Cauchi *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:61
DOI 10.1186/s13634-015-0242-x

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech

Benjamin Cauchi[1,3*], Ina Kodrasi[2,3], Robert Rehr[2,3], Stephan Gerlach[1,3], Ante Jukić[2,3], Timo Gerkmann[2,3], Simon Doclo[1,2,3] and Stefan Goetze[1,3]

Abstract

This paper presents a system aiming at joint dereverberation and noise reduction by applying a combination of a beamformer with a single-channel spectral enhancement scheme. First, a minimum variance distortionless response beamformer with an online estimated noise coherence matrix is used to suppress noise and reverberation. The output of this beamformer is then processed by a single-channel spectral enhancement scheme, based on statistical room acoustics, minimum statistics, and temporal cepstrum smoothing, to suppress residual noise and reverberation. The evaluation is conducted using the REVERB challenge corpus, designed to evaluate speech enhancement algorithms in the presence of both reverberation and noise. The proposed system is evaluated using instrumental speech quality measures, the performance of an automatic speech recognition system, and a subjective evaluation of the speech quality based on a MUSHRA test. The performance achieved by beamforming, single-channel spectral enhancement, and their combination are compared, and experimental results show that the proposed system is effective in suppressing both reverberation and noise while improving the speech quality. The achieved improvements are particularly significant in conditions with high reverberation times.

Keywords: REVERB challenge; Dereverberation; Noise reduction; Beamforming; Spectral enhancement

1 Introduction

In many speech communication applications, such as voice-controlled systems or hearing aids, distant microphones are used to record a target speaker. The microphone signals are often corrupted by both reverberation and noise, resulting in a degraded speech quality and speech intelligibility, as well as in a reduced performance of automatic speech recognition (ASR) systems.

Several algorithms have been proposed in the literature to deal with these issues (cf. [1–3] and the references therein). This paper extends the description and evaluation of the system proposed by the authors in [4], which consists of a commonly used combination of a minimum variance distortionless response (MVDR) beamformer with a single-channel spectral enhancement scheme. In such a combined system, the spectral enhancement scheme typically consists in applying a real-valued spectral gain to the short-time Fourier transform (STFT) of the beamformer output. The computation of this spectral gain relies on estimates of the power spectral densities (PSDs) of the interference to be suppressed, i.e., noise and reverberation, as early reflections are often considered to be beneficial both in terms of speech quality [5] and ASR performance [6].

Different methods have been proposed for estimating the late reverberant and noise PSDs, e.g. relying on assumptions about the sound field or on a voice activity detector (VAD). The PSDs of the noise and reverberation can be estimated using the output signal(s) of a blocking matrix, suppressing the signal to be preserved, in the well-known generalized sidelobe canceller (GSC) structure. The blocking matrix can be designed, e.g., as a delay-and-subtract beamformer cancelling the direct speech component [7, 8] or based on a blind source separation

*Correspondence: benjamin.cauchi@idmt.fraunhofer.de
[1] Fraunhofer IDMT, Hearing Speech and Audio Technology, 26129 Oldenburg, Germany
[3] Cluster of Excellence Hearing4all, Oldenburg, Germany
Full list of author information is available at the end of the article

(BSS) scheme aiming to cancel both the direct speech component and the early reflections [9, 10]. Alternatively, the PSD at a reference position can be obtained using a maximum likelihood estimator (MLE) and a model of the sound field [11]. The PSD to be used in the computation of the spectral postfilter is then obtained by correcting the estimated PSD at the reference position. This correction can be done using an adaptive filter [8], back-projection [9, 10], or the relative transfer functions between the target speaker and the microphones [11].

Other methods estimate the PSD of the interference from the output of the beamformer and thus can in principle also be used if only one microphone is available. In such methods [4, 12], the estimation of the noise PSD is often derived from statistical models of the speech and noise [13, 14]. The estimation of the reverberant PSD can, e.g., be derived from a statistical model of the room impulse response (RIR) and the acoustical properties of the room, such as the reverberation time (T_{60}) or the direct-to-reverberant ratio (DRR) [15, 16].

In the system presented in this paper, the microphone signals are first processed using an MVDR beamformer [17], which aims to suppress sound sources not arriving from the direction of arrival (DOA) of the target speaker, while maintaining a unit gain towards this DOA. The noise coherence matrix used to compute the coefficients of the MVDR beamformer is estimated online using a VAD [18], and the DOA of the target speaker is estimated using the multiple signal classification (MUSIC) algorithm [19, 20]. The beamformer output is processed using a single-channel spectral enhancement scheme, which aims at jointly suppressing the residual noise and reverberation. The main novel contribution of this paper is the combination of the several estimators used in the single-channel spectral enhancement scheme. This spectral enhancement scheme relies on estimates of the PSDs of the noise and the late reverberation, similarly as in [21]. The proposed scheme computes a real-valued spectral gain, combining the clean speech amplitude estimator presented in [22], the noise PSD estimator based on minimum statistics (MS) [13], and an estimator of the (late) reverberant PSD based on statistical room acoustics [15, 23]. In order to reduce the musical noise which is often a byproduct of spectral enhancement schemes, adaptive smoothing in the cepstral domain is used to estimate the speech PSD [24, 25].

The proposed system is evaluated using the REVERB challenge corpus [26], which permits the evaluation of algorithms under realistic conditions in single- and multi-channel scenarios. The single-channel scenario is particularly challenging as illustrated by the results of the REVERB challenge workshop [27], in which most contributions succeeded to reduce reverberation but only a few improved the speech quality [4, 12]. The evaluation is conducted for different configurations of the proposed system in terms of instrumental speech quality measures, improvement of ASR performance, and a subjective evaluation of speech quality and dereverberation using a MUSHRA test [28]. The evaluation results show that the proposed system is able to reduce noise and reverberation while improving the speech quality in both single- and multi-channelscenarios.

This paper is organized as follows. In Section 2, an overview of the proposed system is given. Details about the proposed MVDR beamformer and the single-channel spectral enhancement scheme are presented in Section 3 and in Section 4, respectively. The evaluation corpus is briefly described in Section 5 and the evaluation results are presented in Section 6.

2 System overview

When recording a single speech source in an enclosure using M microphones, the reverberant and noisy mth microphone signal $y_m(n)$ at time index n is given by

$$y_m(n) = s(n) * h_m(n) + v_m(n) \tag{1}$$

$$= x_m(n) + v_m(n), \text{ for } m = 1, \cdots, M, \tag{2}$$

with $s(n)$ denoting the clean speech signal, $h_m(n)$ denoting the RIR between the speech source and the mth microphone, and $x_m(n)$ and $v_m(n)$ denoting the reverberant speech component and the additive noise component in the mth microphone signal, respectively. The STFT representations of $y_m(n)$, $s(n)$, $x_m(n)$, and $v_m(n)$ are denoted by $Y_m(k,\ell)$, $S(k,\ell)$, $X_m(k,\ell)$, and $V_m(k,\ell)$, respectively, with k and ℓ representing the discrete frequency bin and frame indices, respectively.

The proposed system, depicted in Fig. 1, aims at obtaining an estimate $\hat{s}(n)$, with $\hat{\cdot}$ denoting estimated quantities, of the clean speech signal $s(n)$ from the reverberant and noisy microphone signals, $y_m(n)$. This system consists of two stages. First, an MVDR beamformer is applied to the microphone signals. This beamformer aims at reducing noise and reverberation by suppressing the sound sources not arriving from the target DOA, while providing a unity gain in the direction of the target speaker. The noise coherence matrix and the DOA used to compute the MVDR beamformer coefficients are estimated from the received microphone signals $y_m(n)$. The noise coherence matrix is estimated using a VAD [18], whereas the DOA estimation is based on the MUSIC algorithm [19, 20], cf. Section 3. In order to suppress the residual noise and reverberation at the beamformer output $\tilde{x}(n)$, the beamformer output is processed by a single-channel spectral enhancement scheme, cf. Section 4.

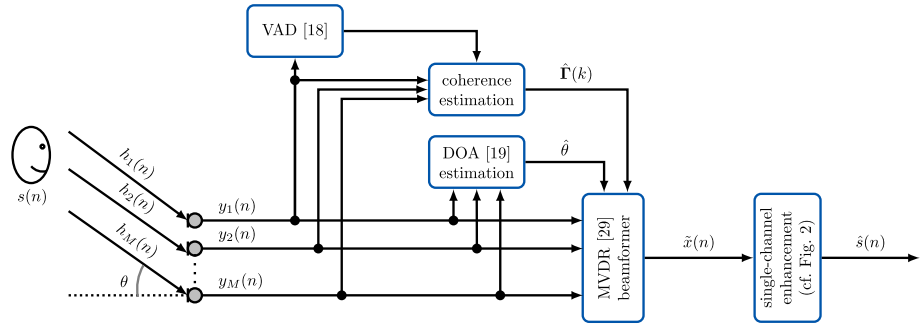

Fig. 1 Overview of the proposed system

3 Beamformer

3.1 MVDR beamforming

In the STFT domain, (2) can be expressed as

$$Y_m(k,\ell) = X_m(k,\ell) + V_m(k,\ell), \text{ for } m = 1,\cdots,M, \quad (3)$$

which in vector notation can be written as

$$\mathbf{Y}(k,\ell) = \mathbf{X}(k,\ell) + \mathbf{V}(k,\ell), \quad (4)$$

with

$$\mathbf{Y}(k,\ell) = [\, Y_1(k,\ell)\ Y_2(k,\ell)\ \ldots\ Y_M(k,\ell)\,]^T, \quad (5)$$

denoting the M-dimensional stacked vector of the received microphone signals and $\mathbf{X}(k,\ell)$ and $\mathbf{V}(k,\ell)$ denoting the stacked vectors of the reverberant speech component and noise component, respectively, defined in the same way as in (5).

In the STFT domain, the beamformer output signal $\tilde{x}(n)$ is denoted by $\tilde{X}(k,\ell)$ and obtained by filtering and summing the microphone signals, i.e.,

$$\begin{aligned}\tilde{X}(k,\ell) &= \mathbf{W}_\theta^H(k)\mathbf{Y}(k,\ell) \\ &= \mathbf{W}_\theta^H(k)\mathbf{X}(k,\ell) + \mathbf{W}_\theta^H(k)\mathbf{V}(k,\ell),\end{aligned} \quad (6)$$

with $\mathbf{W}_\theta(k)$ denoting the stacked filter coefficient vector of the beamformer steered towards the angle θ.

Aiming at minimizing the noise power while providing a unity gain in the direction of the target speaker, the filter coefficients of the MVDR beamformer are computed as [17]

$$\mathbf{W}_\theta(k) = \frac{\Gamma^{-1}(k)\mathbf{d}_\theta(k)}{\mathbf{d}_\theta^H(k)\Gamma^{-1}(k)\mathbf{d}_\theta(k)}, \quad (7)$$

where $\mathbf{d}_\theta(k)$ and $\Gamma(k)$ denote the steering vector of the target speaker and the noise coherence matrix, respectively. Using a far-field assumption, the steering vector $\mathbf{d}_\theta(k)$ is equal to

$$\mathbf{d}_\theta(k) = \left[e^{-j2\pi f_k \tau_1(\theta)}\ e^{-j2\pi f_k \tau_2(\theta)}\ \ldots e^{-j2\pi f_k \tau_M(\theta)}\right], \quad (8)$$

with f_k denoting the center frequency of frequency bin k and $\tau_m(\theta)$ denoting the time difference of arrival of the source at angle θ between the mth microphone and a reference position, which has been arbitrarily chosen as the center of the microphone array.

To compute the MVDR beamformer filter coefficients, an estimate $\hat{\theta}$ of the DOA of the target speaker as well as an estimate of the noise coherence matrix is required.

3.2 Noise coherence matrix estimation

The noise coherence matrix is estimated during noise-only periods detected using the VAD described in [18], as the covariance matrix of the noise-only components, i.e.

$$\hat{\Gamma}(k) = \frac{1}{\overline{\mathbb{L}}_\nu} \sum_{\ell \in \mathbb{L}_\nu} \mathbf{V}(k,\ell)\mathbf{V}^H(k,\ell), \quad (9)$$

with $\mathbb{L}_\nu$ denoting the set of detected noise-only frames and $\overline{\mathbb{L}}_\nu$ its cardinality.

However, if the detected noise-only period is too short for a reliable estimate (cf. Section 5), the coherence matrix $\overline{\Gamma}(k)$ of a diffuse noise field is used instead, i.e., the coherence between two microphones i and i', separated by a distance $l_{i,i'}$, is computed as

$$\overline{\Gamma}_{i,i'}(k) = \frac{\sin\left(2\pi f_k l_{i,i'}/c\right)}{2\pi f_k l_{i,i'}/c}, \quad (10)$$

with c denoting the speed of sound, resulting in the well-known superdirective beamformer [17]. Additionally, a white noise gain constraint $\mathrm{WNG}_{\max}$ is imposed in order to limit the potential amplification of uncorrelated noise, especially at low frequencies. With such a constraint, the used noise coherence matrix is equal to

$$\hat{\Gamma}(k) = \overline{\Gamma}(k) + \varrho(k)\mathbf{I}_M, \quad (11)$$

with $\mathbf{I}_M$ denoting the $M \times M$-dimensional identity matrix and $\varrho(k)$ denoting a frequency-dependent regularization parameter which is computed iteratively such that $\mathbf{W}_\theta^H(k)\mathbf{W}_\theta(k) \leq \mathrm{WNG}_{\max}$ [29].

3.3 DOA estimation

As the beamformer aims at suppressing sources not arriving from the target DOA, an error in the DOA estimate

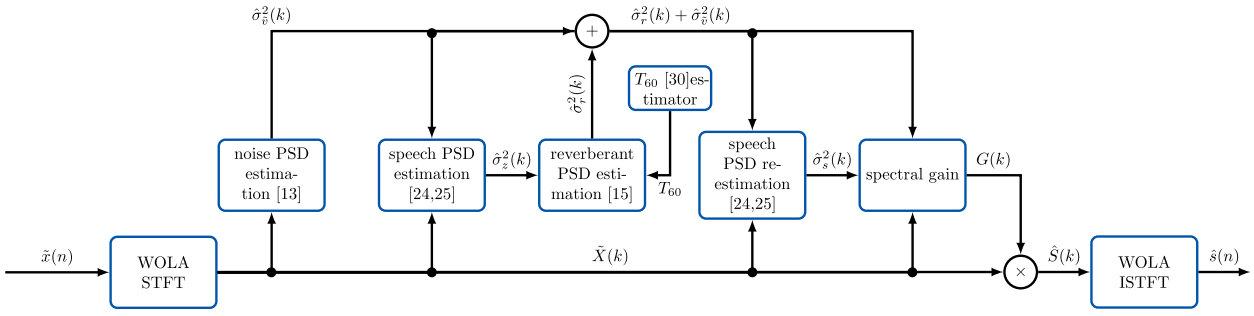

Fig. 2 Overview of the proposed single-channel enhancement scheme for a single frame

may lead to suppression of the desired source by the beamformer. In the proposed system, the subspace-based MUSIC algorithm [19, 20], shown robust in our target application (cf. Section 6.1), has been used to compute the DOA estimate $\hat{\theta}$.

Assuming that speech and noise are uncorrelated, the steering vector corresponding to the true DOA is orthogonal to the noise subspace, which is represented by an $M \times (M-Q)$-dimensional matrix, with Q the number of sources (i.e., $Q = 1$ in this case), defined as

$$\boldsymbol{E}(k,\ell) = \left[\boldsymbol{e}_{Q+1}(k,\ell) \dots \boldsymbol{e}_M(k,\ell)\right]. \tag{12}$$

The noise subspace $\boldsymbol{E}(k,\ell)$ is composed of the eigenvectors of the covariance matrix of $\mathbf{Y}(k,\ell)$ corresponding to the $(M-Q)$ smallest eigenvalues.

The MUSIC algorithm then estimates the DOA as the angle maximizing the sum of the MUSIC pseudo-spectra

$$U_\theta(k,\ell) = \frac{1}{\mathbf{d}_\theta^H(k)\boldsymbol{E}(k,\ell)\boldsymbol{E}^H(k,\ell)\mathbf{d}_\theta(k)}, \tag{13}$$

over a given frequency range, i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{K} \sum_{k_{\text{low}}}^{k_{\text{high}}} U_\theta(k,\ell), \tag{14}$$

with K denoting the total number of considered frequency bins $k = k_{\text{low}} \dots k_{\text{high}}$.

4 Single-channel spectral enhancement

Although the beamformer in Section 3.1 is able to reduce the interference, i.e., noise and reverberation, to some extent, spectral enhancement schemes are able to further reduce reverberation as well as noise. The output signal $\tilde{X}(k,\ell)$ of the MVDR beamformer contains the clean speech signal $S(k,\ell)$ as well as residual reverberation $R(k,\ell)$ and residual noise $\tilde{V}(k,\ell)$, i.e.

$$\tilde{X}(k,\ell) = Z(k,\ell) + \tilde{V}(k,\ell), \tag{15}$$

with

$$Z(k,\ell) = S(k,\ell) + R(k,\ell) \tag{16}$$

the reverberant speech component. Aiming at jointly reducing residual reverberation and noise, the single-channel spectral enhancement scheme summarized in Fig. 2 is proposed, where a real-valued spectral gain $G(k,\ell)$ is applied to the STFT coefficients of the beamformer output, i.e.,

$$\hat{S}(k,\ell) = G(k,\ell)\tilde{X}(k,\ell), \tag{17}$$

with $\hat{S}(k,\ell)$ denoting the STFT of the estimated speech signal.

The spectral gain $G(k,\ell)$ is computed using the minimum mean square error (MMSE) estimator for the clean speech spectral magnitude as proposed in [22] (cf. Section 4.1). This estimator, similarly to the Wiener filter, requires the PSDs of the clean speech, the noise, and the reverberation components.

First, an estimate $\hat{\sigma}_{\tilde{v}}^2(k,\ell)$ of the noise PSD is obtained based on a slight modification of the well-known minimum statistics (MS) approach [13] (cf. Section 4.2) and used to estimate the reverberant speech PSD. The estimate $\hat{\sigma}_z^2(k,\ell)$ of the reverberant speech PSD is computed using temporal cepstrum smoothing [24, 25] (cf. Section 4.3). The estimate $\hat{\sigma}_r^2(k,\ell)$ of the (late) reverberant PSD is computed from the reverberant speech PSD estimate using the approach proposed in [15] (cf. Section 4.4). This approach requires an estimate of the reverberation time T_{60}, which has been obtained using the estimator described in [30]. As the dereverberation task is treated separately from the denoising task, care has to be taken that no reverberation leaks into the noise PSD estimate and vice versa. Thus, a longer minimum search window is used in the MS approach as compared to [13] (cf. Section 5.2).

The estimate $\hat{\sigma}_s^2(k,\ell)$ of the clean speech PSD is finally obtained by a re-estimation, again using temporal cepstrum smoothing. The following subsections give a more detailed description of the different components of the proposed single-channel spectral enhancement scheme.

4.1 Spectral gain

The gain function used in the spectral enhancement scheme has been proposed in [22] to estimate the spectral magnitude of the clean speech. This estimator is derived by modeling the speech magnitude $|S(k,\ell)|$ as a stochastic variable with a chi probability density function (pdf) with shape parameter μ, while the phase of $S(k,\ell)$ is assumed to be uniformly distributed between $-\pi$ and π. Furthermore, the interference $J(k,\ell) = R(k,\ell) + \tilde{V}(k,\ell)$ is modeled as a complex Gaussian random variable with PSD $\sigma_j^2(k,\ell)$. Assuming that $R(k,\ell)$ and $\tilde{V}(k,\ell)$ are uncorrelated, $\sigma_j^2(k,\ell)$ can be expressed as

$$\sigma_j^2(k,\ell) = \mathrm{E}\left\{|J(k,\ell)|^2\right\} = \sigma_{\tilde{v}}^2(k,\ell) + \sigma_r^2(k,\ell), \tag{18}$$

with $\sigma_r^2(k,\ell)$ and $\sigma_{\tilde{v}}^2(k,\ell)$ denoting the PSDs of the reverberation and of the noise, respectively.

The squared distance between the amplitudes (to the power β) of the clean speech $S(k,\ell)$ and the estimated output $\hat{S}(k,\ell)$ is defined as

$$\epsilon(k,\ell) = \left(|S(k,\ell)|^\beta - |\hat{S}(k,\ell)|^\beta\right)^2. \tag{19}$$

The parameter β, typically chosen as $0 < \beta \leq 1$, is a compression factor resulting in a different emphasis given on estimation errors for small amplitudes in relation to large amplitudes. The clean speech magnitude is estimated by optimizing the MMSE criterion

$$\left|\hat{S}(k,\ell)\right| = \underset{\left|\hat{S}(k,\ell)\right|}{\operatorname{argmin}} \mathrm{E}\left\{\epsilon(k,\ell)|\tilde{X}(k,\ell), \sigma_j^2(k,\ell), \xi(k,\ell)\right\}, \tag{20}$$

with $\xi(k,\ell)$ denoting the a priori signal-to-interference ratio (SIR) defined as

$$\xi(k,\ell) = \frac{\sigma_s^2(k,\ell)}{\sigma_r^2(k,\ell) + \sigma_{\tilde{v}}^2(k,\ell)}, \tag{21}$$

with $\sigma_s^2(k,\ell)$ denoting the PSD of the clean speech.

As shown in [22], the solution to (20) leads to the spectral gain $\tilde{G}(k,\ell)$

$$\tilde{G}(k,\ell) = \sqrt{\frac{\xi(k,\ell)}{\mu + \xi(k,\ell)}} \cdot \left[\frac{\mathrm{Gam}\left(\mu + \frac{\beta}{2}\right)}{\mathrm{Gam}\left(\mu\right)} \frac{\Phi\left(1 - \mu - \frac{\beta}{2}, 1; -\nu(k,\ell)\right)}{\Phi\left(1 - \mu, 1; -\nu(k,\ell)\right)}\right]^{1/\beta} \cdot \left(\sqrt{\gamma(k,\ell)}\right)^{-1}, \tag{22}$$

with $\gamma(k,\ell)$ denoting the a posteriori SIR, defined as

$$\gamma(k,\ell) = \frac{|\tilde{X}(k,\ell)|^2}{\sigma_r^2(k,\ell) + \sigma_{\tilde{v}}^2(k,\ell)}, \tag{23}$$

and

$$\nu(k,\ell) = \frac{\gamma(k,\ell)\xi(k,\ell)}{\mu + \xi(k,\ell)}, \tag{24}$$

with $\Phi(\cdot)$ denoting the confluent hypergeometric function and $\mathrm{Gam}(\cdot)$ denoting the complete Gamma function [31]. Depending on the choice of β and μ, the solution in (22) can resemble other well-known estimators, such as the short-time spectral amplitude estimator ($\beta = 1$, $\mu = 1$) [32] or the log-spectral amplitude estimator ($\beta = 0$, $\mu = 1$) [33]. In order to reduce artifacts which may be introduced by directly applying (22), the spectral gain $G(k,\ell)$ in (17) is restricted to values larger than a spectral floor $G_{\min}$ (cf. Section 5.2), i.e.,

$$G(k,\ell) = \max\left(\tilde{G}(k,\ell), G_{\min}\right). \tag{25}$$

To compute the expression in (22), the PSDs $\sigma_s^2(k,\ell)$, $\sigma_{\tilde{v}}^2(k,\ell)$, and $\sigma_r^2(k,\ell)$ have to be estimated from the beamformer output. The used estimators are described in the next subsections.

4.2 Noise PSD estimator

The MS [13] approach has been shown to be a reliable estimator of the noise PSD for moderately time-varying noise conditions. This approach relies on the assumption that the minimum of the noisy speech power, $P_{\tilde{x}}(k,\ell)$, over a short temporal sliding window is not affected by the speech. The noise PSD $\sigma_{\tilde{v}}^2(k,\ell)$ is then estimated by tracking the minimum of $P_{\tilde{x}}(k,\ell)$ over this sliding window, whose usual length corresponds to 1.5 s according to [13].

Figure 3 depicts the powers of anechoic speech, reverberant speech, and additive noise for one frequency bin of their power spectrograms. As illustrated in this figure, the decay time in speech pauses is typically increased in the presence of reverberation. Consequently, a longer tracking window is used in the proposed spectral enhancement scheme (cf. Section 5) in order to avoid reverberant speech affecting the estimation of the noise PSD $\sigma_{\tilde{v}}^2(k,\ell)$.

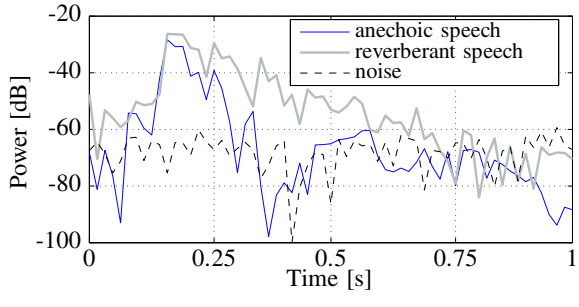

Fig. 3 Power of anechoic speech, reverberant speech, and additive noise at a frequency of 500 Hz for a 1-s signal extracted from the REVERB challenge corpus for a room of T_{60} =0.73 s and a distance of 2 m between the speech source and the microphone

4.3 Speech PSD estimator

Temporal cepstrum smoothing, as proposed in [24], is used to estimate the PSD $\sigma_z^2(k,\ell)$ of the reverberant speech component $Z(k,\ell)$ as well as the PSD $\sigma_s^2(k,\ell)$ of the dereverberated speech signal $S(k,\ell)$. The estimation of $\sigma_z^2(k,\ell)$ only requires the noise PSD estimate $\hat{\sigma}_{\tilde{v}}^2(k,\ell)$ whereas the estimation of $\sigma_s^2(k,\ell)$ additionally requires an estimate of the reverberant PSD $\sigma_r^2(k,\ell)$, as depicted in Fig. 2. The modifications required for the latter case are described at the end of this section.

In order to estimate the reverberant speech PSD $\sigma_z^2(k,\ell)$, the maximum likelihood (ML) estimator of the a priori signal to noise ratio (SNR)

$$\xi_{z_{\mathrm{ml}}}(k,\ell) = \frac{|\tilde{X}(k,\ell)|^2}{\sigma_{\tilde{v}}^2(k,\ell)} - 1 \tag{26}$$

is employed. An estimate $\hat{\sigma}_{z_{\mathrm{ml}}}^2(k,\ell)$ of the reverberant speech PSD can then be obtained as

$$\hat{\sigma}_{z_{\mathrm{ml}}}^2(k,\ell) = \hat{\sigma}_{\tilde{v}}^2(k,\ell)\ \max\left(\xi_{z_{\mathrm{ml}}}(k,\ell),\ \xi_{\mathrm{ml}}^{\min}\right), \tag{27}$$

with $\xi_{\mathrm{ml}}^{\min} > 0$ denoting a lower bound to avoid negative or very small values of $\xi_{z_{\mathrm{ml}}}(k,\ell)$.

In the cepstral domain, $\hat{\sigma}_{z_{\mathrm{ml}}}^2(k,\ell)$ can be represented by

$$\lambda_{z_{\mathrm{ml}}}(q,\ell) = \mathrm{IFFT}\left\{\log\left(\hat{\sigma}_{z_{\mathrm{ml}}}^2(k,\ell)|_{k=0,\cdots,(L-1)}\right)\right\}, \tag{28}$$

with q denoting the cepstral bin index and L denoting the length of the FFT. A recursive temporal smoothing is applied to $\lambda_{z_{\mathrm{ml}}}(q,\ell)$, i.e.,

$$\lambda_z(q,\ell) = \delta(q,\ell)\lambda_z(q,\ell-1) + (1-\delta(q,\ell))\lambda_{z_{\mathrm{ml}}}(q,\ell), \tag{29}$$

with $\delta(q,\ell)$ denoting a time-quefrency-dependent smoothing parameter. Only a mild smoothing is applied to the quefrencies which are mainly related to speech, while for the remaining quefrencies, a stronger smoothing is applied. Consequently, a small smoothing parameter is chosen for the low quefrencies, as they contain information about the vocal tract shape, and for the quefrencies corresponding to the fundamental frequency f_0 in voiced speech. In order to protect these quefrencies, especially the ones corresponding to the fundamental frequency, the parameter $\delta(q,\ell)$ in (29) is adapted. After determining f_0 by picking the highest peak in the cepstrum within a limited search range, $\delta(q,\ell)$ is defined as

$$\delta(q,\ell) = \begin{cases} \delta_{\mathrm{pitch}} & \text{if } q \in \mathbb{Q}, \\ \bar{\delta}(q,\ell) & \text{if } q \in \{0,\cdots,L/2\} \setminus \mathbb{Q}, \end{cases} \tag{30}$$

with $\mathbb{Q}$ denoting a small set of cepstral bins around the quefrency corresponding to f_0 and δ_{pitch} the smoothing parameter for the quefrency bins within $\mathbb{Q}$ [24]. The quantity $\bar{\delta}(q,\ell)$ is given as

$$\bar{\delta}(q,\ell) = \eta\delta(q,\ell-1) + (1-\eta)\bar{\delta}_{\mathrm{const}}(q), \tag{31}$$

where $\bar{\delta}_{\mathrm{const}}(q)$ is time independent and chosen such that less smoothing is applied in the lower cepstral bins. Furthermore, η is a forgetting factor which defines how fast the transition from $\delta(q,\ell)$ to $\bar{\delta}_{\mathrm{const}}(q)$ can occur (cf. Section 5.2). Finally, the reverberant speech PSD estimate $\hat{\sigma}_z^2(k,\ell)$ can be obtained by transforming $\lambda_z(q,\ell)$ back to the spectral domain, i.e.

$$\hat{\sigma}_z^2(k,\ell) = \exp\left(\kappa + \mathrm{DFT}\left\{\lambda_z(q,\ell)\right\}|_{q=0,\cdots,(L-1)}\right), \tag{32}$$

with κ denoting a parameter to compensate for the bias due to the recursive smoothing in the log domain in (29) and is estimated as in [25].

The estimate of the reverberant speech PSD can be used to estimate the reverberant PSD $\sigma_r^2(k,\ell)$ (cf. Section 4.4). After having estimated $\sigma_r^2(k,\ell)$, cepstral smoothing is also used to estimate the dereverberated clean speech PSD $\sigma_s^2(k,\ell)$. In this case, the noise PSD $\sigma_{\tilde{v}}^2(k,\ell)$ in (26) and (27) is replaced by the interference PSD $\sigma_j^2(k,\ell) = \sigma_{\tilde{v}}^2(k,\ell) + \sigma_r^2(k,\ell)$.

4.4 Reverberant PSD estimation

The RIR model presented in [23] represents the RIR as a Gaussian noise signal multiplied by an exponential decay Δ, which depends on the room reverberation time, T_{60}, i.e.,

$$\Delta = \frac{3\ln 10}{T_{60} f_s}. \tag{33}$$

In the proposed spectral enhancement scheme, the approach derived from this model and presented in [15] is used to estimate the reverberant PSD $\sigma_r^2(k,\ell)$ as

$$\hat{\sigma}_r^2(k,\ell) = \mathrm{e}^{-2\Delta T_d f_s}\hat{\sigma}_z^2(k,\ell - T_d/T_s). \tag{34}$$

with

$$\hat{\sigma}_z^2(k,\ell) = \hat{\sigma}_r^2(k,\ell) + \hat{\sigma}_s^2(k,\ell). \tag{35}$$

In (34), T_s denotes the frame shift whereas T_d is the duration of the direct path and early reflections of the RIR, typically assumed to be between 50 and 80 ms. As a result, the estimate $\hat{\sigma}_r^2(k,\ell)$ can be obtained using $\hat{\sigma}_z^2(k,\ell)$ and an estimate of the reverberation time T_{60} obtained using an online estimator such as the one proposed in [30].

Finally, using the estimated PSDs of the reverberation and of the residual noise, an estimate $\hat{\sigma}_s^2(k,\ell)$ of the clean speech PSD is obtained. These estimates are used in (21) to compute the a priori SIR and in (22) to compute the real-valued spectral gain, $\tilde{G}(k,\ell)$.

5 Experimental setup

5.1 Corpus description

The results presented in this paper have been obtained using the evaluation set of the REVERB challenge [26], which consists of a large corpus of speech corrupted

by reverberation and noise. All recordings have been made at a sampling frequency of 16 kHz with a circular microphone array with 20 cm diameter and 8 equidistant microphones. This corpus is divided into simulated and real data. The simulated data is composed of clean speech signals taken from the WSJCAM0 corpus [34], which have been convolved with RIRs recorded in three different rooms and to which measured noise at a fixed SNR of 20 dB have been added. The real data is composed of utterances from the MC-WSJ-AV corpus [35] and contains speech recorded in a room in the presence of noise. The utterances have been spoken from different unknown positions within each room, but the position was constant during each utterance. For each room, two distances (denoted by "near" and "far") between the target speaker and the center of the microphone array have been considered. The combination of a room and a particular distance will be refered to as "condition" in the remainder of this paper. The characteristics of each condition along with the labels used to refer to it are summarize in Table 1.

5.2 Algorithm settings

For the experiments, it has been assumed that the T_{60} and the DOA of the target speaker remain constant for each utterance. Therefore, both T_{60} and DOA have been estimated only once per utterance. The STFT has been computed using a 32-ms Hann window with 50 % overlap and an FFT of length $L = 512$. The DOA has been estimated as the angle minimizing the sum of the MUSIC pseudo-spectra, for $\theta = 0^\circ \ldots 360^\circ$ for every 2°, using all 8 microphones of the circular microphone array for the frequency range from 50 Hz to 5 kHz, cf. Section 3.3.

The MVDR beamformer uses a theoretically diffuse noise coherence matrix and a white noise gain constraint $\mathrm{WNG}_{\max} = -10$ dB if less than 10 frames are detected as noise when applying the VAD, cf. (11). The VAD has been configured similarly as in [18], but its parameters have been adapted in order to apply it to signals with a sampling frequency of 16 kHz. Otherwise, the noise coherence matrix is estimated using all detected noise-only frames, cf. (9). The speech amplitude estimator in Section 4.1 assumes a chi pdf with shape parameter $\mu = 0.5$, a minimum gain $G_{\min}$ of -10 dB, and a compression parameter $\beta = 0.5$. The noise PSD estimator described in Section 4.2 uses the same parameters as in [13], except for the length of the sliding window for minima tracking which has been set to either 1.5 s ($\mathrm{SE}_{1.5}$) or 3 s (SE_3) in our experiments. In (31), $\eta = 0.96$ and all parameters used for the speech PSD estimation, described in Section 4.3, have been set as prescribed in [22]. In (34), T_d has been set to 80 ms.

Table 1 Summary of the testing room conditions and of the labels used for presenting the results

Set	Room	T_{60} [ms]	Distance [cm]	Label
Simulated	Small	250	50	S1, near
			200	S1, far
	Medium	500	50	S2, near
			200	S2, far
	Large	700	50	S3, near
			200	S3, far
Real	Large	700	100	R1, near
			250	R1, far

6 Results

The performance of the proposed system for each condition is evaluated in terms of instrumental speech quality measures (cf. Section 6.2) as well as in terms of word error rate (WER) when using the proposed system as a preprocessing scheme for the REVERB challenge baseline ASR system (cf. Section 6.3). Additionally, the results obtained in a subjective speech quality evaluation are presented for 4 out of 8 conditions in Section 6.4.

The performance of the combined scheme is compared to the performance when applying only the single-channel spectral enhancement scheme to the first microphone signal and when applying only the MVDR beamformer to the multichannel input.

6.1 Observations on beamformer design

The MVDR beamformer used in this paper is steered towards the estimated DOA of the target speech signal. In practice, errors in the DOA estimation can result in speech degradation. Figure 4 (top) depicts the DOA error obtained in all conditions of the simulated data of the REVERB challenge (i.e., a total of 2176 utterances). The true DOA has been considered to be the one stated in the REVERB challenge data documentation [36]. Ignoring outliers, it can be seen that the absolute value of the error is smaller than 5 in room S1 while in room S2, it is smaller than 10° for 50 % of the data and always smaller than 15°. As expected, the largest error in DOA estimation appears in the case of room S3, which has the largest reverberation time. It can be seen that for room S3, in 50 % of the utterances, the absolute value of the DOA error is inferior to 15°. However, it can be as high as 28° for some utterances.

In order to assess the detrimental effect that such DOA error could have on the performance of the MVDR beamformer, one may examine its corresponding beampattern. Figure 4 (bottom) depicts the beampattern of the MVDR beamformer computed using the noise coherence matrix of a theoretically diffuse noise field as in (11), steered towards the zero degrees direction, and using the microphone configuration described in Section 5.1. By observing the width of the main lobe, it appears that the error

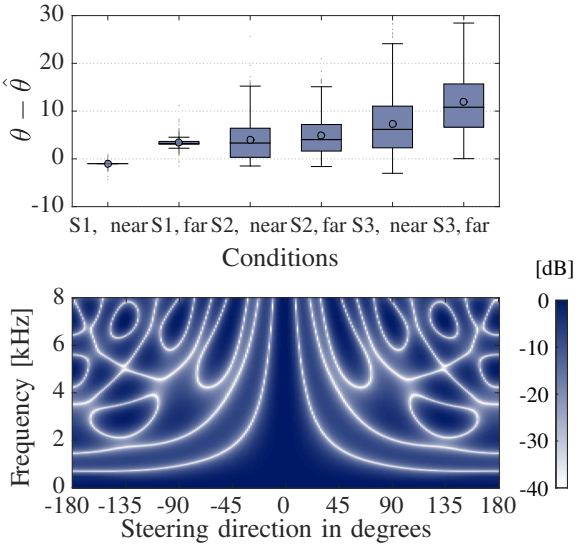

Fig. 4 Error in DOA estimation obtained on the simulated data of the REVERB challenge corpus (*top*) and beampattern of the used MVDR beamformer computed using the noise coherence matrix of a theoretically diffuse noise field (*bottom*)

in DOA is small enough to not introduce distortions in rooms S1 and S2. Some cancellation of the target speech signal may occur in room S3 but should be limited to frequencies higher than 4 kHz.

6.2 Instrumental speech quality measures

The performance in terms of instrumental speech quality measures for the different considered conditions is presented in Table 2 for the simulated data and in Table 3 for the real data. Since various instrumental speech quality measures exist which can be used to assess the quality of denoised and dereverberated signals [37–39] and since it is difficult to assess the quality using only one single measure, the performance of the proposed system has been evaluated using the five signal-based quality measures suggested in [26], i.e., the speech to reverberation modulation energy ratio (SRMR) [40], the cepstral distance (CD) [41], the log likelihood ratio (LLR) [41], the frequency-weighted segmental SNR (FWSSNR) [41], and the perceptual evaluation of speech quality (PESQ) [42]. Among these five quality measures, the SRMR is the only non-intrusive measure, i.e., not requiring a reference signal, and is hence the only measure that can be used to evaluate the performance for real data. The other measures use the clean speech signal $s(n)$ as the reference signal.

For the single-channel case, Tables 2 and 3 compare the quality of the unprocessed (first microphone) signal ("Unp." in tables) to the quality of the signal processed using the proposed spectral enhancement scheme using the standard MS window of 1.5 s ($SE_{1,5}$) as well as a longer

Table 2 Values of the instrumental speech quality measures obtained on the simulated data

	Mean results on all simulated data					
	1 channel			8 channels		
	Unp.	$SE_{1.5}$	SE_3	MVDR	MVDR $+SE_{1.5}$	MVDR $+SE_3$
SRMR [dB]	3.68	*4.42*	4 .39	4.22	*5.01*	4.97
CD [dB]	3.98	3.64	*3.58*	3.81	3.48	*3.41*
LLR	0.58	0.58	*0.57*	*0.59*	0.61	0.6
FWSSNR [dB]	3.62	5.76	*5.92*	4.56	7.1	*7.26*
PESQ	1.48	1.66	*1.67*	1.84	2.03	*2.05*
			S1, near			
SRMR [dB]	4.5	*4.97*	4.95	6.39	*7.21*	7.16
CD [dB]	*1.99*	2.27	2.22	*2.47*	2.65	2.59
LLR	*0.35*	0.4	0.39	*0.37*	0.46	0.45
FWSSNR [dB]	8.12	9.29	*9.52*	10.26	10.28	*10.55*
PESQ	2.14	2.39	*2.4*	2.79	2.84	*2.86*
			S1, far			
SRMR [dB]	4.58	*5.16*	5.13	5.05	*5.76*	5.72
CD [dB]	*2.67*	2.81	2.77	*2.81*	2.9	2.84
LLR	*0.38*	0.44	0.44	*0.39*	0.46	0.45
FWSSNR [dB]	6.68	8.29	*8.48*	8.38	9.69	*9.93*
PESQ	1.61	*1.71*	1.71	2.01	2.12	*2.14*
			S2, near			
SRMR [dB]	3.74	*4.55*	4.52	3.45	*3.97*	3.95
CD [dB]	4.63	3.89	*3.84*	3.78	3.15	*3.09*
LLR	0.49	0.47	*0.46*	*0.51*	0.53	0.52
FWSSNR [dB]	3.35	6.03	*6.19*	2.93	7.01	*7.07*
PESQ	1.4	1.72	*1.73*	2.12	2.55	*2.57*
			S2, far			
SRMR [dB]	2.97	*3.84*	3.81	2.78	*3.49*	3 .46
CD [dB]	5.21	4.68	*4.62*	4.75	4.16	*4.09*
LLR	0.75	0.72	*0.71*	0.77	0.73	*0.72*
FWSSNR [dB]	1.04	3.44	*3.56*	0.6	4.27	*4.3*
PESQ	1.19	1.27	*1.28*	1.33	1.52	*1.53*
			S3, near			
SRMR [dB]	3.57	*4.41*	4 .39	4.52	*5.52*	5.47
CD [dB]	4.38	3.73	*3.67*	4.19	3.61	*3.53*
LLR	0.65	0.64	*0.62*	0.65	0.65	*0.64*
FWSSNR [dB]	2.27	4.84	*4.97*	3.81	6.88	*7.07*
PESQ	1.37	1.65	*1.66*	1.59	1.87	*1.88*
			S3, far			
SRMR [dB]	2.73	*3.6*	3.57	3.14	*4.1*	4.04
CD [dB]	4.96	4.46	*4.38*	4.86	4.41	*4.32*
LLR	0.84	0.82	*0.79*	0.83	0.82	*0.8*
FWSSNR [dB]	0.24	2.7	*2.81*	1.4	4.5	*4.66*
PESQ	1.17	*1.24*	1.24	1.22	1.3	*1.31*

Table 3 SRMR values, in dB, obtained on the real data

	1 channel			8 channels		
SC scheme	Unp.	$SE_{1.5}$	SE_3	MVDR	MVDR $+SE_{1.5}$	MVDR $+SE_3$
all	3.18	*4.76*	4.69	3.57	*4.97*	4.89
R1, near	3.17	*4.81*	4.75	3.58	*5.04*	4.96
R1, far	3.19	*4.7*	4.64	3.56	*4.9*	4.82

window of 3 s (SE_3) for all acoustic conditions (rooms S1, S2, and S3 for positions "near" and "far"). For the 8-channel case, Tables 2 and 3 compare the quality of the output of the MVDR beamformer with and without spectral enhancement scheme, $SE_{1,5}$ and SE_3.

For each condition and for each instrumental quality measure, the best performance is highlighted by means of italic typeface to allow for an easier comparison. As expected, the selected instrumental measures do not always show completely consistent results [37, 38]. Nevertheless, some common tendencies can clearly be observed, which will be summarized next.

The results for all processed signals show an increase in SRMR, except for the MVDR beamformer in the case of room S2 (conditions "S2, near" and "S2, far") of the simulated data. These conditions are also the only ones in which the SRMR is higher in the single-channel case than in the multi-channel case. This performance difference may result from unvalid noise coherence matrix or from error in the DOA estimate for some utterances. The fact that the spectral enhancement scheme, used either alone or in combination with the MVDR beamformer, always increases the SRMR illustrates the ability of the proposed system to reduce the amount of reverberation both in the single- and the multi-channel case.

Additionally, the presented FWSSNR values depict a significant increase in comparison to the unprocessed microphone signal for all processed signals, except for the MVDR beamformer in the case of room S2. This illustrates the noise reduction capabilities of the proposed system. The difference in the FWSSNR values between the single- and the multi-channel scenarios further illustrates the benefit of using an MVDR beamformer aiming at noise reduction in the first stage. It can be noted that using a sliding window of 3 s instead of 1.5 s improves the FWSSNR scores in all simulated conditions, both in the single- and the multi-channel case. The advantage of using this longer sliding window is also illustrated by the lower CD values, both in the single- and in the multi-channel case, suggesting that distortions have been limited by avoiding leakage of the reverberation into the noise PSD estimate. Except for room S1, with the lowest amount of reverberation, both CD and LLR values are lower for the processed signals than for the unprocessed signal.

Finally, the improvement in the overall perceptual quality of the processed signal is illustrated by means of the PESQ score, which increases up to 0.19 and 0.49 for the single- and multi-channel scenarios, respectively. The PESQ score is increased in all conditions, with the largest improvement being obtained by the combined system MVDR + SE_3.

6.3 Word error rate

In order to evaluate the potential benefit of the proposed signal enhancement scheme on the performance of an ASR system, the processed signals have been used as the input for the baseline speech recognition system provided by the REVERB challenge [26]. This system is based on the hidden Markov model toolkit (HTK) [43], using mel-frequency cepstral coefficients, including Deltas and double Deltas, as features and acoustic models with tied-state hidden Markov models with 10 Gaussian components per state. The ASR models provided by the REVERB challenge [26] have been trained on clean data containing 7861 sentences uttered by 92 speakers for a total of approximately 17.5 h. The achieved ASR performance is measured in terms of WER, as depicted in Fig. 5, for the different signal enhancement schemes and acoustic conditions.

Compared to the scores obtained using the unprocessed signals (cf. horizontal black lines in Fig. 5), the WER increases slightly for the conditions with the lowest reverberation time (room S1). This indicates that spectral coloration introduced by the enhancement scheme may reduce the performance of the ASR system while the benefit of dereverberation is limited for small reverberation times. In all other conditions, the single-channel spectral enhancement scheme reduces the WER, with SE_3 yielding larger improvements than $SE_{1.5}$. Except for room S3, the MVDR beamformer yields better results than the single-channel scheme. The combination of the MVDR beamformer with SE_3 yields the largest improvement: absolute WER improvement up to 44.28 % for the simulated data (condition "S2, far") and up to 29.48 % for the real data (condition "R1, near").

6.4 Subjective evaluation of the speech quality

Since instrumental quality assessment, especially for the task of assessing dereverberation performance, may not always correlate well with the opinion of human listeners [37], we conducted a listening experiment in addition to the instrumental quality assessment described before.

The subjective evaluation is based on a multi-stimulus test with hidden reference and anchor (MUSHRA) following the specifications described in [28]. Four acoustic conditions have been tested, "S2, near'; "S2, far"; "R1, near"; and "R1, far". These conditions have been chosen to match the conditions used in the online MUSHRA test

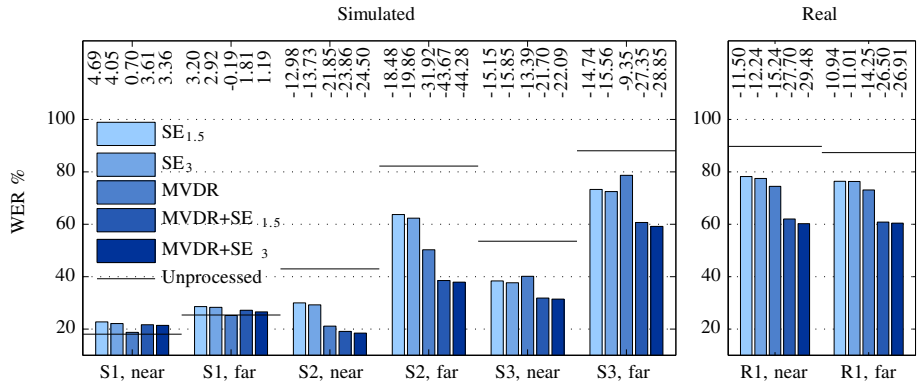

Fig. 5 WER obtained using the baseline recognizer of the REVERB challenge trained on clean data. Numbers indicate the difference with the WER obtained on unprocessed data

conducted in [27]. We have carried out a subjective evaluation for the unprocessed signal and for 3 processing schemes, namely, the single-channel scheme applied to the first microphone signal (SE_3), the MVDR beamformer using 8 microphones (MVDR), and the combination of the MVDR beamformer with the spectral enhancement scheme (MVDR + SE_3). In addition to these signals, a hidden reference and an anchor have been presented to the subjects. The hidden reference was the anechoic speech signal in the case of simulated data and the signal recorded by a headset microphone in the case of real data. The anchor consisted of the first microphone signal, low-pass filtered with a cut-off frequency of 3.5 kHz.

A total of 21 self-reported normal-hearing listeners participated in the MUSHRA listening test. The listening test was conducted in a soundproof booth and the subjects listened to diotic signals through headphones (Seinheiser HD 380 pro). Each subject evaluated 3 utterances per condition (i.e., 12 uterances per subject), in terms of two different attributes: "overall quality" and "perceived amount of reverberation", on a scale ranging from 0 to 100. For each subject, the utterances to be evaluated were randomly picked from the REVERB challenge database. All signals were normalized in amplitude and presented at a sampling frequency of 16 kHz and a quantization of 16 bit using a Roland sound card (model UA-25EXCW). The listening test was divided into three stages. In the first stage, the subjects were asked to listen to all files that would be presented to them during a training phase. This training phase allowed the subjects to get familiar with the data to be evaluated and to adjust the sound volume to a comfortable level. In the second stage, the subjects had to evaluate the overall quality of the signals and finally, the third stage consisted in the evaluation of the perceived amount of reverberation. The order of presentation of algorithms and conditions were randomized between all stages and all subjects.

The obtained MUSHRA scores are summarized in Fig. 6. The anchor appears to be the least satisfactory

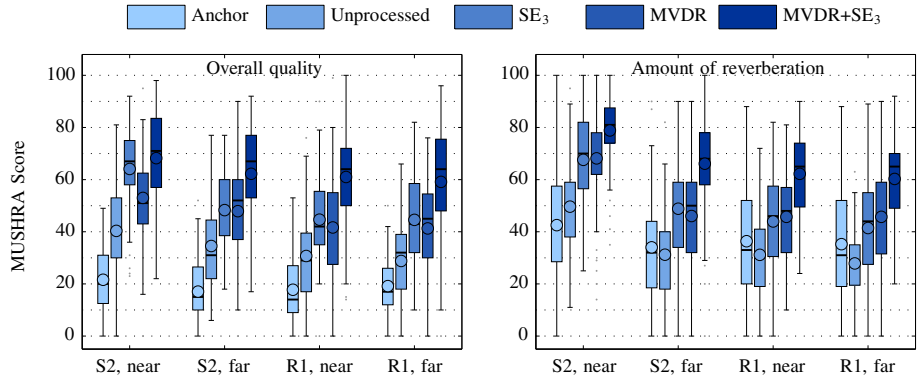

Fig. 6 MUSHRA scores for three processing schemes, the unprocessed signal and the low-pass filtered anchor. The highest score, 100, was labeled as "excellent" or "no reverberation" for the attributes "overall quality" and "perceived amount of reverberation", respectively. The means over all files and all subjects are displayed by circles. The scores of the hidden reference, close to 100 with small variance, are not displayed

Table 4 Results of the Friedman's test for both tested attributes. The value $p < 0.01$ indicates the significance of the results and χ^2 denotes the Friedman's chi square statistic

		S2, near	S2, far	R1, near	R1, far
Overall quality	χ^2	99.6	77.1	98.9	90.6
	p	<0.01	<0.01	<0.01	<0.01
Amount of reverberation	χ^2	93.9	120.8	98.6	104.7
	p	<0.01	<0.01	<0.01	<0.01

for the attribute "overall quality," suggesting that the subjects used the full extent of the grading scale. However, this is not the case for the attribute "perceived amount of reverberation", illustrating the difficulty of evaluating this attribute. The three considered processing schemes yielded an improvement compared to the unprocessed signal both in terms of "overall quality" and of "perceived amount of reverberation". As expected, the largest reduction of the "perceived amount of reverberation" is observed for the combination MVDR + SE_3. The combination MVDR + SE_3 improves the overall quality as well, although the improvement, compared to the single-channel scheme, is lower than for the attribute "perceived amount of reverberation". The use of an MVDR beamformer alone reduces the "perceived amount of reverberation" but does not improve the performance compared to the single-channel processing scheme (SE_3).

Since the scores of the MUSHRA test were not normally distributed, a Friedman's test [44] was used to examine the significance of the results, excluding the scores of the anchor and the reference. The results of the Friedman's test are presented in Table 4. The p value, $p < 0.01$, shows that at least one significant pairwise difference can be observed in all conditions and for all attributes. In order to examine the significance of the pairwise difference in performance between the processing schemes, a Wilcoxon rank sum test [45] has been used for each condition separately. A Bonferroni correction has been applied resulting in significant effects being considered for $p < 0.05/6$. For the attribute "perceived amount of reverberation", the differences in performance between the unprocessed signal and all processing schemes are significant but no significant differences were present between the different processing schemes. The same conclusion holds for the attribute "overall quality", except for the room R1 and the condition "S2, near", where the differences between the unprocessed signal and the output of the MVDR beamformer do not appear to be significant.

Even though the statistical significance criterion is not always satisfied, the trend of the results confirm the benefits of combining a beamformer with a single-channel spectral enhancement scheme for reducing reverberation and noise and for improving the overall speech quality.

7 Conclusions

In this paper, we have presented the combination of an MVDR beamformer with a single-channel spectral enhancement scheme, aiming at joint dereverberation and noise reduction. In the MVDR beamformer, the noise coherence matrix is estimated online using a VAD, whereas the DOA of the target speaker is estimated using the MUSIC algorithm. The output of this beamformer is processed using a spectral enhancement scheme combining statistical estimators of the speech, noise, and reverberant PSDs and aiming at joint residual reverberation and noise suppression. The evaluation of the proposed system, carried out using instrumental speech quality measures, a speech recognizer trained on clean data and subjective listening tests, illustrates the benefits of the proposed scheme.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The research leading to these results has received funding from the EU Seventh Framework Programme project DREAMS under grant agreement ITN-GA-2012-316969 as well as by the DFG-Cluster of Excellence EXC 1077/1, Hearing4all.

Author details

[1]Fraunhofer IDMT, Hearing Speech and Audio Technology, 26129 Oldenburg, Germany. [2]University of Oldenburg, Department of Medical Physics and Acoustics, 26111 Oldenburg, Germany. [3]Cluster of Excellence Hearing4all, Oldenburg, Germany.

Received: 22 February 2015 Accepted: 18 June 2015
Published online: 23 July 2015

References

1. J Benesty, J Chen, Y Huang, *Microphone Array Signal Processing*. (Springer, Berlin, Germany, 2008)
2. S Gannot, I Cohen, in *Springer Handbook of Speech Processing. Chap. 47*, ed. by Benesty J, MM Sondhi, and Y Huang. Adaptive beamforming and postfiltering (Springer Berlin, 2008)
3. Naylor PA, Gaubitch ND, *Speech Dereverberation*. (Springer, Berlin, 2010)
4. B Cauchi, I Kodrasi, R Rehr, S Gerlach, Jukić, T Gerkmann, S Doclo, S Goetze, in *Proc. REVERB Challenge Workshop*. Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme (Florence, Italy, 2014)
5. JS Bradley, H Sato, M Picard, On the importance of early reflections for speech in rooms. J. Acoust. Soc. Am. **113**(6), 3233–3244 (2003)
6. R Maas, EAP Habets, A Sehr, W Kellermann, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. On the application of reverberation suppression to robust speech recognition, vol. 1 (Kyoto, Japan, 2012), pp. 297–300
7. EAP Habets, S Gannot, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Dual-microphone speech dereverberation using a reference signal, vol. IV (Honolulu, USA, 2007), pp. 901–904
8. S Braun, EAP Habets, in *Proc. European Signal Processing Conference (EUSIPCO)*. Dereverberation in noisy environments using reference signals and a maximum likelihood estimator (Marrakech, Morocco, 2013)
9. A Schwarz, K Reindl, W Kellermann, in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*. On blocking matrix-based dereverberation for automatic speech recognition (Aachen, Germany, 2012), pp. 1–4
10. A Schwarz, K Reindl, W Kellermann, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A two-channel reverberation suppression scheme based on blind signal separation and wiener filtering (Kyoto, Japan, 2012), pp. 113–116

11. A Kuklasiński, S Doclo, SH Jensen, J Jensen, in *Proc. European Signal Processing Conference (EUSIPCO)*. Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids (Lisbon, Portugal, 2014), pp. 61–65
12. S Wisdom, T Powers, L Atlas, J Pitton, in *Proc. REVERB Challenge Workshop*. Enhancement of reverberant and noisy speech by extending its coherence (Florence, Italy, 2014)
13. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
14. T Gerkmann, RC Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012)
15. K Lebart, JM Boucher, PN Denbigh, A new method based on spectral subtraction for speech de-reverberation. Acta Acoustica. **87**, 359–366 (2001)
16. EAP Habets, S Gannot, I Cohen, Late reverberant spectral variance estimation based on a statistical mode. IEEE Signal Process. Lett. **16**(9), 770–774 (2009)
17. J Bitzer, KU Simmer, in *Microphone Arrays*, Digital Signal Processing, ed. by Brandstein M, Ward D. Superdirective microphone arrays (Springer Berlin, 2001), pp. 19–38
18. J Ramırez, JC Segura, C Benıtez, A De La Torre, A Rubio, Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3), 271–287 (2004)
19. RO Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
20. N Madhu, *Acoustic source localization: Algorithms, applications and extensions to source separation*. (Ph.D, Thesis, Ruhr-Universität Bochum, May 2009)
21. HW Löllmann, P Vary, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A blind speech enhancement algorithm for the suppression of late reverberation and noise (Taipei, Taiwan, 2009), pp. 3989–3992
22. C Breithaupt, M Krawczyk, R Martin, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech (Las Vegas, Nevada, USA, 2008), pp. 4037–4040
23. JD Polack, Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. Appl. Acoustics. **38**(2), 235–244 (1993)
24. C Breithaupt, T Gerkmann, R Martin, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing (Las Vegas, Nevada, USA, 2008), pp. 4897–4900
25. T Gerkmann, R Martin, On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling. IEEE Trans. Signal Process. **57**(11), 4165–4174 (2009)
26. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, EAP Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech (New Paltz, NY, USA, 2013)
27. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, EAP Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, Summary of the REVERB challenge (2014). [Online] Available: http://reverb2014.dereverberation.com/workshop/slides/reverb_summary.pdf. Accessed 07/07/15
28. ITU (ITU-R), Recommendation BS.1534–3: Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems. Online, available at: http://www.itu.int/rec/R-REC-BS.1534-2-201406-I/en, access date 07/07/15
29. H Cox, RM Zeskind, MM Owen, Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1376 (1987)
30. J Eaton, ND Gaubitch, PA Naylor, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost (Vancouver, Canada, 2013), pp. 161–165
31. IS Gradshteyn, IM Ryzhik, *Table of Integrals, Series, and Products*. (Academic Press, Inc., Boston, 1994)
32. Y Ephraim, D Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
33. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
34. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. WSJCAMO: a british english speech corpus for large vocabulary continuous speech recognition (Detroit, Michigan, USA, 1995), pp. 81–84
35. M Lincoln, I McCowan, J Vepa, HK Maganti, in *Proc. IEEE Workshop Autom. Speech Recognition and Understanding (ASRU)*. The multichannel Wall Street Journal audio–visual corpus (MC-WSJ-AV): Specification and initial experiments (Cancún, Mexico, 2005), pp. 357–362
36. REVERB Challenge, Documentation about the room impulse responses and noise data used for the REVERB challenge SimData. [Online] Available: http://reverb2014.dereverberation.com/tools/Document_RIR_noise_recording.pdf, Accessed: June 27, 2015
37. S Goetze, *On the Combination of Systems for Listening-Room Compensation and Acoustic Echo Cancellation in Hands-Free Telecommunication Systems*. (PhD thesis, Dept. of Telecommunications, University of Bremen (FB-1), Bremen, Germany, 2013)
38. S Goetze, A Warzybok, I Kodrasi, JO Jungmann, B Cauchi, J Rennies, E Habets, A Mertins, T Gerkmann, S Doclo, B Kollmeier, in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*. A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms (Antibes, France, 2014)
39. PC Loizou, *Speech Enhancement Theory and Practice*. (Taylor & Francis, New York, 2007)
40. T Falk, C Zheng, W-Y Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1766–1774 (2010)
41. Y Hu, PC Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008)
42. ITU-T, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs. Online, available at: https://www.itu.int/rec/T-REC-P.862-200102-I/en, access date 07/07/15
43. S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book*, 3.4.1 edn. (Cambridge University Engineering Dept, Cambridge, 2009). http://htk.eng.cam.ac.uk/prot-docs/HTKBook/htkbook.html
44. M Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937)
45. JD Gibbons, S Chakraborti, *Nonparametric Statistical Inference*. (Springer, Berlin, 2011)