed this assumption and also assessed the overall quality of the different dereverberation schemes as shown later in Section V-B. It has to be mentioned that the spectral subtraction is performed in the frequency domain where the phase of the input signal is kept. Hence, no modification of the ITD cue results for this specific algorithm.

In order to ensure unaffected source localization, it is important to preserve the binaural cues to a certain extent. A preservation of the binaural cues can be ensured basically in two different ways. A first method would be to reconstruct the binaural cues after the processing using a binaural postfilter. Another method, which is considered here, is to incorporate the cue preservation into the processing algorithm. An overview about binaural reproduction suitable for the application to blind source separation (BSS) can be found in [34]. The problem of binaural cue preservation in the context of binaural artificial bandwidth extension (BWE) has been addressed, e.g., in [35].

III. BINAURAL CUE PRESERVING DEREVERBERATION

A. Binaural Dereverberation Concept

The proposed model-based dereverberation concept consists of two independent components as depicted in Fig. 5. The considered algorithms are realized by short-term spectral weighting using the weighted overlap–add method [36]. For the transformation into the frequency domain, the disturbed input signals $x_{l\,|\,r}(k)$ are first segmented into overlapping frames of length L. After windowing (e.g., applying a Hann window), these frames are transformed via fast Fourier transform (FFT) of length M into the short-term spectral domain. At discrete frequency bins μ, the distorted signals for right and left channel are $X_{l\,|\,r}(\lambda,\mu)$. The enhanced spectra $\hat{S}_{l\,|\,r}(\lambda,\mu)$ can be obtained by multiplying the coefficients $X_{l\,|\,r}(\lambda,\mu)$ with the weighting gains $G_{\mathrm{late}}(\lambda,\mu)$ and $G_{\mathrm{coh}}(\lambda,\mu)$ of the two stages. The enhanced time domain signals $\hat{s}_{l\,|\,r}(k)$ are obtained by using the inverse fast Fourier transform (IFFT) and overlap-add.

Applying different weighting gains to each channel can cause unwanted modifications in the spatial impression as described in Section II-B and the listening test in Section V-B. Therefore, the same weighting gains are applied to each channel and hence, the ILD is unaffected. In order to ensure an unaffected interaural phase difference (which is used by the human auditory system for the determination of the ITD), the phase of the disturbed input signals is kept. Additionally, each channel shows the same algorithmic delay. This concept is also used in binaural noise reduction algorithms, cf. [12], [15], [17].

In the following, the calculation of the weighting gains is derived successively. The cascade of the two stages is mainly motivated by the fact that each stage requires different properties for the input signal in terms of the DRR. The first stage comprises an estimation of the late reverberant energy, where the underlying statistical model of the RIR requires a low DRR [18]. After the first processing step, the DRR increases since late reverberation is attenuated while keeping the direct and early speech component unaffected. This first stage does not influence the coherence between both channels since the same spectral weights are applied to both channels. The second stage estimates the (direct) speech power spectral density (PSD), which requires a high

DRR in order to reduce estimation errors. Thus, it is beneficial to increase the DRR in a previous step. Since the second stage attenuates all non-coherent parts, and hence early and late reverberation, a great DRR increase can be expected as well. Consequently, a reversed order of the two stages would be less effective, as also confirmed by our experiments.

B. Stage I: Dereverberation Based on a Statistical Model of Late Reverberation

An efficient dereverberation algorithm based on a statistical model of late reverberation has been proposed first in [1] and was later refined in [18]. The basic idea is to estimate the variance of the late reverberant speech components and to formulate a weighting rule that aims to suppress late reverberant components while leaving the direct path and early reflections unaltered. This subsection describes the original single-channel algorithm and discusses the extension to binaural outputs.

A representation of a room impulse response $h(k)$ of length T_r (in s) can be divided into its direct and early as well as its late components by

$$h(k) = \begin{cases} 0, & \text{for } k < 0 \\ h_{\text{early}}(k), & \text{for } 0 \leq k < \mathrm{T}_l \cdot f_s \\ h_{\text{late}}(k), & \text{for } \mathrm{T}_l \cdot f_s \leq k \leq \mathrm{T}_r \cdot f_s \end{cases} \tag{9}$$

where $h_{\text{early}}(k)$ refers to the direct and early path, $h_{\text{late}}(k)$ to the late path, and T_l marks the time span after which the late reverberation begins. The range of T_l usually lies in the range of 50–100 ms, cf. [37].

The reverberant signal $x(k)$ can now be decomposed into its early and late reverberant speech components $x_{\text{early}}(k)$ and $x_{\text{late}}(k)$ by

$$x(k) = \underbrace{\sum_{n=0}^{\mathrm{T}_l f_s - 1} s(k-n)h(n)}_{x_{\text{early}}(k)} + \underbrace{\sum_{n=\mathrm{T}_l f_s}^{\mathrm{T}_r f_s} s(k-n)h(n)}_{x_{\text{late}}(k)} \tag{10}$$

where the corresponding DFT spectra are named $X_{\text{early}}(\lambda, \mu)$ and $X_{\text{late}}(\lambda, \mu)$, respectively.

An estimate for the variances of the late reverberant speech can be obtained by means of a simple statistical model for the room impulse response (RIR) [18]

$$\hat{h}_{\text{late}}(k) = \begin{cases} n(k)\, e^{-\rho k f_s^{-1}}, & \text{for } k \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $n(k)$ is a sequence of i.i.d. random variables with zero mean and normal distribution. The decay rate ρ is linked to the reverberation time T_{60} by

$$\rho = \frac{3\ln(10)}{\mathrm{T}_{60}}. \tag{12}$$

Based on (11) it can be shown that the late reverberant component $x_{\text{late}}(k)$ (or $X_{\text{late}}(\lambda, \mu)$) can be modeled as an uncorrelated noise process [18]. It has to be mentioned that this model is valid only when the direct path energy is smaller than the energy of all reflections (low DRR), cf. [3].

An estimator for the variance of the late reverberant speech is given by

$$\sigma^2_{x_{\text{late}}}(\lambda, \mu) = e^{-2\rho \mathrm{T}_l} \cdot \sigma^2_x(\lambda - N_l, \mu) \tag{13}$$

with the variance $\sigma^2_x(\lambda, \mu)$ of the reverberant speech and N_l the number of frames corresponding to T_l. In order to estimate the *a posteriori* signal-to-interference ratio (SIR)

$$\eta(\lambda, \mu) = \frac{|X(\lambda, \mu)|^2}{\sigma^2_{x_{\text{late}}}(\lambda, \mu)} \tag{14}$$

the spectral variance of the reverberant speech is calculated by recursive averaging

$$\sigma^2_x(\lambda, \mu) = \alpha_2 \cdot \sigma^2_x(\lambda - 1, \mu) + (1 - \alpha_2) \cdot |X(\lambda, \mu)|^2 \tag{15}$$

with a smoothing factor $0 \leq \alpha_2 \leq 1$. The weights for the suppression of the late reverberant components are calculated by the spectral magnitude subtraction rule

$$G'_{\text{late}}(\lambda, \mu) = 1 - \frac{1}{\sqrt{\eta(\lambda, \mu)}}. \tag{16}$$

Additionally, a lower bound $G^{\text{late}}_{\text{min}}$ is applied to all weighting gains to counter overestimation of $\sigma^2_{x_{\text{late}}}(\lambda, \mu)$.

Robust estimation of the reverberation time T_{60} is quite challenging and not considered in this paper, cf. [38]. Here, we use a T_{60} estimate obtained directly from the RIR by applying the Schroeder integral [39].

In the next step, the presented single-channel algorithm is extended into a binaural-input binaural-output algorithm. A delay-and-sum beamformer (DSB) is used to generate a reference signal which has the advantage of a low computational complexity. The reference signal is calculated from the average of both time-aligned signals according to

$$X_{\text{ref}}(\lambda, \mu) = \frac{1}{2} \cdot (X'_l(\lambda, \mu) + X'_r(\lambda, \mu)). \tag{17}$$

The estimation of the time delays in the beamformer is performed by means of the generalized cross-correlation with phase transform (GCC-PHAT) as described in [40]. From this reference signal, the spectral variance of the late reverberations as well as the weighting gains $G'_{\text{late}}(\lambda, \mu)$ are computed by the previously described spectral subtraction rule.

It has to be mentioned that the use of the DSB itself performs already a reduction of reverberation. Therefore, the resulting reverberation time and hence, the estimated variance of the late reverberations is only an approximation of the estimates directly from the input signals. Since the DSB provides only a small amount of reverberation reduction and as a small variation in the estimated reverberation time is not critical [4], this approximation is still feasible. For reducing the amount of musical tones of the spectral subtraction approach of (16), spectral smoothing of the magnitudes $G'_{\text{late}}(\lambda, \mu)$ is performed [19]. The main idea is to reduce the annoying musical tones especially in low signal-to-interference ratio (SIR) regions requiring a reliable and robust detector. In order to obtain a good indication whether a frame contains speech or not, the power ratio between the enhanced reference signal $X_{\text{ref}}(\lambda, \mu) \cdot G'_{\text{late}}(\lambda, \mu)$

and the reference signal $X_{\mathrm{ref}}(\lambda,\mu)$ is calculated for each frame λ as follows:

$$\zeta(\lambda)=\frac{\sum_{\mu=0}^{M-1}|G'_{\mathrm{late}}(\lambda,\mu)\cdot X_{\mathrm{ref}}(\lambda,\mu)|^2}{\sum_{\mu=0}^{M-1}|X_{\mathrm{ref}}(\lambda,\mu)|^2}. \tag{18}$$

If the frame mainly contains anechoic speech (high SIR), the power of the processed frame is equal or only slightly lower to the power of the input frame, i.e., $\zeta(\lambda)\approx 1$. By contrast, the speech enhancement system is supposed to strongly attenuate the input signal in low SIR conditions, resulting in a power ratio $\zeta(\lambda)\approx 0$. Based on $\zeta(\lambda)$, the magnitudes of the weighting gains $G'_{\mathrm{late}}(\lambda,\mu)$ of frame λ are adaptively smoothed over frequency μ using a moving average window. The odd window length $N_s(\lambda)$ is set to

$$N_s(\lambda)=\begin{cases}1, & \text{if } \zeta(\lambda)\ge\zeta_{\mathrm{thr}}(\lambda)\\ 2\cdot\mathrm{round}\left[\left(1-\frac{\zeta(\lambda)}{\zeta_{\mathrm{thr}}}\right)\cdot\Psi\right]+1, & \text{else.}\end{cases} \tag{19}$$

In order to detect only low SIR regions, a threshold ζ_{thr} is required that controls the trade-off between speech distortions and musical noise reduction. The term $1-(\zeta(\lambda)/\zeta_{\mathrm{thr}})$ provides a soft-decision that states the reliability of the low SIR detection. The function round$[\cdot]$ rounds the element to the nearest integer and Ψ is a scaling factor that determines the maximum degree of smoothing. Equation (19) ensures that the more reliable a low SIR frame was detected, the longer the window length resulting in stronger smoothing of the weighting gains. Applying a moving average window of length $N_s(\lambda)$ is equivalent to a linear filtering with the impulse response $H_s(\lambda,\mu)$ as follows:

$$H_s(\lambda,\mu)=\begin{cases}\frac{1}{N_s(\lambda)}, & \text{if } \mu<N_s(\lambda)\\ 0, & \text{else}\end{cases} \tag{20}$$

where $\mu\in\{0,1,\ldots,M-1\}$. Within the smoothing procedure, the weighting gain magnitudes are convoluted over frequency μ by the low-pass filter $H_s(\lambda,\mu)$ in every frame λ

$$G_{\mathrm{late}}(\lambda,\mu)=G'_{\mathrm{late}}(\lambda,\mu)*H_s(\lambda,\mu). \tag{21}$$

Finally, the smoothed weighting gains $G_{\mathrm{late}}(\lambda,\mu)$ are applied to the disturbed input spectra by

$$\tilde{S}_l(\lambda,\mu)=X_l(\lambda,\mu)\cdot G_{\mathrm{late}}(\lambda,\mu) \tag{22a}$$
$$\tilde{S}_r(\lambda,\mu)=X_r(\lambda,\mu)\cdot G_{\mathrm{late}}(\lambda,\mu). \tag{22b}$$

C. Stage II: Dereverberation Based on Sound Field Coherence

The motivation for a second processing step is that the spectral subtraction rule described in the previous Section III-B aims at reducing late reverberation only and hence, residual reverberation remains. The subsequent coherence-based dereverberation algorithm exploits the low coherence of the sound field between different microphones to estimate the (direct) speech PSD and to remove all non-coherent signal parts while keeping the coherent parts unaffected. Since only the direct speech shows a high coherence among sensors as shown later, this approach also reduces early reverberation. A further advantage is that no estimation of room acoustic parameters (e.g., T_{60}) is required and that *a priori* information about the sound field can significantly improve the effectiveness of the algorithm. For the sake of clarity, we first describe the background of the method in terms of two general microphone signals $x_{1|2}(k)$. When it comes to the application in the two-stage system, all calculations are performed on the pre-dereverberated signal spectra $\tilde{S}_{l\,|\,r}(\lambda,\mu)$.

For the derivation of this method, it is assumed that the source–microphone distance is smaller than the critical distance. Therefore, the speech signals captured by the two microphones $x_{1|2}(k)$ are mutually correlated, i.e., the magnitude squared coherence (MSC) between the two microphone signals is close to one. This assumption can be fulfilled mostly for hearing devices and close-talking telephone devices.

The coherence between the two signals $x_{1|2}(k)$ is defined as

$$\Gamma_{x_1x_2}(\Omega)=\frac{\Phi_{x_1x_2}(e^{j\Omega})}{\sqrt{\Phi_{x_1x_1}(e^{j\Omega})\cdot\Phi_{x_2x_2}(e^{j\Omega})}} \tag{23}$$

where $\Phi_{x_1x_1}(e^{j\Omega})$ and $\Phi_{x_2x_2}(e^{j\Omega})$ represent the auto-power spectral densities (APSD) of $x_1(k)$ and $x_2(k)$, respectively. The cross-power spectral density (CPSD) between $x_1(k)$ and $x_2(k)$ is denoted by $\Phi_{x_1x_2}(e^{j\Omega})$. The frequently used term magnitude squared coherence (MSC) is referred to the square of (23).

The coherence between two microphones of an ideal spherically isotropic (diffuse) sound field can be expressed as [37]

$$\Gamma^{(\mathrm{diff})}_{x_1x_2}(f)=\mathrm{sinc}\left(\frac{2\pi f d_{\mathrm{mic}}}{c}\right) \tag{24}$$

with distance d_{mic} between two omnidirectional microphones with a line-of-sight and f denoting the frequency. The sound field in a reverberant room can be approximated by a diffuse sound field, cf. [37]. This has been shown in experiments with a dummy head in reverberant rooms, e.g., in [29].

In contrast to the decomposition of the reverberant signal in (10), we will now consider a division into its direct components (<2 ms) and reverberant components (≥2 ms). For the sake of simplicity, we will give the decomposition for the monaural case only, as an extension for each of the binaural channels can be performed in the same manner. The decomposed input signal $x(k)$ can be expressed by

$$x(k)=\underbrace{\sum_{n=0}^{\mathrm{T}_d f_s-1}s(k-n)h(n)}_{x_{\mathrm{direct}}(k)}+\underbrace{\sum_{n=\mathrm{T}_d f_s}^{\mathrm{T}_r f_s}s(k-n)h(n)}_{x_{\mathrm{reverb}}(k)} \tag{25}$$

where the time span of the direct sound (including sound propagation) is given by T_d. Since we assume a high portion of direct sound, we can simply determine T_d by the global maximum of the RIR plus a few reflections (here: 2 ms).

While in Section III-B the early speech component $x_{\mathrm{early}}(k)$ was the target signal, now the direct speech component $x_{\mathrm{direct}}(k)$ is the target signal.

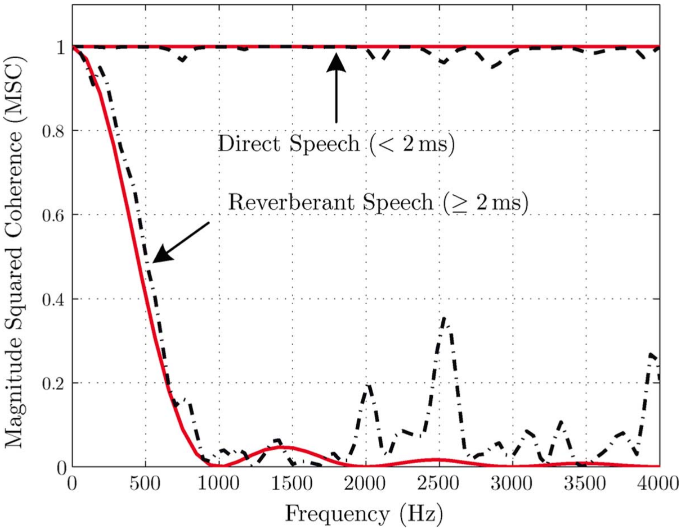


Fig. 6. MSC of direct and reverberant speech (without head). Theoretical curves (solid) and measured curves (dashed) for the parameters: $d_{\mathrm{mic}} = 0.17$ m, $\mathrm{T}_{60} = 0.72$ s.

Let us now regard the MSC of direct and reverberant speech as illustrated in Fig. 6, where the curves have been generated as follows. First, two measured room impulse responses (without dummy head) have been decomposed into direct and reverberant components. Afterwards, speech data of 8-s duration from the NTT database has been convolved with each of the RIRs resulting in separate direct and reverberant signals for each channel. Finally, the MSC between the two channels (left and right) has been calculated for both direct and reverberant speech using the Welch periodogram approach [41] and is plotted in the range 1–4 kHz.

The upper dashed line in Fig. 6 shows the MSC of the direct speech component ($\Gamma^2_{x_{\mathrm{direct},1}x_{\mathrm{direct},2}}(\Omega)$) while the lower dashed curve shows the MSC of the reverberant speech component ($\Gamma^2_{x_{\mathrm{reverb},1}x_{\mathrm{reverb},2}}(\Omega)$). The solid lines give the corresponding theoretical coherence function. As a high amount of direct speech is assumed, the theoretical coherence for the direct speech is one for all frequencies. The lower solid line gives the theoretical curve for an ideal diffuse sound field according to (24). As seen from the figure, the assumptions having made about the coherence of direct and reverberant speech are valid. Since the reverberant components received by the microphones can be represented by two additive, uncorrelated noise sources, the terms noise and reverberant components are used interchangeably in the following. As a further remark, the first stage of our dereverberation system does not cause any influence on the coherence when using $\tilde{s}_{l\,|\,r}(k)$ for the calculation. Applying identical linear filtering to both channels does not alter the coherence.

Having described the basic idea of the coherence-based dereverberation algorithm, we will now derive a dual-channel Wiener filter which takes these considerations into account, and use the notation according to Fig. 5. A common framework for speech enhancement is based on the optimal minimum mean square error (MMSE) criterion, cf. [42]. It turns out that the optimal weighting gains are given by the Wiener solution

$$G_{\mathrm{coh}}(\lambda,\mu) = \frac{\Phi_{ss}(\lambda,\mu)}{\Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu)} \tag{26}$$

where $\Phi_{ss}(\lambda,\mu)$ denotes the APSD of the original (undisturbed) signal and $\Phi_{nn}(\lambda,\mu)$ the APSD of the additive noise component. As discussed previously, the term $\Phi_{nn}(\lambda,\mu)$ is referred to the APSD of the reverberant component.

For calculating the optimal postfilter coefficients in multichannel systems, several approaches have been presented in the past. They all have in common that the estimation procedure is optimized for a specific sound field model. A well-known technique by Zelinski assumes a perfectly incoherent sound field and hence, uncorrelated noise at different sensors [43]. Since this assumption does not hold in real sound fields, an improved approach was presented by McCowan in [44] who proposed to use a model of the coherence for spherically isotropic (diffuse) sound field.

First, a brief derivation of this algorithm will be given and second, the estimation of the required power spectra is discussed. Assuming the same noise power spectrum across sensors as well as time-aligned signals, the power spectra reads

$$\Phi_{\tilde{s}_r\tilde{s}_r}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu) \tag{27}$$
$$\Phi_{\tilde{s}_l\tilde{s}_l}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Phi_{nn}(\lambda,\mu) \tag{28}$$
$$\Phi_{\tilde{s}_l\tilde{s}_r}(\lambda,\mu) = \Phi_{ss}(\lambda,\mu) + \Gamma_{\tilde{s}_l\tilde{s}_r}(\Omega)\,\Phi_{nn}(\lambda,\mu). \tag{29}$$

An estimate of the original (undistorted) signal APSD is calculated by (30), shown at the bottom of the page, [44] where the hat-operator $\{\hat{\cdot}\}$ indicates an estimate as shown later. The function $\mathrm{Re}\{\,\cdot\,\}$ returns the real part of its argument. Since the estimate of the signal APSD may not be negative or singular, a maximum threshold $\Gamma_{\max}$ for the coherence function has to be applied to ensure that $1 - \mathrm{Re}\{\Gamma_{\tilde{s}_l\tilde{s}_r}(\Omega)\} > 0$ holds for the denominator. The resulting spectral weights of the Wiener filter can now be calculated by

$$G_{\mathrm{coh}}(\lambda,\mu) = \frac{\hat{\Phi}_{ss}(\lambda,\mu)}{\frac{1}{2}\cdot\left(\hat{\Phi}_{\tilde{s}_l\tilde{s}_l}(\lambda,\mu) + \hat{\Phi}_{\tilde{s}_r\tilde{s}_r}(\lambda,\mu)\right)}. \tag{31}$$

The spectral weights are further confined by a lower threshold $G^{\mathrm{coh}}_{\min}$ for robustness against overestimation errors and to control

$$\hat{\Phi}_{ss}(\lambda,\mu) = \frac{\mathrm{Re}\{\hat{\Phi}_{\tilde{s}_l\tilde{s}_r}(\lambda,\mu)\} - \frac{1}{2}\mathrm{Re}\{\Gamma_{\tilde{s}_l\tilde{s}_r}(\Omega)\}\left(\hat{\Phi}_{\tilde{s}_l\tilde{s}_l}(\lambda,\mu) + \hat{\Phi}_{\tilde{s}_r\tilde{s}_r}(\lambda,\mu)\right)}{1 - \mathrm{Re}\{\Gamma_{\tilde{s}_l\tilde{s}_r}(\Omega)\}} \tag{30}$$

the amount by which reverberation is attenuated. The spectral weights are applied to each of the two channels by

$$\hat{S}_l(\lambda,\mu) = \tilde{S}_l(\lambda,\mu) \cdot G_{\mathrm{coh}}(\lambda,\mu) \tag{32a}$$
$$\hat{S}_r(\lambda,\mu) = \tilde{S}_r(\lambda,\mu) \cdot G_{\mathrm{coh}}(\lambda,\mu). \tag{32b}$$

The calculation of the weighting gains $G_{\mathrm{coh}}(\lambda,\mu)$ comprises an estimation of the APSDs $\Phi_{\tilde{s}_l\tilde{s}_l}(\lambda,\mu)$, $\Phi_{\tilde{s}_r\tilde{s}_r}(\lambda,\mu)$ and CPSD $\Phi_{\tilde{s}_l\tilde{s}_r}(\lambda,\mu)$ of the two time-aligned input channels. This is performed by means of an recursive periodogram approach according to

$$\begin{aligned}\hat{\Phi}_{\tilde{s}_l\tilde{s}_l|\tilde{s}_r\tilde{s}_r}(\lambda,\mu) &= \alpha_3\hat{\Phi}_{\tilde{s}_l\tilde{s}_l|\tilde{s}_r\tilde{s}_r}(\lambda-1,\mu)\\ &\quad + (1-\alpha_3)|\tilde{S}_{l\,|\,r}(\lambda,\mu)|^2 \end{aligned} \tag{33}$$
$$\begin{aligned}\hat{\Phi}_{\tilde{s}_l\tilde{s}_r}(\lambda,\mu) &= \alpha_3\hat{\Phi}_{\tilde{s}_l\tilde{s}_r}(\lambda-1,\mu)\\ &\quad + (1-\alpha_3)\tilde{S}_l(\lambda,\mu)\cdot\tilde{S}_r^*(\lambda,\mu) \end{aligned} \tag{34}$$

with smoothing factor $0 \leq \alpha_3 \leq 1$.

As mentioned before, the derivation of this dual-channel Wiener filter assumes time-aligned signals at both sensors. The weighting gains $G_{\mathrm{coh}}(\lambda,\mu)$ are calculated on the time-aligned signals (again, using GCC-PHAT for time delay estimation) and applied to the non time-aligned spectra $\tilde{S}_{l\,|\,r}(\lambda,\mu)$. By this, we ensure that the algorithm works effectively for different azimuth angles as shown later. Since the maximum time difference between the right and left channel is limited by the head geometry (see (1)), the maximum ITD range of ±750 μ s is small compared to a typical frame length of 10–30 ms. For the sake of brevity, we omitted an extra index for the time-aligned signals.

The crucial point is now to select a suitable model for the sound field coherence in (30). For an ideal diffuse sound field with a line-of-sight between two microphones, the optimal solution is the model in (24). However, when it comes to binaural signal processing where no line-of-sight between the microphones can be assumed, this model is not appropriate. Since the head-shadowing has a severe impact on the coherence, we propose to use the coherence model for a binaural sound field as described in the following Section IV.

IV. Binaural Coherence Model

It is well-known that the coherence between two microphones changes [compared to (24)] when an object is in the line-of-sight. This has been shown theoretically and in experiments on measured data with a dummy head in a crowded cafeteria in [45]. Investigations in reverberant rooms have recently been published in [29].

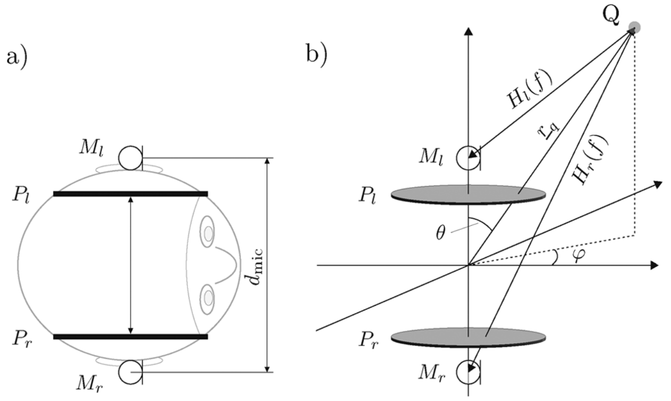


Fig. 7. Simplified geometrical model of the human head. (a) Head with the two plates. (b) Geometrical model.

This section describes an improved coherence model for a diffuse sound field compared to (24) which takes the shadowing effect of the head into account. The resulting model is then used in the coherence-based dereverberation algorithm (30). In order to describe the complex geometry of the human head, a simplified configuration with two circular plates (P_l and P_r) as depicted in Fig. 7 will be assumed in the following [45], [46]. The corresponding coherence of the sound field can now be calculated by integration over all azimuth and elevation angles (θ, φ) according to (35), shown at the bottom of the page.

Here, $\{\cdot\}^*$ denotes the complex conjugate and H_l and H_r represent the transfer functions between a punctual sound source at the position $\underline{r}_q$ and the two microphones M_l and M_r. It is also assumed that the distance of the sound source is large compared to the microphone distance (far-field assumption). A solution of this equation invokes the use of the Helmholtz–Kirchhoff Integral theorem and is described in greater detail in [45]. Since this requires the calculation of the integrals over every angle θ and φ, a simple curve-fitting is proposed as an efficient alternative. Based on the sum of Gaussians, an approximation of the sound field coherence can be expressed by

$$\hat{\Gamma}_{x_lx_r}^{(\mathrm{head})}(f) = \sum_{p=1}^{P} a_p \cdot \exp\left(-\left(\frac{f-b_p}{c_p}\right)^2\right) \tag{36}$$

with coefficients a_p, b_p, c_p, and the model order P. Since a natural ear spacing of $d_{\mathrm{mic}} = 0.15-0.17$ m is assumed, this coherence function needs to be evaluated only once. The coefficients

$$\Gamma_{x_lx_r}^{(\mathrm{head})}(f) = \frac{\left|\int\limits_{\varphi=0}^{2\pi}\int\limits_{\theta=0}^{\pi} H_l(f,\theta,\varphi)H_r^*(f,\theta,\varphi)\sin\theta d\theta d\varphi\right|}{\sqrt{\int\limits_0^{2\pi}\int\limits_0^{\pi}|H_l(f,\theta,\varphi)|^2\sin\theta d\theta d\varphi\int\limits_0^{2\pi}\int\limits_0^{\pi}|H_r(f,\theta,\varphi)|^2\sin\theta d\theta d\varphi}} \tag{35}$$

TABLE II
COEFFICIENTS OF THE BINAURAL COHERENCE MODEL FOR $d_{\text{mic}} = 0.17$ m USING A NONLINEAR LEAST-SQUARES FITTING

p	a_p	b_p	c_p
1	1	18.97	291.1
2	$14.5 \cdot 10^{-3}$	875.2	105.7
3	$2.38 \cdot 10^{-3}$	1371	151.5

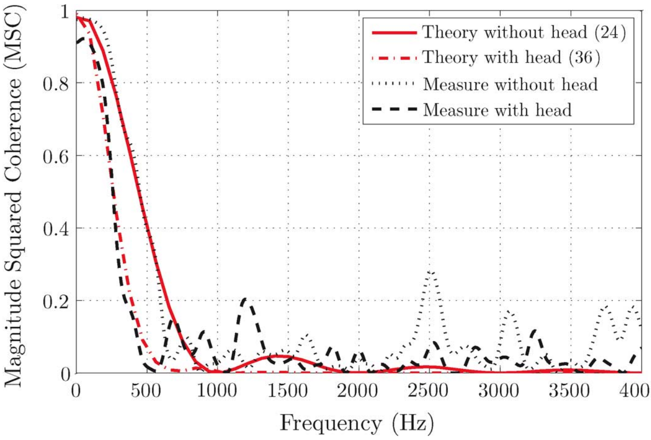


Fig. 8. Magnitude squared coherence of ideal diffuse sound field and shadowing influence. Plotted are the theoretical curves and results from measurements in a reverberant environment ($T_{60} = 0.72$ s).

for $d_{\text{mic}} = 0.17$ m and a mixture of $P = 3$ Gaussians are calculated using the MATLAB Curve Fitting Toolbox and listed in Table II. The root mean squared error (RMSE) between the solution of (35) and the approximation (36) is $2.4 \cdot 10^{-3}$ in the frequency range 1–48000 Hz.

Fig. 8 shows the corresponding curves for two microphones at a distance of $d_{\text{mic}} = 0.17$ m. The functions are plotted as the squared magnitudes of the coherence function $\Gamma^2_{x_l x_r}(\Omega)$ bounded above 0.99. The theoretical curves represent the ideal diffuse sound field without head (24) and the sound field with head shadowing (36). The measured curves have been obtained by a set of measured binaural room impulse responses of a lecture room, with and without a dummy head. We conclude that the proposed coherence model greatly matches the measured data and is appropriate for binaural dereverberation.

It could be assumed that the influence of the head can be modeled by scaling d_{mic} of the ideal diffuse coherence in (24). However, it turned out in several experiments that this does not lead to a sufficient solution compared to the model of (36).

V. EXPERIMENTS

In order to evaluate the performance of the discussed dereverberation concept, experiments with three different binaural room impulse responses from the Aachen Impulse Response (AIR) database [29] were carried out. The degraded speech files are generated using eight anechoic speech files (four female and four male speakers) from the NTT database each convolved with the different BRIRs.

For an objective evaluation, the non-intrusive measurement based on the speech to reverberation modulation energy ratio (SRMR) is employed [47]. Furthermore, the Bark spectral distortion (BSD) is used as a perceptually motivated spectral distance measure [48]. The reference signal for the BSD is the direct path signal [see (25)]. All signal levels are normalized to −26 dBov using the ITU-T Rec. P.56 speech voltmeter [49]. Silence periods have been removed before evaluation using the voice activity detector (VAD) of the AMR-WB speech codec [50]. Further simulation parameters are listed in Table III.

TABLE III
MAIN SIMULATION PARAMETERS

Parameter	Settings
Sampling frequency	$f_s = 16\,\text{kHz}$
Frame length	$L = 256$
FFT length	$M = 256$
Frame overlap	50% (Hann window)
Smoothing factors	$\alpha_2 = 0.9, \alpha_3 = 0.8$
Coherence threshold	$\Gamma_{\text{max}} = 0.99$
Gain factor thresholds	$G^{\text{late}}_{\text{min}} = G^{\text{coh}}_{\text{min}} = 0.3$
Late reverberant time span	$T_l = 0.1\,\text{s}$
Gain smoothing threshold	$\zeta_{thr} = 0.4$
Gain smoothing scaling factor	$\Psi = 25$

Section V-A gives the evaluation results for the proposes two-stage system as well as an independent analysis of both stages. In a listening experiment related to the discussions in Section II-B, we show how bilateral dereverberation influences the perceived spatial impression compared to binaural dereverberation.

For the experiments, we use the weighting gains of the dual-channel algorithm by Allen *et al.* [51] as a reference. This algorithm is related to the aforementioned Wiener filter (Stage II) since it uses directly the estimated coherence. The corresponding gains are calculated by

$$G_{\text{allen}}(\lambda, \mu) = \frac{|\hat{\Phi}_{x_l x_r}(\lambda, \mu)|}{\sqrt{\hat{\Phi}_{x_l x_l}(\lambda, \mu) \cdot \hat{\Phi}_{x_r x_r}(\lambda, \mu)}} \tag{37}$$

and applied to each channel according to the proposed binaural dereverberation concept. A similar approach using the magnitude squared coherence weighting gains has been performed in [15]. However, our experiments have shown that this exhibits higher processing artifacts compared to (37). The algorithms in [16] and [17] cannot be taken into account for a fair comparison since they perform a directional filtering and different weighting gains to both channels. Besides that, an extension of the LP residual enhancement modules in the two-stage algorithms [5], [6] to a binaural-input binaural-output structure is not straightforward and out of the scope of this paper.

A. *Binaural Dereverberation Experiments*

This subsection gives the results of five different binaural dereverberation algorithms. The corresponding weighting gains are calculated as follows.

- Dereverberation based on a statistical model of late reverberation (Stage I only) using
  1) LATE: the weighting gains of (16) calculated from the reference signal (17).
- Coherence-based dereverberation (Stage II only) using

2) COH-ALLEN: the estimated coherence as weighting gains (37).
3) COH-MCCOWAN: the ideal diffuse sound field model (24).
4) COH-HEAD: the proposed binaural diffuse sound field model taking the shadowing effects into account (36).

- Two-stage system (Fig. 5) (Stages I and II) using

5) TWO-STAGE: the proposed two-stage algorithm.

For estimating the reverberation time (required for LATE and TWO-STAGE), several algorithms exist in the literature. Since we focus on the application of the technique to binaural processing, we simply calculate T_{60} directly from the impulse response. An overview of blind and semi-blind methods for a direct estimation out of the reverberant speech can be found, e.g., in [38].

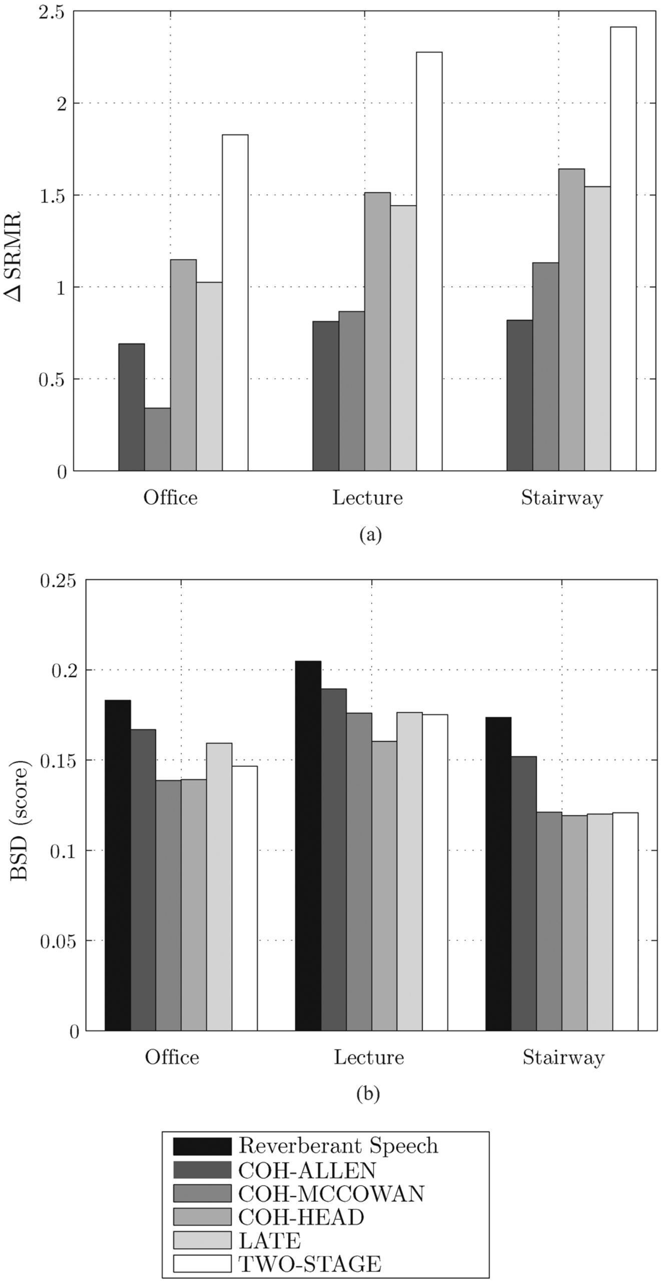
For each channel (left and right), the measurements are calculated separately and averaged. The ΔSRMR gives the enhancement compared to the reverberant speech, averaged over all dereverberated files. The BSD score measures the distortion between the direct speech component and the processed signal. In case of the stairway hall RIR, all simulation results are averaged over the azimuth angles $0, 15, \ldots, 90^\circ$ and are depicted in Fig. 9.

It turns out that among the compared coherence-based techniques, the proposed algorithm (COH-HEAD) shows the highest amount of reverberation reduction in terms of SRMR improvement. It even outperforms the LATE-algorithm for all tested scenarios. The BSD improvements are almost equivalent for all coherence-based techniques, which means that using the proposed coherence model (COH-HEAD) does not cause more distortions compared to the COH-ALLEN and COH-MCCOWAN algorithm. The two-stage system (TWO-STAGE) further increases the SRMR measure and outperforms all other algorithms.

This tendency of the objective measurements was verified by informal listening tests. It could be observed that the LATE algorithm reduces greatly the late reverberant tail but no early reverberation. The coherence-based algorithms made the processed signal to sound more “clear” (reduction in coloration). This effect was audible in particular for the two-stage system, where a reduction of both coloration and overlap-masking results in the best listening comfort without audible distortions among all tested approaches.

To illustrate the reduction of room reverberation, the spectrograms of a clean, reverberant and processed (TWO-STAGE) speech segment are shown in Fig. 10. It can be seen that the proposed two-stage algorithm is capable of enhancing reverberant speech. Due to the selected moderate simulation settings, a certain amount of reverberation still remains. However, a more aggressive setting could further reduce reverberation at the cost of audible speech distortions.

In a further experiment, we evaluated the need for the time-alignment used for the coherence-based algorithms (COH-X). The dereverberation performance was measured in terms of the SRMR in dependency of the azimuth angle (using the stairway hall BRIRs). It was observed, that without the time-alignment, a sufficient enhancement can only be obtained in the range $0^\circ \leq \theta \leq 30^\circ$. The time-alignment ensures a similar dereverberation performance over the entire azimuth range.



(a)

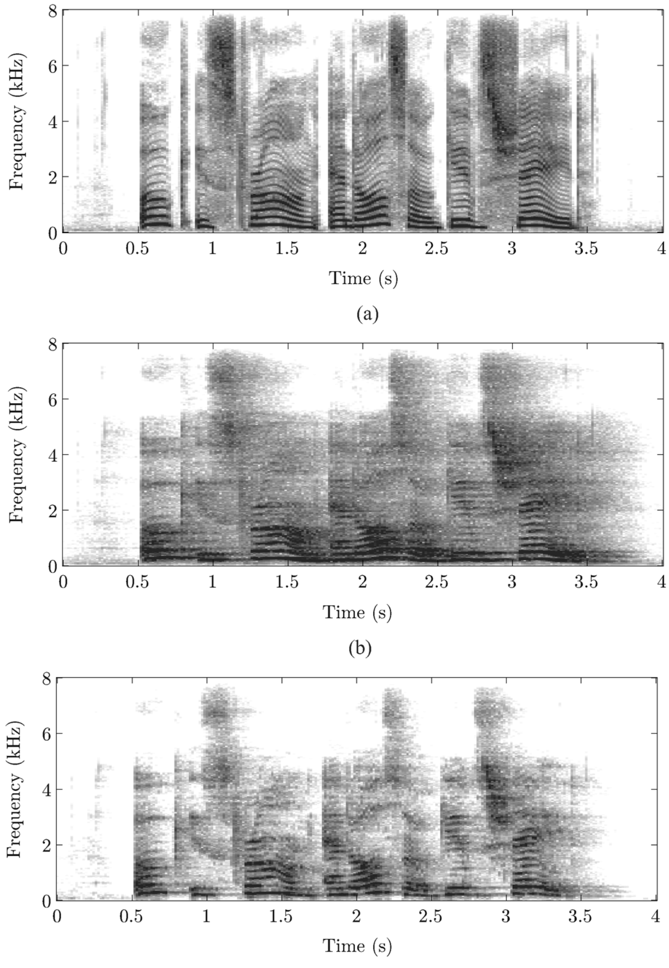
(b)

(c)

Fig. 9. Simulation results of the different dereverberation techniques. SRMR entries (a) give the difference to the reverberant speech, BSD (b) indicates the spectral distortions between the direct speech signal and the processed signal. (a) Speech-to-reverberation modulation energy ratio (SRMR). (b) Bark spectral distortion (BSD). (c) Legend.

B. Influence of Bilateral Processing by A Listening Test

The degradation of the binaural cues due to bilateral dereverberation (see Section II-B) has also been investigated with an informal listening experiment. During the test with 17 experienced listeners, three different signals were presented to the participants: the reverberant speech, the processed signal using a binaural dereverberation algorithm (A) and the processed signal after bilateral dereverberation (B). The test signals (A) are processed using the binaural algorithm (LATE) as described



(a)

(b)

(c)

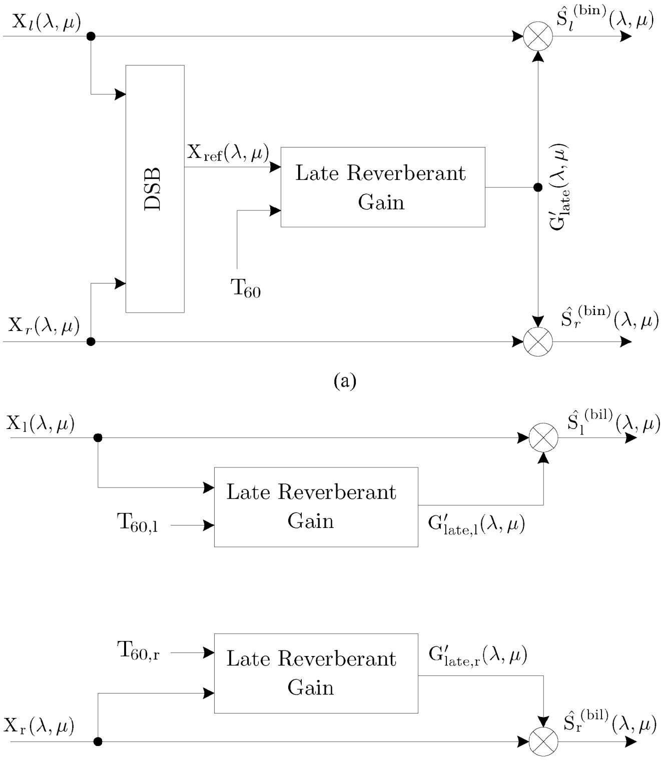
Fig. 10. Spectrograms of (a) clean, (b) reverberant, and (c) processed speech (TWO-STAGE) using BRIRs of the lecture room. (a) Clean speech. (b) Reverberant speech. (c) Processed speech (TWO-STAGE).

in Section III-B without gain smoothing [see Fig. 11(a)]. The bilateral signals (B) are generated using the same algorithm in a bilateral configuration, which means that the gains of (16) are calculated and applied individually to each channel [see Fig. 11(b)].

Since the other binaural algorithms (COH-X and TWO-STAGE) all require two input channels, a bilateral processing is not possible and hence, a comparison to the LATE algorithm would not be fair.

For each of the sentences, the listeners were asked to judge the overall speech quality as well as the audible modifications in the interaural time and level differences (compared to the provided reverberant speech). The listeners could choose between "A sounds better than B," "B sounds better than A," and "no preference." The samples could be played ad libitum before the probands had to make their judgments. The reverberant signals are generated using binaural room impulse responses of the stairway hall at different azimuth angles and distances.

In order to ensure high quality audio and to avoid distortions due to the headphone, a calibrated HEAD Acoustics PEQ V digital equalizer in combination with a Sennheiser HD600 headphone was used. The test took place in a low-reverberant studio booth having a high sound isolation of 42 dB.



(a)

(b)

Fig. 11. Block diagrams of (a) binaural processing and (b) bilateral processing used to generate the audio files for the listening test. (a) Binaural processing. (b) Bilateral processing.

TABLE IV
RESULTS OF THE LISTENING TEST

Simulation setup	No preference	Binaural dereverberation Fig. 11(a)	Bilateral dereverberation Fig. 11(b)
$\theta = 0°$, $d = 2\,\mathrm{m}$	5.9 %	82.4 %	11.7 %
$\theta = 60°$, $d = 1\,\mathrm{m}$	11.7 %	64.8 %	23.5 %
Average	**8.8 %**	**73.6 %**	**17.6 %**

The results for two different azimuth angles are stated in Table IV. It can be seen that for both setups most participants preferred the binaural dereverberation method (A) over the bilateral algorithm (B). This corresponds to the objective evaluation and conclusions of Section II-B.

VI. Conclusion

This paper proposed a novel two-stage speech enhancement algorithm for binaural dereverberation which is based on a model of the room impulse response (RIR) and a model of the sound field coherence. The algorithm operates in the frequency domain and consists of two components: The first stage of the algorithm is based on a statistical model of the RIR and comprises a spectral subtraction rule which depends on the variance of the late reverberant speech. It includes a frequency smoothing process of the spectral gains to reduce musical tones. In a second stage, the residual reverberation is attenuated by a dual-channel Wiener filter which is based on a new coherence model taking into account head shadowing. The overall binaural input-output structure does not affect the

TABLE V
PROPERTIES OF THE DIFFERENT ROOMS

Room	T_{60}	d_{LM}	θ
Office room	0.37 s	1 m	0° (frontal)
Lecture room	0.72 s	5.5 m	0°
Stairway hall	0.69 s	2 m	$0, 15, \ldots, +90°$

most important binaural cues, i.e., interaural time difference (ITD) and interaural level difference (ILD), and hence, keeps the localization ability. This was motivated by investigations how state-of-the art dereverberation algorithms influence the binaural cues in bilateral processing. In simulations with measured binaural room impulse responses, the proposed system achieves a significant reduction of early and late reverberation, which was confirmed by informal listening tests. A further enhancement, especially in rooms with moderate reverberation, could be obtained by means of an adaptive coherence model based on a measure of the "diffusiveness."

APPENDIX
AACHEN IMPULSE RESPONSE (AIR) DATABASE

The AIR database[1] consists of binaural room impulse responses which have been measured in different realistic environments [29]. All measurements have been performed with and without a dummy head (HEAD Acoustics HMS II.4). Reverberation times T_{60}, loudspeaker-microphone distances d_{LM}, and azimuth angles θ between head and loudspeaker are as stated in Table V. The binaural room impulse responses (BRIR) of the stairway hall are normalized for all angles such that the direct path energy of left and right channel is equal for $\theta = 0°$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments which helped to improve the presentation of this paper. Additionally, the authors would like to thank H. W. Löllmann, C. Nelke, and C. Norrenbrock for valuable contributions. A part of this research has been carried out during the stay of M. Jeub in the Speech and Audio Processing Lab at Imperial College London, London, U.K. Many thanks for fruitful discussions and suggestions to P. A. Naylor, E. A. P. Habets, and M. R. P. Thomas.

REFERENCES

[1] K. Lebart, "Speech dereverberation applied to automatic speech recognition and hearing aids," Ph.D. dissertation, Univ. de Rennes, Rennes, France, 1999.

[2] P. Naylor and N. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2005.

[3] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[4] H. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Appl. Signal Process.*, vol. 1, 2009.

[5] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[6] N. Gaubitch, E. Habets, and P. Naylor, "Multimicrophone speech dereverberation using spatiotemporal and spectral processing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2008, pp. 3222–3225.

[7] T. van den Bogeart, T. Klasen, M. Moonen, L. van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 515–526, 2005.

[8] J. Blauert, *Spatial Hearing—The Psychophysics of Human Sound Localization*, Rev. ed. Cambridge, MA: MIT Press, 1996.

[9] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, vol. 18, pp. 2915–2929, 2005.

[10] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output. I. Fixed-processing systems," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 529–542, Jun. 1997.

[11] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output. II. A two microphone adaptive system," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[12] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–14, 2006.

[13] T. van den Bogeart, "Preserving binaural cues in noise reduction algorithms for hearing aids," Ph.D. dissertation, Katholieke Univ. Leuven, Leuven, Belgium, 2008.

[14] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. 4th Int. Symp. Commun., Control, Signal Process. (ISCCSP)*, Limassol, Cyprus, 2010.

[15] J. Peissig, "Binaurale hörgerätestrategien in komplexen störschallsituationen," Ph.D. dissertation, Univ. Göttingen, Göttingen, Germany, 1992.

[16] T. Wittkopp, "Two-channel noise reduction algorithms motivated by models of binaural interaction," Ph.D. dissertation, Universität Oldenburg, Oldenburg, Germany, 2001.

[17] T. Wittkopp and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, pp. 111–138, 2003.

[18] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. United With Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[19] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 4409–4412.

[20] M. Jeub and P. Vary, "Binaural dereverberation based on a dual-channel wiener filter with optimized noise field coherence," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 4710–4713.

[21] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, 1999.

[22] W. M. Hartmann, "How we localize sound," *Phys. Today*, pp. 24–29, Nov. 1999.

[23] L. A. Jeffress, "A place theory of sound localization," *J. Compar. Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, Feb. 1948.

[24] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006.

[25] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[26] M. Slaney, "Auditory toolbox," Interval Research Corp., Palo Alto, CA, 1998, Tech. Rep..

[27] R. Patterson, M. Allerhand, and C. Gigure, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Amer.*, vol. 98, no. 4, pp. 1890–1894, Oct. 1995.

[28] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[29] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conf. Digital Signal Process. (DSP)*, Santorini, Greece, 2009.

[30] Multi-lingual speech database for telephonometry, 1994, NTT-Corporation.

[1] http://www.ind.rwth-aachen.de/AIR

[31] B. Wright and Y. Zhang, "A review of the generalization of auditory learning," *Phil. Trans. R. Soc. B*, vol. 364, pp. 301–311, 2009.

[32] M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor, "A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2007, pp. 50–53.

[33] M. Jeub and P. Vary, "Enhancement of reverberant speech using the CELP postfilter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 3993–3996.

[34] S. Wehr, H. Puder, and W. Kellermann, "Blind source separation and binaural reproduction with hearing aids: An overview," in *ITG Fachtagung Sprachkommunikation*, Aachen, Germany, 2008.

[35] L. Laaksonen and J. Virolainen, "Binaural artificial bandwidth extension (B-ABE) for speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 4009–4012.

[36] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, Jan. 1980.

[37] H. Kuttruff, *Room Acoustics*, 5th ed. Oxon, U.K.: Spon, 2009.

[38] H. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Seattle, WA, 2008.

[39] M. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[40] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[41] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AE-15, no. 2, pp. 70–73, Jun. 1967.

[42] P. Vary and R. Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment*. Chichester, U.K.: Wiley, 2006.

[43] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New York, 1988, vol. 5, pp. 2578–2581.

[44] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[45] M. Dörbecker, "Mehrkanalige signalverarbeitung zur verbesserung akustisch gestörter sprachsignale am beispiel elektronischer hörhilfen," Ph.D. dissertation, RWTH Aachen Univ., Aachen, Germany, 1998.

[46] M. Dörbecker, "Sind kohärenzbasierte störgeräuschreduktionsverfahren für elektronische hörhilfen geeignet? Modelle zur beschreibung der kohärenzeigenschaft," in *Proc. ITG-Fachtagung Sprachkommunikation*, Dresden, Germany, 1998.

[47] T. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Seattle, WA, 2008.

[48] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[49] Objective measurement of active speech level, 1993, ITU-T Rec. P.56, ITU.

[50] Adaptive multi-rate—wideband speech codec, voice activity detector, vol. 6.0.0, 2004, 3GPP TS 26.194.

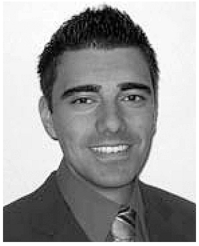
[51] J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, 1977.

Marco Jeub (S'08) joined a dual studies program of the European Aeronautic Defence and Space Company (EADS), Munich, Germany, in 2002, where he received the Dipl.-Ing. (BA) degree in communications engineering in 2005. He received the M.Sc. degree from the Technical University Berlin, Berlin, Germany, in 2007, where he participated in a one year exchange program with the Shanghai Jiao Tong University, Shanghai, China. He is currently pursuing the Ph.D. degree at the Institute of Communication Systems and Data Processing, RWTH Aachen University, Aachen, Germany.

His research interests cover the areas of single-and multichannel speech enhancement algorithms including noise suppression and dereverberation.

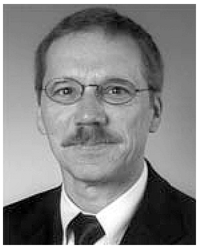
Magnus Schäfer received the Dipl.-Ing. degree in information and communication technology from RWTH Aachen University, Aachen, Germany, in 2006. He is currently pursuing the Ph.D. degree at the Institute of Communication Systems and Data Processing, RWTH Aachen University.

His research interests cover the areas of single- and multichannel speech and audio coding as well as speech enhancement.



Thomas Esch (S'07) received the Dipl.-Ing. degree in information and communication technology from RWTH Aachen University, Aachen, Germany, in 2005, where he is currently pursuing the Dr.-Ing. degree.

He is currently with the Institute of Communication Systems and Data Processing, RWTH Aachen University. His main research interests are digital speech and audio processing, including speech enhancement in noisy environments.



Peter Vary (M'85–SM'04–F'09) received the Dipl.-Ing. degree in electrical engineering from the University of Darmstadt, Darmstadt, Germany, in 1972 and the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Erlangen, Germany, in 1978.

In 1980, he joined Philips Communication Industries (PKI), Nuremberg, Germany, where he became head of the Digital Signal Processing Group. Since 1988, he has been a Professor at RWTH Aachen University, Aachen, Germany, and head of the Institute of Communication Systems and Data Processing. His main research interests are speech coding, joint source-channel coding, error concealment, and speech enhancement including noise suppression, acoustic echo cancellation, and artificial wideband extension.