Binaural dereverberation based on interaural coherence histograms

Adam Westermann\textsuperscript{b,c} and Jörg M. Buchholz\textsuperscript{b}

\textit{National Acoustic Laboratories, Australian Hearing, 16 University Avenue, Macquarie University, New South Wales 2109, Australia}

Torsten Dau

Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, DK-2800 Kgs. Lyngby, Denmark

(Received 23 July 2012; revised 4 March 2013; accepted 18 March 2013)

A binaural dereverberation algorithm is presented that utilizes the properties of the interaural coherence (IC) inspired by the concepts introduced in Allen \textit{et al.} [J. Acoust. Soc. Am. \textbf{62}, 912–915 (1977)]. The algorithm introduces a non-linear sigmoidal coherence-to-gain mapping that is controlled by an online estimate of the present coherence statistics. The algorithm automatically adapts to a given acoustic environment and provides a stronger dereverberation effect than the original method presented in Allen \textit{et al.} [J. Acoust. Soc. Am. \textbf{62}, 912–915 (1977)] in most acoustic conditions. The performance of the proposed algorithm was objectively and subjectively evaluated in terms of its impacts on the amount of reverberation and overall quality. A binaural spectral subtraction method based on Lebart \textit{et al.} [Acta Acust. Acust. \textbf{87}, 359–366 (2001)] and a binaural version of the original method of Allen \textit{et al.} were considered as reference systems. The results revealed that the proposed coherence-based approach is most successful in acoustic scenarios that exhibit a significant spread in the coherence distribution where direct sound and reverberation can be segregated. This dereverberation algorithm is thus particularly useful in large rooms for short source-receiver distances. \textcopyright 2013 Acoustical Society of America.

\url{http://dx.doi.org/10.1121/1.4799007}

PACS number(s): 43.60.Mn, 43.66.Pn, 43.66.Yw, 43.55.Hy [SAF] Pages: 2767–2777

I. INTRODUCTION

When communicating inside a room, the speech signal is accompanied by multiple reflections originating from the surrounding surfaces. The impulse response of the room is characterized by early reflections (first 50–80 ms of the room response) and late reflections or reverberation (Kuttruff, 2000).

In terms of auditory perception, early reflections mainly introduce coloration (Salomons, 1995), are beneficial for speech intelligibility (Bradley \textit{et al.}, 2003), and are typically negligible with regard to sound localization (Blauert, 1996). In contrast, reverberation smears the temporal and spectral features of the signal; this commonly deteriorates speech intelligibility (Moncur and Dirks, 1967), listening comfort (Ljung and Kjellberg, 2010), and localization performance. Some of the preceding negative effects are partly compensated for in normal-hearing listeners by auditory mechanisms such as the precedence effect (Litovsky \textit{et al.}, 1999), monaural/binaural de-coloration, and binaural dereverberation (e.g., Zurek, 1979; Blauert, 1996; Buchholz, 2007). However, in hearing-impaired listeners, reverberation can be detrimental because of reduced hearing sensitivity as well as decreased spectral and/or temporal resolution (e.g., Moore, 2012). In addition, a hearing impairment may affect the auditory processes that otherwise help listening in reverberant environments (e.g., Akeroyd and Guy, 2011; Goverts \textit{et al.}, 2001).

Thus suppressing reverberation by utilizing a dereverberation algorithm, e.g., in hands-free devices, binaural telephone headsets, and digital hearing aids, might improve speech intelligibility, localization performance, and ease of listening.

Several dereverberation algorithms have been proposed in the literature. They address either early reflections or reverberation, are blind or non-blind, or use single or multiple input channels. Typical methods for suppressing early reflections include inverse filtering (e.g., Neely and Allen, 1979; Mourjopoulos, 1992) and linear prediction residual processing (e.g., Gillespie \textit{et al.}, 2001; Yegnanarayana \textit{et al.}, 1999).

Processing methods for suppressing reverberation are typically based on spectral enhancement techniques, which decompose the speech signal in time and frequency and suppress components that are estimated to be mainly reverberant. Different approaches have been proposed to realize this estimation.

Allen \textit{et al.} (1977) proposed a binaural approach where gain factors are determined by the diffuseness of the sound field between two spatially separated microphones. They suggested two methods for calculating gain factors, one of which represented the coherence function of the two channels. However, because of a cophase-and-add stage, which combined the binaural channels, only a monaural output was provided.

Kollmeier \textit{et al.} (1993) extended the original approach of Allen \textit{et al.} (1977) by applying the original coherence gain factor separately to both channels, thus providing a binaural output. Jeub
and Vary (2010) demonstrated that synchronized spectral weighting across binaural channels is important for preserving binaural cues. In Simmer et al. (1994), a coherence-based Wiener filter was suggested that estimates the reverberation noise from a model of coherence between two points in a diffuse field. Their method was further refined in McCowan and Bourland (2003) and Jeub and Vary (2010) where acoustic shadow effects from a listener’s head and torso were included.

Single-channel spectral enhancement techniques employ different methods for reverberation noise estimation. Wu and Wang (2006) proposed that the reverberation noise can be estimated in the time-frequency domain from the power spectrum of preceding speech. Lebart et al. (2001) assumed an exponential decay of reverberation with time. In their model, the signal-to-reverberation noise ratio in each time frame is determined by the energy in the current frame compared to that of the previous. Common problems with these methods are the so-called “musical noise” effects and the suppression of signal onsets, both caused by an overestimation of the reverberation noise. Tsilfidis and Mourjopoulos (2009) introduced a gain-adaptation technique that incorporates knowledge of the auditory system to suppress musical noise. They also proposed a power relaxation criterion to maintain signal onsets. Alternative modifications based on the signal direct-to-reverberant energy ratio (DRR) have been proposed by Habets (2010). An overview of dereverberation methods can be found in Naylor and Gaubitch (2010).

In the present study, a binaural dereverberation algorithm is introduced that utilizes the properties of the interaural coherence (IC), inspired by the concepts introduced in Allen et al. (1977). Applying the method of Allen et al. (1977) to different acoustic scenarios revealed that the dereverberation performance strongly varied between scenarios. To better understand this behavior, an investigation of the IC in different acoustic scenarios was performed, showing how IC distributions varied over frequency as a function of distance and reverberation time. Because the linear coherence-to-gain mapping of the previous coherence-based methods [such as Allen et al. (1977)] cannot account for this behavior, a non-linear sigmoidal coherence-to-gain mapping is proposed here that is controlled by an online estimate of the inherent coherence statistics in a given acoustical environment. In this way, frequency-specific processing and weighting characteristics are applied that result in an improved dereverberation performance, especially in acoustic scenarios where the coherence varies strongly over time and frequency. The performance of the proposed algorithm is evaluated objectively and subjectively, assessing the amount of reverberation and overall signal quality. The performance is compared to two reference systems, a binaural spectral subtraction method, inspired by Lebart et al. (2001), and a binaural version of the original method of Allen et al. (1977).

II. THE COHERENCE-BASED DEREVERBERATION ALGORITHM

A. Signal processing

The signal processing of the proposed binaural dereverberation method is illustrated in Fig. 1. Two reverberant time signals, recorded at the left and right ear of a person or a dummy head, \( x_l(n) \) and \( x_r(n) \), are transformed to the time-frequency domain using the short-time Fourier transform (STFT) (Allen and Rabiner, 1977). This results in the complex-valued short-term spectra \( X_l(m,k) \) and \( X_r(m,k) \), where \( m \) denotes the time frame and \( k \) the frequency band.

For the STFT, a Hanning window of length \( L \) (including zero-padding of length \( L/2 \)) and a 75% overlap (i.e., applying a time shift of \( L/4 \) samples) between successive windows are used. For each time-frequency bin, the absolute value of the interaural coherence (IC or coherence from here) is calculated, and third-octave smoothing is applied (Hatziantoniou and Mourjopoulos, 2000). A sigmoidal mapping stage is subsequently applied to the coherence estimates to realize a coherence-to-gain mapping. This mapping realizes a time-varying filter that attenuates time-frequency regions with a low IC (i.e., that are strongly affected by reverberation) and leaves regions untouched with high IC (i.e., where the direct sound is dominant). The parameters of the sigmoidal

![Block diagram of the proposed signal processing method.](image)

FIG. 1. Block diagram of the proposed signal processing method. The signals recorded at the ears, \( x_l(n) \) and \( x_r(n) \), are transformed via the STFT to the time-frequency domain, resulting in \( X_l(m,k) \) and \( X_r(m,k) \). The IC is calculated for each time-frequency bin, and third-octave smoothing is applied. Statistical long-term properties of the IC are used to derive parameters of a sigmoidal mapping stage. The mapping is applied to the IC to realize a coherence-to-gain relationship, and subsequent temporal windowing is performed. The derived gains (or weights) are applied to both channels \( X_l(m,k) \) and \( X_r(m,k) \). The dereverberated signals, \( \hat{s}_l(n) \) and \( \hat{s}_r(n) \), are reconstructed by applying the inverse STFT.
coherence-to-gain mapping are calculated based on an online estimate of the statistical properties of the IC (i.e., applying frequency-dependent coherence histograms). To suppress potential aliasing artifacts that may be introduced by applying this filtering process, temporal windowing is applied (Kates, 2008). This is realized by applying an inverse STFT to the derived filter gains and then truncating the resulting time-domain representation to a length of \( L/2 + 1 \). This filter response is then zero-padded to a length of \( L \) and another STFT is performed. The resulting filter gains are applied to both channels \( X_i(m, k) \) and \( X_j(m, k) \). The dereverberated signals, \( \hat{s}_i(n) \) and \( \hat{s}_j(n) \), are finally reconstructed by applying the inverse STFT and then adding the resulting (overlapping) signal segments (Allen and Rabiner, 1977).

### B. Signal decomposition and coherence estimation

From the time-frequency signals \( X_i(m, k) \) and \( X_j(m, k) \), the IC is calculated as

\[
C_{ij}(m, k) = \frac{|\Phi_{ij}(m, k)|}{\sqrt{|\Phi_{ii}(m, k)| |\Phi_{jj}(m, k)|}},
\]

where \( \Phi_{ij}(m, k) \) and \( \Phi_{ji}(m, k) \) represent the exponentially weighted short-term cross-correlation and auto-correlation functions

\[
\Phi_{ij}(m, k) = x\Phi_{ii}(m, k - 1) + |X_i(m, k)|^2, \quad \Phi_{ji}(m, k) = x\Phi_{jj}(m, k - 1) + |X_j(m, k)|^2, \quad \Phi_{ij}(m, k) = x\Phi_{ij}(m, k - 1) + X_i(m, k)X_j^*(m, k),
\]

where \( x \) is the recursion constant and * indicates the complex conjugate. These coherence estimates yield values between 0 (for fully incoherent signals) and 1 (for fully coherent signals). If the time window applied in the STFT exceeds the duration of the room impulse responses (RIR) between a sound source and the two ears, the coherence approaches unity (Jacobsen and Roisin, 2000). When shorter time windows than the duration of the involved RIRs are applied in the STFT (which is typically the case), the estimated coherence is highly influenced by the used window length (Scharerr, 2010).

The recursion constant \( x \) determines the temporal integration time \( \tau \) of the coherence estimate, which is given by

\[
\tau = \frac{L}{4f_s \cdot \ln(x)},
\]

where \( f_s \) is the sampling frequency. The integration time needs to be short enough to follow the changes in the involved signals (i.e., speech), but long enough to provide reliable coherence estimates. In this study, an STFT window length of 6.4 ms (identical to that of Allen et al., 1977 and corresponding to 282 samples) and a recursion constant of \( x = 0.97 \) (corresponding to a time constant \( \tau \approx 100 \) ms) are used. The applied time constant is similar to the ones used in previous work (e.g., Kollmeier et al., 1993) and is able to follow syllabic changes.

### C. Coherence-to-gain mapping

To cope with the different frequency-dependent distributions of the IC observed in different acoustic scenarios (see Sec. IV), a coherence-distribution dependent coherence-to-gain mapping is introduced. This is realized by a sigmoid function which is controlled by an (online) estimate of the statistical properties of the IC in each frequency channel. The resulting filter gains are

\[
G_{\text{sig}}(m, k) = \frac{1 - g_{\min}}{1 + \exp\{-k_{\text{slope}}(k)[C_{LR}(m, k) - k_{\text{shift}}(k)]\}} + g_{\min},
\]

where \( k_{\text{slope}} \) and \( k_{\text{shift}} \) control the sigmoidal slope and the position. The minimum gain \( g_{\min} \) is introduced to limit signal processing artifacts associated with applying infinite attenuation.

To calculate the frequency-dependent parameters of the sigmoidal mapping function, coherence samples for a duration, defined by \( f_{\text{sig}} \), are gathered in a histogram. For constant source-receiver location, \( f_{\text{sig}} \) of several seconds was found to provide a good compromise between stable parameter estimates and as short as possible adaptation time. For moving sources and changing acoustic environments, the method for updating the sigmoidal parameters might need revision.

A coherence histogram (shown as a Gaussian distribution for illustrative purposes) is exemplified in Fig. 2 (gray curve) together with the corresponding first (\( Q_1 \)) and second (\( Q_2 \) or median) quartile. An example sigmoidal coherence-to-gain mapping is represented by the black solid curve. The linear mapping applied by Allen et al. (1977) is indicated by the black dashed curve. When applying a linear mapping, the gain (given by \( C_{LR} \)) is smoothly turned down with decreasing IC (i.e., increasing amount of reverberation), and thus almost all samples are attenuated to a certain degree. In contrast, the sigmoidal mapping strongly suppresses samples with low IC (which is only limited by \( g_{\min} \)) and leaves samples with higher IC untouched. In this way, a much stronger suppression of reverberation is achieved.

![FIG. 2. Idealized IC histogram distribution in one frequency-channel (gray curve). The coherence-to-gain relationship in the specific channel is calculated to intersect \( G_{\text{sig}}(C_{LR}) = g_{\min} + k_p \) and \( G_{\text{sig}}(C_{LR}) = 1 - k_p \). Thereby, \( g_{\min} \) denotes the maximum attenuation and \( k_p \) determines the processing degree.


Westermann et al.: Binaural dereverberation
The degree of processing is determined by $k_p$, which directly controls the slope of the sigmoidal mapping. The parameters $k_{\text{slope}}$ and $k_{\text{shift}}$ of the sigmoidal mapping are derived by inserting the two points $G_{\text{sig} C_0 = Q_1} = g_{\text{min}} + k_p$ and $G_{\text{sig} C_0 = Q_2} = 1 - k_p$ into Eq. (6) and then solving the resulting two equations for $k_{\text{slope}}$ and $k_{\text{shift}}$ (see Fig. 2), i.e.,

$$k_{\text{shift}}(k) = \frac{\ln(G_{\text{sig} C_0 = Q_1}^{-1})}{\ln(G_{\text{sig} C_0 = Q_2}^{-1})} Q_2(k) + Q_1(k),$$

$$k_{\text{slope}}(k) = \frac{\ln(G_{\text{sig} C_0 = Q_1}^{-1}) - 1}{Q_1(k) - k_{\text{shift}}},$$

where $Q_1(k)$ and $Q_2(k)$ are estimated in each frequency channel as the first and second quartile of the measured coherence histograms and $g_{\text{min}}$ and $k_p$ are predetermined parameters. Following such approach, $k_p$ provides the only free parameter, which directly controls the slope of the sigmoidal function and thus, determines the degree (or aggressiveness) of the dereverberation processing. For speech presented in an auditorium with source-receiver distances of 0.5 and 5 m (see Sec. IV), examples of sigmoidal mappings are shown in Fig. 3 for different values of $k_p$ in the 751.7 Hz frequency channel. It can be seen that the coherence-to-gain function steepens as $k_p$ decreases (i.e., as the processing degree increases). In addition, as the distribution broadens (from 5 to 0.5 m), the slope of the coherence-to-gain function decreases. Hence, in contrast to the original coherence-based dereverberation approach in Allen et al. (1977), which considered a fixed linear coherence-to-gain mapping (Fig. 2, dashed line), the proposed approach provides a flexible mapping function, which automatically adapts to any given acoustic condition.

D. Reference systems

To compare the performance of the proposed algorithm to the state-of-the-art algorithms described in the relevant literature, two additional dereverberation methods were implemented: The IC-based algorithm proposed by Allen et al. (1977) and the spectral subtraction-based algorithm described by Lebart et al. (2001). This approach relies on the estimation of reverberation noise in speech based on a model of the room impulse response (RIR). This model was derived from an estimation of the reverberation time. The binaural extension was realized by third-octave smoothing and temporal windowing (Sec. II A) were added. Hence, the same processing as shown Fig. 1 was applied except that the sigmoidal coherence-to-gain mapping was replaced by a linear mapping (see Fig. 2, dashed line). The same recursion constant and window length as in the first algorithm (1) were used.

(3) A binaural extension of the spectral subtraction approach described by Lebart et al. (2001). This approach relies on the estimation of reverberation noise in speech based on a model of the room impulse response (RIR). This model was derived from an estimation of the reverberation time. The binaural extension was realized by

### Table I. Processing parameters used for the proposed algorithm.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>$f_s$</td>
<td>44.1 kHz</td>
</tr>
<tr>
<td>Frame length</td>
<td>$L$</td>
<td>6.4 ms</td>
</tr>
<tr>
<td>Frame overlap</td>
<td>$\alpha$</td>
<td>75%</td>
</tr>
<tr>
<td>Recursion constant</td>
<td>$\beta$</td>
<td>0.97</td>
</tr>
<tr>
<td>Gain threshold</td>
<td>$g_{\text{min}}$</td>
<td>0.1</td>
</tr>
<tr>
<td>Processing degrees</td>
<td>$k_p$</td>
<td>{0.01; 0.2; 0.35}</td>
</tr>
<tr>
<td>Sigmoidal updating time</td>
<td>$t_{\text{sig}}$</td>
<td>3 s</td>
</tr>
</tbody>
</table>
(a) averaging the reverberation time estimates for the left and right channel and (b) synchronizing the spectral weighting in both channels. The latter was realized by calculating the weights for the left and right channel in each time-frequency bin and then applying the minimum value to both channels. The original processing parameters of Lebart et al. (2001) were used.

III. EVALUATION METHODS

To evaluate the performance of the proposed dereverberation algorithm, objective as well as subjective measures were applied. Reverberant speech was created by convolving anechoic speech with binaural room impulse responses (BRIRs), recorded at 0.5 and 5 m distances in an auditorium (see Appendix). The auditorium had a reverberation time of $T_{60} = 1.9 \text{s}$ at 2 kHz and DRRs of $-9.34$ and $-28 \text{ dB}$, respectively. Two anechoic sentences from the Danish speech database, recorded by Christiansen and Henrichsen (2011), were used, each spoken by both a male and a female talker, resulting in two sentences for each position.

A. Objective evaluation methods

Several metrics have been suggested to predict the performance and quality of dereverberation algorithms (Kokkinakis and Loizou, 2011; Goetze et al., 2010; Naylor and Gaubitch, 2010). Two commonly used objective measures were applied here to evaluate different aspects of the proposed dereverberation algorithm.

1. Signal-to-reverberation ratio

The segmental signal-to-reverberation (segSRR) ratio estimates the amount of direct signal energy compared to reverberant energy (e.g., Wu and Wang, 2006; Tsilfidis and Mourjopoulos, 2011) and was given by

\[
\text{segSRR} = \frac{10}{K} \log_{10} \left[ \frac{\sum_{n=0}^{N-1} (k_{\text{path}} s_d(n))^2}{\sum_{n=0}^{N-1} (k_{\text{path}} \delta(n) - \bar{\delta}(n))^2} \right],
\]

(9)

where $s_d(n)$ denotes the direct path signal, $\delta(n)$ the (reverberant) test signal, $k_{\text{path}}$ is a normalization constant, $N$ the frame-length (here 10 ms), $k = 0, ..., W - 1$ and $W$ the total number of frames. The direct sound was derived by convolving the anechoic speech signal with a modified (time-windowed) version of the applied BRIR, which only contained the direct sound component. The denominator provides an estimate of the reverberation energy by subtracting the waveform of the direct sound from the waveform of the tested signal (which includes the direct sound). The improvement in SRR was then calculated by

\[
\Delta \text{segSRR} = \text{segSRR}_{\text{proc}} - \text{segSRR}_{\text{ref}}.
\]

(10)

Thereby, segSRR$_{\text{ref}}$ was calculated from the original reverberant speech signal by convolving the anechoic speech with a given BRIR. The segSRR$_{\text{proc}}$ was calculated from the same reverberant speech signal but processed by the considered dereverberation algorithm. Hence, an algorithm that successfully suppresses reverberation should achieve SRR improvements of $\Delta \text{segSRR} > 0 \text{ dB}$.

Because time-based quality measures, such as the segSRR, are sensitive to any applied normalization, all signals were normalized to equal root mean square (RMS) levels before the actual segSRR was calculated. In addition, the level of the direct path signal was multiplied by the factor $k_{\text{path}}$, in such a way that the energy in the direct path was equal to the direct path component of the processed signal. The appropriate $k_{\text{path}}$ was determined numerically by minimizing the denominator in Eq. (9) for the case that the unprocessed (reference) reverberant signal was applied. Only frames with segSRR$_{\text{segSRR}} < -10 \text{ dB}$ were included in calculating the total segSRR from Eq. (9). This was done because the segSRR measure would otherwise be dominated by frames that mainly contain direct sound energy while frames that mainly contain reverberation energy would provide only a minor contribution.

2. Noise-mask ratio

The noise-mask ratio (NMR) is often used as an objective measure for evaluating the sound quality produced by dereverberation methods (e.g., Furuya and Kataoka, 2007; Tsilfidis et al., 2008). The measure is related to human auditory processing as only audible noise components (or artifacts) are considered. According to Brandenburg (1987), the NMR is defined as

\[
\text{NMR} = 10 \log_{10} \left( \frac{1}{W} \sum_{i=0}^{W-1} \sum_{b=0}^{B-1} \sum_{m=\omega_{hb}}^{\omega_{lb}} |R(\omega, m)|^2 \right),
\]

(11)

with $W$ denoting the total number of frames, $B$ the number of critical bands (or auditory frequency channels), and $C_b$ the number of frequency bins inside the critical band with index $b$. The power spectrum of the reverberation, $|R(\omega, m)|^2$, was calculated by subtracting the power spectrum of the anechoic signal from that of the test signal where $\omega$ is the angular frequency and $m$ is the time frame. The upper and lower cut-off frequencies were given by $\omega_{lb}$ and $\omega_{hb}$, respectively, and the masked threshold by $T_b(m)$, which depends on the spectral magnitude in the $b$th critical band (for details, see Brandenburg, 1987). The difference between the reverberant (reference) and processed NMR was then defined as

\[
\Delta \text{NMR} = \text{NMR}_{\text{proc}} - \text{NMR}_{\text{ref}}.
\]

(12)

As the amount of audible noise increases (i.e., NMR$_{\text{proc}}$ decreases), the resulting $\Delta \text{NMR}$ decreases. Thus smaller values of $\Delta \text{NMR}$ indicate a quality improvement.

B. Subjective evaluation methods

A subjective evaluation method similar to the multiple stimuli with hidden reference test (MUSHRA) was applied to subjectively evaluate the performance of the different
dereverberation algorithms (see ITU, 2003). These types of experiments have been widely applied to efficiently extract specific signal features even in cases where differences are very subtle (e.g., Lorho, 2010). A graphical user interface (GUI) was presented to the subjects to judge the attributes “amount of reverberation” and “overall quality” on a scale from 0 to 100 with descriptive adjectives: “Very little,” “little,” “medium,” “much,” and “very much.” The subjects could switch among six different processing methods: The original IC-based method, the proposed IC-based method with $k_p = 0.01$, 0.2, and 0.35, the spectral subtraction method, and an anchor. Anchors are an inherent trait of MUSHRA experiments to increase the reproducibility of the results and to prevent contraction bias (e.g., Bech and Zacharov, 2006). Additionally, subjects had access to the reference (unprocessed) stimulus via a “reference button.” Two different source-receiver positions (0.5 and 5 m) were considered, and each condition was repeated once. For an intuitive comparison with the objective evaluation results, the subjective scores were transformed to 100 - scores. The resulting scores were named “strength of dereverberation” and “overall loss of quality.”

To evaluate the quality of speech, the anchor was realized by distorting the reference signal using an adaptive multi-rate (AMR) speech coder (available from 3GPP TS26.073, 2008) with a bit-rate of 7.95 kbit/s. The resulting distortions were similar to the artifacts produced by the different dereverberation methods. Anchors for judging the amount of reverberation were created by applying a temporal half cosine window with a length of 600 ms to the BRIRs and thereby artificially reducing the resulting reverberation while keeping direct sound and early reflections. The unprocessed reference stimulus was not included as a hidden anchor because pilot experiments showed that this resulted in a significant compression bias of the subjects’ responses (for further details, see Bech and Zacharov, 2006). All experiments were carried out in a double-walled sound insulated booth, using a MATLAB GUI, Sennheiser HD-650 circumaural headphones and a computer with a RME DIGI96/8 PAD high-end sound card. The measurement setup was calibrated to produce a sound pressure level of 65 dB, measured in an artificial ear coupler (B&K 4153).

Ten (self-reported) normal-hearing subjects participated in the experiment. All subjects were either engineering acoustics students or sound engineers and were considered as experienced listeners. An instruction sheet was handed out to all subjects. Prior to the test, a training session was carried out to introduce the GUI and the applied terminology. There was no time limit for the experiment but, on average, the subjects required 1 h to complete the experiment.

IV. RESULTS

A. Effects of reverberation on speech in different acoustic environments

1. Spectrogram representations

The effects of reverberation on speech in a room are shown in the spectrograms in Fig. 4. The anechoic speech sample for a male speaker is shown in Fig. 4(a). The anechoic signal, convolved with one channel of a BRIR recorded in an auditorium at a 0.5 m distance (see Sec. IV) is shown in Fig. 4(b). A comparison of Figs. 4(a) and 4(b) reveals that a large number of the dips in the anechoic speech representation are filled due to the reverberation, i.e., the reverberation leads to a smearing both in the temporal and spectral domain.

2. Interaural coherence

The lowest levels of coherence exist in an isotropic diffuse sound field, where the coherence measured between two points is given by a sinc-function

$$C_{\text{diff}} = \frac{\sin(2\pi f d_{\text{mic}}/c)}{2\pi f d_{\text{mic}}},$$

with $c$ representing the speed of sound and $d_{\text{mic}}$ the distance between the two measuring points (Martin, 2001). In such a case, the coherence approaches unity at low frequencies and exhibits zero-crossings at frequencies corresponding to the distance between the two measurement points, as indicated by the solid curve in Fig. 5(a). A similar behavior is found for the IC but altered by the interference of the torso, head, and pinna of a listener (Jeub et al., 2009).
Figure 5(a) shows IC histograms for speech presented in a reverberation chamber, calculated from the binaural recordings of Hansen and Munch (1991). The algorithm defined in Sec. II A was first applied to describe the short-term IC of the binaural representation of an entire sentence spoken by a male talker. From the resulting IC values, the coherence histograms were derived. Gray scale reflects the number of occurrences (height of the histogram) in a given frequency channel. As expected from the ideal diffuse sound field, an increased coherence is observed below 1 kHz. Above 1 kHz, most coherence values are between 0.1 and 0.3. The lower limit of the obtained IC values and the IC spread of the distribution are caused by the non-stationarity of the input speech signal and the temporal resolution of the coherence estimation (i.e., the window length $L$ and the recursion constant $\alpha$).

Figures 5(b)–5(d) show example coherence histograms for 0.5, 5, and 10 m source-receiver distances in an auditorium with a reverberation time of $T_{60} = 1.9$ s at 2 kHz and a volume of 1150 m$^3$ (see Appendix for recording details). The overall coherence decreases with increasing distance between the source and the receiver. This results from the decreased direct-to-reverberant energy ratio at longer source-receiver distances. At very small distances [Fig. 5(b)], most coherence values are close to 1, indicating that mainly direct sound energy is present. In addition, the coherence arising from the diffuse field (with values between 0.1 and 0.3) is separable from that arising from the direct sound field. For the 5 m distance, substantially fewer frames with high coherence values are observed. This is because frames containing direct sound information are now affected by reverberation, and there is no clear separability anymore between frames with direct and reverberant energy. At a distance of 10 m, this trend continues as the coherence values further drop and the distribution resembles that found in the diffuse field, i.e., very little direct sound is available.

For small source-receiver distances, where the direct sound is separable from the diffuse sound field, a dereverberation algorithm that directly applies the short-term coherence as a gain [i.e., applying a linear coherence-to-gain mapping as proposed by Allen et al. (1977)] should suppress reverberant time-frequency segments and preserve direct sound elements. However, with increasing source-receiver distance, the effectiveness of such an algorithm can be expected to decrease, since direct sound elements will be increasingly “contaminated” by diffuse reverberation. Moreover, the observed different coherence histograms suggest that the optimal coherence-to-gain mapping depends on frequency as well as the specific acoustic condition. Because the dereverberation algorithm proposed in Allen et al. (1977) applies a fixed coherence-to-gain mapping, it can only provide a significant suppression of reverberation in very specific acoustic conditions. In addition, because of the limited coherence range at lower frequencies (where all IC values are rather high), a linear coherence-to-gain relationship would result in a high gain at lower frequencies for all acoustical conditions and would effectively act as a low-pass filter.

**B. Effects of dereverberation processing on speech**

The spectrogram shown in Fig. 4(c) illustrates the effect of dereverberation on speech. The proposed algorithm was applied with a moderate processing degree (i.e., $k_p = 0.2$). It
can be seen that a substantial amount of the smearing caused by the reverberation in the room [Fig. 4(b)] was reduced by the dereverberation processing.

1. Signal-to-reverberation ratio

Figure 6 (gray bars) shows the signal-to-reverberation ratio, $\Delta \text{segSRR}$ [Eq. (10)], for the different processing schemes. All algorithms show a significant reduction in the amount of reverberation (i.e., all exhibit positive values).

For the 0.5 m distance (left panel), the proposed algorithm (for $k_p = 0.2$) provides the best performance. For the lowest degrees of processing ($k_p = 0.35$), the performance is slightly below that attained for the spectral subtraction algorithm. For the 5 m distance (right panel), the proposed method for the highest processing degree ($k_p = 0.01$) performs comparably with the spectral subtraction method. As expected, the performance of the proposed method generally drops with decreasing processing degree (i.e., increasing $k_p$ value). The original IC-based method generally shows the poorest performance and provides essentially no reverberation suppression in the 0.5 m condition.

2. Noise-mask ratio

In Fig. 6, $\Delta \text{NMR}$ (white bars) is shown where smaller values correspond to less audible noise or better sound quality. For the different processing conditions, the original IC-based approach shows the best overall performance for both source-receiver distances. Considering the very small amount of dereverberation that is provided by this algorithm (see Sec. IV B 1 and Fig. 6), this observation is not surprising because the algorithm only has a minimal effect on the signal. The performance of the proposed method for high degrees of processing (i.e., $k_p = 0.01$) is similar or slightly better than that obtained with the spectral subtraction approach. For decreasing degrees of processing (i.e., $k_p = 0.2$ and 0.35), the performance of the proposed method increases, but at the same time, the strength of dereverberation (as indicated by segSRR) also decreases (see gray bars in Fig. 6). Considering both measures, segSRR and the NMR, the proposed method is superior for close sound sources (i.e., the 0.5 m condition with $k_p = 0.2$) and exhibits performance similar to the spectral subtraction method for the 5 m condition.

3. Subjective evaluation

The results from the subjective evaluation for each processing method are shown in Fig. 7. For better comparison with the objective results, the measured data were inverted (i.e., shown as $100 - \text{measured score}$). The attributes “amount of reverberation” and “overall quality” were consequently changed to “strength of dereverberation” and “loss of quality.” Considering the strength of dereverberation, indicated by the gray bars, the proposed approach exhibited the best performance for $k_p = 0.01$ at both distances. As the degree of processing decreases (i.e., for increasing values of $k_p$), the strength of dereverberation decreases. The improvement relative to the spectral subtraction approach is considerably higher for the 0.5 m distance (left panel) than for the 5 m distance (right panel). The original approach of Allen et al. (1977) produced the lowest strength of dereverberation for both source-receiver distances. The differences in scores between the original approach and the others were noticeably larger for the 0.5 m distance than for 5 m. This indicates
that for very close sound sources, the other methods are more efficient than the original IC approach.

The loss of quality of the signals processed with the proposed IC-based method were found to be substantially smaller for the 0.5 m condition than for the 5 m condition. This difference is not as large with the original approach as well as the spectral subtraction method, indicating that the proposed IC-based method is particularly successful for very close sound sources. As in the objective quality evaluation, increasing the degree of dereverberation processing (i.e., by decreasing \( k_p \)) results in a drop of the overall quality. However, this effect is not as prominent when decreasing \( k_p \) from 0.35 to 0.2 at the 0.5 m distance.

Considering both subjective measures, the proposed method with \( k_p = 0.2 \) clearly exhibits the best overall performance at the 0.5 m distance. Even when applying the highest degree of processing (i.e., \( k_p = 0.01 \)), the quality is similar to that obtained with spectral subtraction but the strength of dereverberation is substantially higher. For the 5 m distance, increasing the degree of processing has a negligible effect on the strength of dereverberation but is detrimental for the quality. However, for \( k_p = 0.35 \), the performance of the proposed method is comparable to that obtained with the spectral subtraction approach.

An analysis of variance (ANOVA) showed significance for the sample effect at source-receiver distances of 0.5 m \([F = 97.65, p < 0.001]\) and 5 m \([F = 41.31, p < 0.001]\). No significant subject effect was found.

V. DISCUSSION

According to the subjective results of the present study, the proposed method outperformed the two reference methods in all conditions. The original IC-based (reference) method proposed by Allen et al. (1977) did not provide any substantial effect on the considered signals and consequently resulted in very low dereverberation scores and very high quality scores. The spectral subtraction-based dereverberation method based on Lebart et al. (2001) generally provided a significant amount of dereverberation but always reduced the overall quality. For the 0.5 m distance, the proposed method provided the strongest dereverberation effect as well as best quality for all processing degrees \( (k_p) \). In the 5 m condition, the proposed method slightly outperformed the reference methods, both in terms of dereverberation and quality, but only for the lowest processing degree \( (k_p = 0.35) \).

The subjective evaluation method employed here is particularly sensitive to small differences between processing methods. However, the subjective data for the 0.5 and 5 m conditions cannot directly be compared because they are presented with different unprocessed reference signals. Due to the substantially different characteristics in the two conditions, a simultaneous presentation would result in scores at either end of the scale, which is known as compression bias (Bech and Zahorik, 2006). For comparisons on an absolute scale, the objective measures applied here are more suitable.

When comparing the objective results between the 0.5 and the 5 m conditions from Fig. 6, the strength of dereverberation (i.e., segSRR) was higher for all methods in the nearer condition. In terms of quality loss (NMR difference), all algorithms performed better in the 0.5 m condition. There are two main reasons for the differences between the 0.5 and 5 m conditions. First, at 0.5 m, where the DRR is substantially higher than at 5 m, the amount of required processing is lower, resulting in a signal of higher quality. Second, the high coherence arising from the direct sound and the early reflections is distinguishable from the diffuse sound field with low coherence [Fig. 5(b)], i.e., a bimodal coherence distribution can be observed. Considering the narrow coherence distribution for the 5 m condition in Fig. 5(c), no high coherence values are present that clearly separate the direct and the diffuse field.

A good overall correspondence of the subjective and objective results was found (Sec. IV B). Considering the strength of dereverberation, the segSRR slightly underpredicted the effectiveness of the proposed approach when compared to the subjective results. A likely reason is that the subjects used cues for reverberation estimation that are not reflected in the objective measures. For instance, when using the original implementation of the segSRR without thresholding, a very poor correlation with the subjective data was found. This is because the contribution from non-reverberant frames substantially alter the segSRR estimates. When the thresholding was introduced, the correspondence with the perceptual results increased dramatically. However, additional modifications or different methods need to be derived to further improve correspondence between subjective and objective results. In the quality evaluation, the NMR seemed to overestimate the distortion and artifacts introduced by the proposed method at 0.5 m and to underestimate them at 5 m. Moreover, the subjects showed higher sensitivity to the distortions and artifacts produced by the proposed method than the NMR measure. As pointed out by Tsilfidis and Mourjopoulos (2011), none of the quality measures (including the NMR measure) was developed to cope specifically with dereverberation and the artifacts introduced by such processing. Generally, none of the commonly applied objective quality measures are well correlated with subjective scores (Wen et al., 2006).

From the results of the present study, it can be concluded that the effectiveness of the proposed approach strongly depends on the coherence distribution in a given acoustical scenario and the applied coherence-to-gain mapping. The coherence estimation mainly depends on the window length of the STFT analysis and the recursion constant \( \alpha \). A window length consistent with literature was chosen here, but this could perhaps be optimized. The temporal resolution is reflected in the recursion constant \( \alpha \) [Eq. (5)], which here was also chosen according to the relevant literature. Lowering the integration time (decreasing the recursion constant) increases the noisiness of the coherence estimates and results in a higher limit for the lowest obtainable coherence values. This effectively reduces the processing range of the dereverberation algorithm and thus, its effectiveness. If larger integration times were chosen, the spread of coherence would be lost, again reducing the effective processing range. An alternative approach, for instance, would be to change...
the recursion constant dynamically. As in dynamic-range compression (e.g., Kates, 2008), the concept of an attack time and release time could be adopted to improve the temporal resolution at signal onsets while maintaining robust coherence estimates in case of signal decays.

The proposed coherence-to-gain mapping had a substantial effect on the performance both for dereverberation and quality (see Sec. IV). For close source-receiver distances, a high processing degree should be applied for best performance (e.g., $k_p = 0.01$). For larger distances, the processing degree should be decreased (i.e., increasing $k_p$). Hence the $k_p$ value should adapt based on source-receiver distance, which should be considered in future algorithm improvements. With reference to Fig. 5, the average coherence across frequency seems to correlate well with source-receiver distance and thus may be used as a measure for automatically adjusting the value of $k_p$. However, other source-receiver distance measures may be even more appropriate for controlling $k_p$ (Vesa, 2009).

Roman and Woodruff (2011) investigated intelligibility with ideal binary masks (IBMs) applied to reverberant speech both in noise and concurrent speech. They found significant improvements in intelligibility especially when reverberation and noise were suppressed while early reflections were preserved. The IBMs, however, require a priori information about the time-frequency representation of the reverberation and noise. With reference to the proposed coherence-based method, for very low values of $k_p$ and narrow distributions of IC, the mapping steepens and it resembles a binary mask. In future studies, IC could be used as a measure for determining time-frequency bins in a binary mask framework.

The coherence-to-gain mapping was directly defined by the histograms and only the slope was controlled by the single free parameter $k_p$. However, shifting the function may allow better tuning of the coherence-to-gain mapping relative to the IC histograms and, thus, may further improve performance. This could be an effective addition to the processing proposed here. Furthermore, the shape of the mapping function could be adapted based on the current coherence distribution. The sigmoidal parameters are currently updated at a rate of $t_{sig} = 3\ s$. However, in some acoustic scenarios, the coherence distribution may change at a different rate. Hence, $t_{sig}$ may need to be changed or controlled by a measure of the changes in the overall coherence statistics.

VI. SUMMARY AND CONCLUSION

An interaural-coherence based dereverberation method was proposed. The method applies a sigmoidal coherence-to-gain mapping function that is frequency dependent. This mapping is controlled by an (online) estimate of the present interaural coherence statistics that allows an automatic adaptation to a given acoustic scenario. By varying the overall processing degree with the parameter $k_p$, a trade-off between the amount of dereverberation and sound quality can be adjusted. The objective measures segSRR and NMR were applied and compared to subjective scores associated with “amount of reverberation” and “overall quality,” respectively. The objective and the subjective evaluation methods showed that when a significant spread in coherence is provided by the binaural input signals, the proposed dereverberation method exhibits superior performance compared to existing methods both in terms of reverberation reduction and overall quality.

ACKNOWLEDGMENTS

The authors would like to thank Dr. A. Tsilfidis (University of Patras, Greece) for his contribution to the evaluation of the dereverberation methods. This work was supported by an International Macquarie University Research Excellence Scholarship (IMQRES) and Widex A/S.

APPENDIX MEASURING BINAURAL IMPULSE RESPONSES

To evaluate the coherence as a function of source-receiver distance, binaural room impulse responses (BRIRs) were recorded in an auditorium using a Bruel & Kjaer head and torso simulator (HATS) in conjunction with a computer running MATLAB for playback and recording. The auditorium had a reverberation time of $T_{50} = 1.9\ s$ at $2\ kHz$ and a volume of $1150\ m^3$. The corresponding reverberation distance is $1.4\ m$ (see Kuttruff, 2000).

A DynAudio BM6P two-way loudspeaker was used as the sound source. This speaker-type was chosen to roughly approximate the directivity pattern of a human speaker while providing an appropriate signal-to-noise ratio. The BRIRs were measured using logarithmic upward sweeps (for details, see Müller and Massarani, 2001). Anechoic speech samples with a male speaker (taken from Hansen and Munch, 1991) were convolved with the BRIRs to simulate reverberant signals.


Acoustical Society of America. Journal

Title Details

**Basic Description**

<table>
<thead>
<tr>
<th>Title</th>
<th>Acoustical Society of America. Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISSN</td>
<td>0001-4966</td>
</tr>
<tr>
<td>Publisher</td>
<td>Acoustical Society of America</td>
</tr>
<tr>
<td>Country</td>
<td>United States</td>
</tr>
<tr>
<td>Status</td>
<td>Active</td>
</tr>
<tr>
<td>Start Year</td>
<td>1929</td>
</tr>
<tr>
<td>Frequency</td>
<td>Monthly</td>
</tr>
<tr>
<td>Volume Ends</td>
<td>Jun - Dec</td>
</tr>
<tr>
<td>Language of Text</td>
<td>Text in: English</td>
</tr>
<tr>
<td>Refereed</td>
<td>Yes</td>
</tr>
<tr>
<td>Abstracted / Indexed</td>
<td>Yes</td>
</tr>
<tr>
<td>Serial Type</td>
<td>Journal</td>
</tr>
<tr>
<td>Content Type</td>
<td>Academic / Scholarly</td>
</tr>
<tr>
<td>Format</td>
<td>Print</td>
</tr>
<tr>
<td>Website</td>
<td><a href="http://asadl.org/jasa/">http://asadl.org/jasa/</a></td>
</tr>
<tr>
<td>Description</td>
<td>Covers all phases of research and engineering of interest to acoustical scientists and engineers like linear and nonlinear acoustics; aeroacoustics, underwater sound and acoustical oceanography; ultrasonics and quantum acoustics; architectural and structural acoustics and vibration; speech, music and noise; psychology and physiology of hearing; engineering acoustics, sound transducers and measurements; bioacoustics, animal bioacoustics and bioreponse to vibration.</td>
</tr>
</tbody>
</table>

**Subject Classifications**

**Additional Title Details**

**Publisher & Ordering Details**

**Price Data**

**Abstracting & Indexing**

**Other Availability**

**Demographics**

**Reviews**