JOINT DEREVERBERATION AND NOISE REDUCTION USING BEAMFORMING AND A SINGLE-CHANNEL SPEECH ENHANCEMENT SCHEME

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

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
benjamin.cauchi@idmt.fraunhofer.de

ABSTRACT

The REVERB challenge provides a common framework for the evaluation of speech enhancement algorithms in the presence of both reverberation and noise. This contribution proposes a system consisting of a commonly used combination of a beamformer with a single-channel speech enhancement scheme aiming at joint dereverberation and noise reduction. First, a minimum variance distortionless response beamformer with an on-line estimated noise coherence matrix is used to suppress the noise and possibly some reflections. The beamformer output is then processed by a single-channel speech enhancement scheme, incorporating temporal cepstrum smoothing which suppresses both reverberation and residual noise. Experimental results show that improvements are particularly significant in conditions with high reverberation times.

Index Terms— REVERB challenge, dereverberation, noise reduction.

1. INTRODUCTION

In teleconferencing applications, voice-controlled systems and hearing aids, the recorded speech signals are often corrupted by both reverberation and noise, resulting in speech quality and speech intelligibility degradation, as well as deterioration in automatic speech recognition (ASR) performance. Several algorithms have been proposed in the literature to deal with these issues [1–6], but the lack of a common evaluation framework made the comparison between different approaches difficult. The REVERB challenge proposes an evaluation framework aiming to facilitate the progress of speech enhancement algorithms for noisy and reverberant environments [7].

The research leading to these results has received funding from the EU Seventh Framework Programme project DREAMS under grant agreement ITN-GA-2012-316969, from the DFG Cluster of Excellence 1077 Hearing4All, from a GIF grant, and from the MWK PhD Program Signals and Cognition.

The system proposed in this contribution consists of a commonly used combination of a beamformer and a single-channel speech enhancement scheme. First, the multi-channel input signals are processed using a minimum variance distortionless response (MVDR) beamformer [8], which aims to suppress sound sources not arriving from the direction of arrival (DOA) of the target speaker. The noise coherence matrix in the MVDR beamformer is estimated from noise-only periods, determined using a voice activity detector (VAD) [9], and the DOA of the target speaker is estimated using the multiple signal classification (MUSIC) algorithm [10, 11].

The beamformer output is then processed using a single-channel speech enhancement scheme, which aims at jointly suppressing the remaining noise and reverberation and relies on estimates of the power spectral densities (PSDs) of the noise and of the reverberation similarly as in [5]. The proposed scheme computes a real-valued gain function combining the clean speech amplitude estimator in [12], the noise PSD estimator based on minimum statistics in [13], and the estimator of the PSD of the late reverberation based on statistical room acoustics in [14]. In addition, adaptive smoothing in the cepstral domain is used to estimate the speech PSD in order to reduce the musical noise which is often a byproduct of spectral enhancement schemes [15].

This paper is organized as follows. In Section 2 the notation is introduced and the proposed system is described. A description of the used beamformer is provided in Section 3, while the single-channel spectral enhancement scheme is described in Section 4. The challenge and the evaluation corpus are introduced in Section 5 and the results achieved by the proposed system are presented in Section 6.

2. NOTATION AND CONFIGURATION

Consider an acoustic system with a single speech source and M microphones. The reverberant and noisy m-th microphone

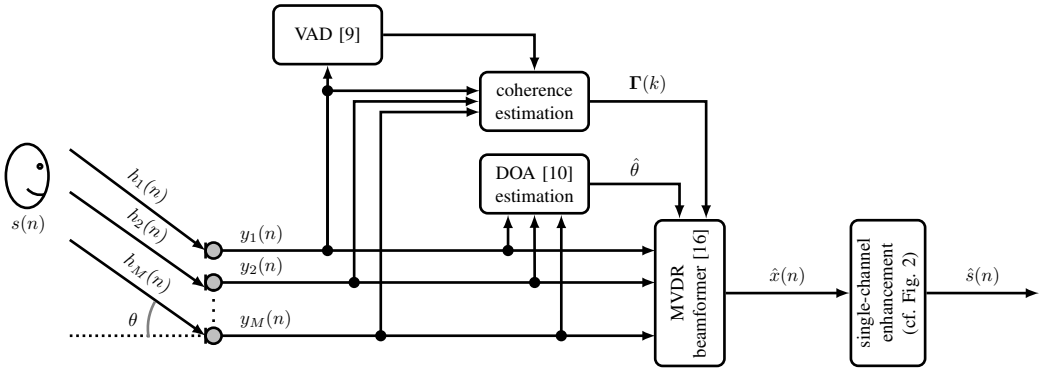

Fig. 1. Overview of the proposed system.

signal $y_m(n)$ at time index n is given by

$$\begin{aligned} y_m(n) &= s(n) * h_m(n) + v_m(n) \\ &= x_m(n) + v_m(n),\ m = 1,\ \cdots,\ M, \end{aligned} \tag{1}$$

with $s(n)$ being the clean speech signal, $h_m(n)$ being the room impulse response (RIR) between the source and the m-th microphone, and $x_m(n)$ and $v_m(n)$ denoting the reverberant speech component and the additive noise component of the m-th microphone signal, respectively. Aiming to obtain an estimate $\hat{s}(n)$ of the clean speech signal $s(n)$, the speech enhancement scheme depicted in Fig. 1 is proposed.

First, the received microphone signals $y_m(n)$ are used to estimate the noise coherence matrix and the DOA of the target speech signal. The DOA estimation is based on the MUSIC algorithm [10, 11] which will be briefly described in Section 3.2. The estimated noise coherence matrix and DOA are then used to design an MVDR beamformer, which aims at noise reduction and some dereverberation by suppressing the sound sources not arriving from the target DOA while providing a unity gain in the direction of the target speaker. Finally, the beamformer output $\hat{x}(n)$ is processed by a single-channel speech enhancement scheme, described in Section 4, which aims at joint noise and reverberation suppression.

In the remainder of this paper, the short-time Fourier transform (STFT) representations of $s(n)$, $x_m(n)$, $v_m(n)$, $y_m(n)$ and $\hat{x}(n)$ are denoted by $S(k,\ell)$, $X_m(k,\ell)$, $V_m(k,\ell)$, $Y_m(k,\ell)$ and $\hat{X}(k,\ell)$, respectively, with k and ℓ representing the frequency bin and frame indices.

3. BEAMFORMER

3.1. MVDR beamformer

The M-dimensional stacked vector of the received microphone signals $\mathbf{Y}(k,\ell)$ can be written as

$$\mathbf{Y}(k,\ell) = \mathbf{X}(k,\ell) + \mathbf{V}(k,\ell), \tag{2}$$

with

$$\mathbf{Y}(k,\ell) = [Y_1(k,\ell)\ Y_2(k,\ell)\ \ldots\ Y_M(k,\ell)]^T, \tag{3}$$

and $\mathbf{X}(k,\ell)$ and $\mathbf{V}(k,\ell)$ defined similarly as in (3). The beamformer output signal $\hat{X}(k,\ell)$ is obtained by filtering and summing the microphone signals, i.e.,

$$\begin{aligned} \hat{X}(k,\ell) &= \mathbf{W}_\theta^H(k)\mathbf{Y}(k,\ell) \\ &= \mathbf{W}_\theta^H(k)\mathbf{X}(k,\ell) + \mathbf{W}_\theta^H(k)\mathbf{V}(k,\ell), \end{aligned} \tag{4}$$

with $\mathbf{W}_\theta(k)$ denoting the stacked filter coefficient vector of the beamformer steered towards the angle θ. Aiming at minimizing the noise output power while providing a unity gain in the direction of the target speaker, the filter coefficients of the MVDR beamformer are computed as

$$\mathbf{W}_\theta(k) = \frac{\mathbf{\Gamma}^{-1}(k)\mathbf{d}_\theta(k)}{\mathbf{d}_\theta^H(k)\mathbf{\Gamma}^{-1}(k)\mathbf{d}_\theta(k)}, \tag{5}$$

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

$$\mathbf{d}_\theta(k) = [e^{-j2\pi f_k \tau_1(\theta)}\ e^{-j2\pi f_k \tau_2(\theta)}\ \cdots e^{-j2\pi f_k \tau_M(\theta)}], \tag{6}$$

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

As can be clearly seen from equations (5) and (6), in order to compute the beamformer filter coefficients, the DOA of the target speaker as well as the noise coherence matrix need to be estimated. Estimation of the target DOA will be discussed in Section 3.2. To estimate the noise coherence matrix $\mathbf{\Gamma}(k)$, the VAD described in [9] is used and $\mathbf{\Gamma}(k)$ is computed using all detected noise-only frames. However, if the length of the detected noise-only period is too short (cf. Section 5), the coherence matrix $\mathbf{\Gamma}_{\text{diff}}(k)$ of a diffuse noise field is used instead, resulting in the well-known superdirective beamformer [8]. Since superdirective beamformers are known to be sensitive to uncorrelated noise, a white noise gain constraint $\text{WNG}_{\max}$

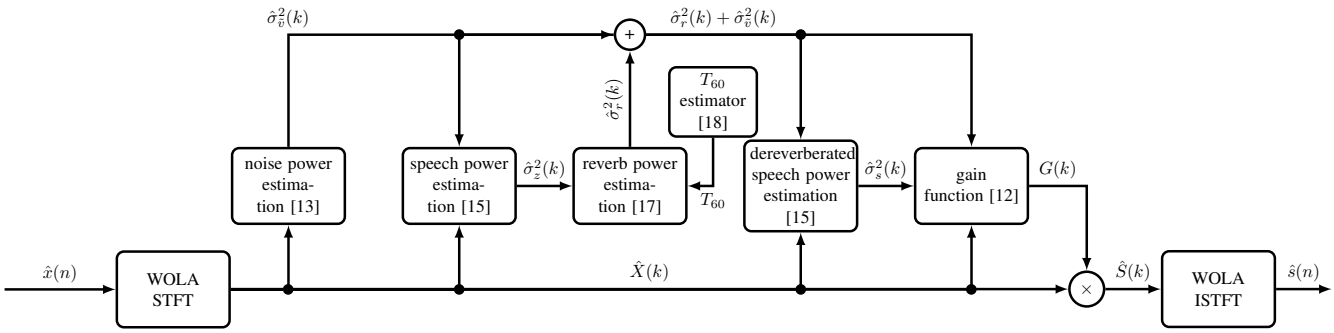

Fig. 2. Overview of the proposed single-channel enhancement scheme for a single frame.

is imposed in this case. With such a constraint the filter coefficients are computed as

$$\mathbf{W}_\theta(k) = \frac{\left(\mathbf{\Gamma}_{\text{diff}}(k) + \varrho(k)\mathbf{I}_M\right)^{-1}\mathbf{d}_\theta(k)}{\mathbf{d}_\theta^H(k)\left(\mathbf{\Gamma}_{\text{diff}}(k) + \varrho(k)\mathbf{I}_M\right)^{-1}\mathbf{d}_\theta(k)} \tag{7}$$

with $\mathbf{I}_M$ the $M \times M$-dimensional identity matrix and $\varrho(k)$ a frequency-dependent regularization parameter which is iteratively computed such that $\mathbf{W}_\theta^H(k)\mathbf{W}_\theta(k) \leq \text{WNG}_{\max}$ [16] .

3.2. DOA estimation

Since an error in the DOA estimate $\hat{\theta}$ of the target speech signal can lead to the beamformer suppressing the desired signal, a robust subspace-based algorithm (MUSIC) has been used to estimate the DOA of the target speaker [10, 11]. Using the MUSIC algorithm, this DOA can be estimated as

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \frac{1}{K}\sum_{k_{\text{low}}}^{k_{\text{high}}} U_\theta(k,\ell), \tag{8}$$

where K denotes the total number of considered frequency bins $k = k_{\text{low}} \dots k_{\text{high}}$ and $U(k,\ell,\theta)$ denotes the so-called MUSIC pseudo-spectra, which are calculated as

$$U_\theta(k,\ell) = \frac{1}{\mathbf{d}_\theta^H(k)\boldsymbol{E}(k,\ell)\boldsymbol{E}^{\text{H}}(k,\ell)\mathbf{d}_\theta(k)}. \tag{9}$$

The noise subspace

$$\boldsymbol{E}(k,\ell) = \left[\boldsymbol{e}_{Q+1}(k,\ell) \dots \boldsymbol{e}_M(k,\ell)\right], \tag{10}$$

is an $M \times (M-Q)$-dimensional matrix, with Q being the number of sources (i.e. $Q = 1$ in this case), composed of the eigenvectors of the covariance matrix of $\mathbf{Y}(k,\ell)$ corresponding to the $(M-Q)$ smallest eigenvalues.

Assuming that speech and noise are uncorrelated, the steering vector corresponding to the true DOA is orthogonal to the noise subspace such that the DOA of the target speaker can be estimated as the angle maximizing the sum of the MUSIC-pseudo-spectra in equation (9).

4. SINGLE-CHANNEL ENHANCEMENT

The single-channel enhancement scheme which is applied to the output signal $\hat{x}(n)$ of the MVDR-beamformer is summarized in Fig. 2. The signal $\hat{x}(n)$ is assumed to contain the clean speech signal $s(n)$ as well as residual reverberation $r(n)$ and noise $\tilde{v}(n)$, i.e.

$$\hat{x}(n) = s(n) + r(n) + \tilde{v}(n). \tag{11}$$

In the STFT domain, (11) is expressed as

$$\hat{X}(k,\ell) = Z(k,\ell) + \tilde{V}(k,\ell), \tag{12}$$

where

$$Z(k,\ell) = S(k,\ell) + R(k,\ell). \tag{13}$$

Aiming at jointly reducing reverberation and noise, a real-valued gain $G(k,\ell)$ is applied to the beamformer output signal, i.e.

$$\hat{S}(k,\ell) = G(k,\ell)\hat{X}(k,\ell), \tag{14}$$

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

The gain is computed by using the minimum mean square error (MMSE) estimator for the clean speech magnitude proposed in [12]. This estimator, similarly to the Wiener filter, requires the PSDs of the clean speech, of the noise and of the reverberation components, which have to be estimated from the beamformer output signal. First, an estimate of the noise PSD is obtained using minimum statistics [13] and further used to estimate the reverberant speech PSD. 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, in order to avoid the reverberation leaking into the noise PSD estimate, a longer minimum search window is used in the minimum statistics approach as compared to [13] (cf. Section 5).

The PSD of the reverberant speech is estimated using temporal cepstrum smoothing [15] and the late reverberation PSD is estimated from the reverberant speech PSD using the approach proposed in [17]. This approach requires an estimate of the reverberation time T_{60}, which has been obtained using

the estimator described in [18]. The PSD of the clean speech component is finally obtained by a re-estimation, again using temporal cepstrum smoothing. The following sections give a more detailed description of the different components of the proposed single-channel enhancement scheme.

4.1. Gain function

In [12], it is assumed that the speech magnitude $|S(k,\ell)|$ follows a chi probability density function (PDF) with a 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 by 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

$$\sigma_j^2(k,\ell) = \mathrm{E}\left\{|J(k,\ell)|^2\right\} = \sigma_{\tilde{v}}^2(k,\ell) + \sigma_r^2(k,\ell), \tag{15}$$

with $\sigma_r^2(k,\ell)$ and $\sigma_{\tilde{v}}^2(k,\ell)$ denoting the PSDs of the reverberation and of the noise component, respectively. With $\sigma_s^2(k,\ell)$ denoting the PSD of the clean speech and $\xi(k,\ell)$ denoting the *a priori* signal-to-interference ratio (SIR) defined as,

$$\xi(k,\ell) = \frac{\sigma_s^2(k,\ell)}{\sigma_r^2(k,\ell) + \sigma_{\tilde{v}}^2(k,\ell)}, \tag{16}$$

the clean speech magnitude is estimated by optimizing the MMSE criterion

$$\left|\hat{S}(k,\ell)\right| = \underset{\left|\hat{S}(k,\ell)\right|}{\operatorname{argmin}} \mathrm{E}\left\{\epsilon(k,\ell)|\hat{X}(k,\ell), \sigma_j^2(k,\ell), \xi(k,\ell)\right\}, \tag{17}$$

with

$$\epsilon(k,\ell) = \left(|S(k,\ell)|^\beta - \left|\hat{S}(k,\ell)\right|^\beta\right)^2, \tag{18}$$

where the parameter β is a compression factor such that a different emphasis is given on estimation errors for small amplitudes in relation to large amplitudes.

According to [12], with $\gamma(k,\ell)$ denoting the *a posteriori* SIR defined as

$$\gamma(k,\ell) = \frac{|X(k,\ell)|^2}{\sigma_r^2(k,\ell) + \sigma_{\tilde{v}}^2(k,\ell)}, \tag{19}$$

and

$$\nu(k,\ell) = \frac{\gamma(k,\ell)\xi(k,\ell)}{\mu + \xi(k,\ell)}, \tag{20}$$

the solution to (17) leads to the gain function $\tilde{G}(k,\ell)$

$$\tilde{G}(k,\ell) = \sqrt{\frac{\xi(k,\ell)}{\mu + \xi(k,\ell)}} \times \left[\frac{\mathrm{Gam}\left(\mu + \frac{\beta}{2}\right)}{\mathrm{Gam}\left(\mu\right)} \frac{\Phi\left(1 - \mu - \frac{\beta}{2}, 1; -\nu(k,\ell)\right)}{\Phi\left(1 - \mu, 1; -\nu(k,\ell)\right)}\right]^{1/\beta} \times \left(\sqrt{\gamma(k,\ell)}\right)^{-1}, \tag{21}$$

where $\Phi(\cdot)$ denotes the confluent hypergeometric function while $\mathrm{Gam}\left(\cdot\right)$ denotes the complete Gamma function. Depending on the choice of β and μ, this estimator can resemble other well known estimators, such as the short-time spectral amplitude estimator [19] or the log-spectral amplitude estimator [20]. To compute the expression in (21), 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.

In order to reduce artifacts which may be introduced by (21), the gain $G(k,\ell)$ used in (14) is restricted to values larger than a spectral floor G_{min} (cf. Section 5), i.e.,

$$G(k,\ell) = \max\left(\tilde{G}(k,\ell), G_{\text{min}}\right). \tag{22}$$

4.2. Noise PSD estimator

The noise PSD $\sigma_{\tilde{v}}^2(k,\ell)$ is estimated using the minimum statistics approach [13], which tracks the minima of the input signal PSD over a sliding window and has been shown to be reliable for slowly varying and stationary noises. A realization of the PSD of the noise signal is first estimated as the smoothed periodogram of the input signal which is obtained as

$$P_{\tilde{v}}(k,\ell) = \alpha P_{\tilde{v}}(k,\ell-1) + (1-\alpha)|X(k,\ell)|^2, \tag{23}$$

with α denoting a smoothing parameter. The PSD $\sigma_{\tilde{v}}^2(k,\ell)$ is estimated as the minimum of $P_{\tilde{v}}(k,\ell)$ over a short temporal sliding window, with a usual length corresponding to 1.5 s. This technique relies on the assumption that the minimum of $P_{\tilde{v}}(k,\ell)$ within a 1.5 s window is not affected by speech, allowing for an inference of the noise PSD. In reverberant environments, however, the decay time in speech pauses may be increased. Thus, in order to avoid reverberant speech to affect the noise PSD estimate $\sigma_{\tilde{v}}^2(k,\ell)$, a longer tracking window is used (cf. Section 5).

Tracking the minimum of the PSD using a time-frequency independent α can lead to an inaccurate estimate of $P_{\tilde{v}}(k,\ell)$ and a delay in detecting augmentation of the noise power. In order to circumvent these issues, a time-frequency variant smoothing constant $\alpha(k,\ell)$ has been derived in [13], which aims to minimize the MMSE

$$\alpha(k,\ell) = \underset{\alpha(k,\ell)}{\operatorname{argmin}} \mathrm{E}\left\{P_{\tilde{v}}(k,\ell) - \sigma_{\tilde{v}}^2(k,\ell)|P_{\tilde{v}}(k,\ell-1)\right\}. \tag{24}$$

The solution to (24) is given by

$$\alpha(k,\ell) = \frac{1}{1 + \left(\frac{P_{\tilde{v}}(k,\ell-1)}{\sigma_{\tilde{v}}^2(k,\ell)} - 1\right)^2}, \tag{25}$$

in which $\sigma_{\tilde{v}}^2(k,\ell)$ is in practice unavailable and replaced by $\sigma_{\tilde{v}}^2(k,\ell-1)$. In order to compensate for the delay in the

adaptation of $\sigma_{\tilde{v}}^2(k,\ell)$, which could lead to overestimation or underestimation, the smoothing parameter is corrected as

$$\alpha(k,\ell) = \frac{\alpha_{\max}\alpha_c(k,\ell)}{1+\left(\frac{P_{\tilde{v}}(k,\ell-1)}{\sigma_{\tilde{v}}^2(k,\ell-1)}-1\right)^2}, \tag{26}$$

where $\alpha_{\max}$ is the maximum allowable smoothing constant and $\alpha_c(k,\ell)$ is given by

$$\begin{aligned}\alpha_c(k,\ell) &= 0.7\,\alpha_c(k,\ell-1) + 0.3\max\left(\tilde{\alpha}_c(k,\ell), 0.7\right),\\ \tilde{\alpha}_c(k,\ell) &= \frac{1}{1+\left(\frac{\sum_{k=0}^{L-1} P_{\tilde{v}}(k,\ell-1)}{\sum_{k=0}^{L-1}|X(k,\ell)|^2}-1\right)^2}.\end{aligned} \tag{27}$$

Additionally, a lower limit $\alpha_{\min}(k,\ell)$ is applied to $\alpha_c(k,\ell)$ in order to improve the performance of the estimator in high levels of non-stationary noise. The resulting estimate $\hat{\sigma}_{\tilde{v}}^2(k,\ell)$ of $\sigma_{\tilde{v}}^2(k,\ell)$ is used to estimate both $\sigma_s^2(k,\ell)$ and $\sigma_r^2(k,\ell)$ as described in the following section.

4.3. Speech PSD estimator

Once the estimate $\hat{\sigma}_{\tilde{v}}^2(k,\ell)$ is available, temporal cepstrum smoothing, proposed in [15], is used to estimate the PSD $\sigma_z^2(k,\ell)$ of the reverberant speech component $Z(k,\ell)$. The same method can also be used to estimate the dereverberated speech component $\sigma_s^2(k,\ell)$ if an estimate of the reverberation power $\sigma_r^2(k,\ell)$ is available. The modifications of the formula required for the latter case are described in 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)

$$\xi_z^{\mathrm{ml}}(k,\ell) = \frac{|X(k,\ell)|^2}{\sigma_{\tilde{v}}^2(k,\ell)} - 1 \tag{28}$$

is employed. The speech power $P_z(k,\ell)$ can then be obtained as

$$P_z(k,\ell) = \sigma_{\tilde{v}}^2(k,\ell)\max\left(\xi_z^{\mathrm{ml}}(k,\ell),\,\xi_{\min}^{\mathrm{ml}}\right), \tag{29}$$

where $\xi_{\min}^{\mathrm{ml}} > 0$ is a lower bound to avoid negative or very small values of $\xi_z^{\mathrm{ml}}(k,\ell)$. In the cepstral domain, $P_z(k,\ell)$ can be represented by $\lambda_z^{\mathrm{ml}}(q,\ell)$ as

$$\lambda_z^{\mathrm{ml}}(q,\ell) = \mathrm{IDFT}\left\{\log\left(P_z(k,\ell)|_{k=0,\cdots,(L-1)}\right)\right\}, \tag{30}$$

where q is the cepstral bin index. A recursive temporal smoothing is applied to $\lambda_z^{\mathrm{ml}}(q,\ell)$, i.e.,

$$\lambda_z(q,\ell) = \delta(q,\ell)\lambda_z(q,\ell-1) + (1-\delta(q,\ell))\lambda_z^{\mathrm{ml}}(q,\ell), \tag{31}$$

With $\delta(q,\ell)$ being a time-quefrency dependent smoothing parameter. Finally, $\hat{\sigma}_z^2(k,\ell)$ can be obtained by transforming $\lambda_z(q,\ell)$ into the spectral domain as

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

	Simulated	Real
# of sentences	2176 ($\sim$ 4.8 hrs.)	372 ($\sim$ 0.6 hrs.)
# of speakers	28	10

Table 1. Quantity of data in the evaluation set.

where κ, estimated as in [21], is a constant introduced to compensate for the bias due to the recursive smoothing in the log-domain in (31). Only little smoothing is applied to the cepstral bins which are mainly related to speech while for the remaining coefficients a stronger smoothing is employed. Consequently, small smoothing constants are chosen for the low quefrencies, as they contain information about the vocal tract shape. The same holds for the coefficients 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 constant $\delta(q,\ell)$ in (31) 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

$$\delta(q,\ell) = \begin{cases}\delta_{\mathrm{pitch}} & \text{if } q \in \mathbb{Q}\\ \bar{\delta}(q,\ell) & \text{if } q \in \{0,\cdots,L/2\}\setminus\mathbb{Q}\end{cases} \tag{33}$$

where $\mathbb{Q}$ is a small set of cepstral bins around the quefrency corresponding to f_0 and δ_{pitch} is the smoothing constant for the pitch coefficients [15]. The quantity $\bar{\delta}(q,\ell)$ is given as

$$\bar{\delta}(q,\ell) = \eta\delta(q,\ell-1) + (1-\eta)\bar{\delta}^{\mathrm{const}}(q), \tag{34}$$

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.

The reverberant speech PSD can be used to estimate the PSD of the late reverberation $\sigma_r^2(k,\ell)$ as shown in the following section. After having estimated $\sigma_r^2(k,\ell)$, cepstral smoothing is also used to estimate the dereverberated speech PSD $\sigma_s^2(k,\ell)$. In this case, the noise PSD $\sigma_{\tilde{v}}^2(k,\ell)$ in equation (28) and (29) is replaced by the interference PSD $\sigma_j^2(k,\ell) = \sigma_{\tilde{v}}^2(k,\ell) + \sigma_r^2(k,\ell)$. The dereverberated speech PSD $P_s(k,\ell)$ can finally be computed and used to obtain the estimate $\hat{\sigma}_s^2(k,\ell)$ of $\sigma_s^2(k,\ell)$.

4.4. Reverberation estimation

The reverberant PSD $\sigma_r^2(k,\ell)$ is estimated using the method described in [17], which is based on the RIR model suggested in [14] that represents the RIR as a Gaussian stationary noise signal multiplied by an exponential decay rate Δ dependent of the T_{60}

$$\Delta = \frac{3\ln 10}{T_{60} f_s}. \tag{35}$$

This estimator represents the PSD of the reverberant speech $\sigma_z^2(k,\ell)$ as

$$\sigma_z^2(k,\ell) = \sigma_r^2(k,\ell) + \sigma_s^2(k,\ell), \tag{36}$$

		Simulated Data							Real Data		
		T60=250ms		T60=500ms		T60=700ms		Mean	T60=700ms		Mean
		near	far	near	far	near	far		near	far	
1 channel	SRMR [dB]	4.7(4.5)	4.8(4.6)	4.3(3.8)	3.9(3.0)	4.3(3.6)	3.9(2.7)	4.3(3.7)	4.9(3.2)	4.8(3.2)	4.8(3.2)
	FWSSNR [dB]	10.3(8.2)	8.9(6.8)	6.2(3.3)	3.5(1.0)	4.9(2.3)	2.8(0.3)	6.0(3.6)			
	CD [dB]	2.0(2.0)	2.7(2.7)	3.8(4.6)	4.7(5.2)	3.7(4.4)	4.4(5.0)	3.6(4.0)			
	LLR	0.5(0.4)	0.4(0.4)	0.5(0.5)	0.8(0.8)	0.6(0.7)	0.8(0.9)	0.6(0.6)			
	PESQ	2.4(2.2)	1.7(1.6)	1.7(1.4)	1.3(1.2)	1.6(1.4)	1.3(1.2)	1.7(1.5)			
8 channels	SRMR [dB]	6.7(4.5)	5.3(4.6)	3.7(3.8)	3.4(3.0)	5.3(3.6)	4.4(2.7)	4.8(3.7)	4.9(3.2)	4.8(3.2)	4.8(3.2)
	FWSSNR [dB]	11.3(8.2)	10.4(6.8)	6.9(3.3)	4.1(1.0)	7.0(2.3)	4.5(0.3)	7.3(3.6)			
	CD [dB]	2.4(2.0)	2.8(2.7)	3.0(4.6)	4.1(5.2)	3.6(4.4)	4.4(5.0)	3.4(4.0)			
	LLR	0.5(0.5)	0.5(0.4)	0.6(0.5)	0.8(0.8)	0.7(0.7)	0.8(0.9)	0.6(0.6)			
	PESQ	2.9(2.2)	2.1(1.6)	2.5(1.4)	1.5(1.2)	1.8(1.4)	1.3(1.2)	2.0(1.5)			

Table 2. Mean signal-based measures over all utterances using either 1 or 8 channels. The scores of the unprocessed signals are displayed between parentheses.

which leads to the estimate of $\sigma_r^2(k,\ell)$,

$$\hat{\sigma}_r^2(k,\ell) = \mathrm{e}^{-2\Delta T_d f_s}\sigma_z^2(k,\ell - T_d/T_s). \quad (37)$$

In (37), T_s denotes the frame shift whereas T_d is the duration of the direct path and early reflections of the RIR, typically set between 50 ms and 80 ms. As a result, $\hat{\sigma}_r^2(k,\ell)$ can be estimated using $\hat{\sigma}_z^2(k,\ell)$ and the reverberation time.

5. EXPERIMENTAL SETUP

5.1. Corpus description

The results presented in this contribution have been obtained using the evaluation set of the REVERB challenge [7], which consists of a large corpus of speech corrupted by reverberation and noise. This corpus is divided into simulated and real data as described in Table 1. All recordings have been made at a sampling frequency of 16 kHz with a circular array with 20 cm diameter and 8 equidistant microphones.

The simulated data is composed of close talk speech taken from the WSJCAM0 corpus [22] which has been convolved with recorded RIRs and to which measured noise signals at a fixed SNR of 20 dB have been added. The RIRs have been measured in three different rooms with reverberation times of 250, 500 and 700 ms. The distance between the source and the array is either 0.5 m (condition "near") or 2 m (condition "far").

The real data is composed of utterances from the MC-WSJ-AV corpus [23] and contains speech recorded in a noisy reverberant room with $T_{60} \approx 700$ ms at a distance between the source and the array of either 1 m (condition "near") or 2.5 m (condition "far"). Utterances have been spoken from different unknown positions within the room but the position was constant during each utterance.

5.2. Algorithm settings

The proposed system, described in Section 2, has been applied using utterance-based processing, assuming that the T_{60} and the DOA of the target speaker remained constant within each utterance. The STFT has been computed using a 32 ms Hann window with 50 % overlap. The DOA as 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 array for a frequency range from 50 Hz to 5 kHz. The beamformer, described in Section 3, uses a theoretically diffuse noise field and a white noise constraint $\mathrm{WNG}_{\max}$ =-10 dB if less than 10 frames of noise-only period have been detected within the utterance. The speech amplitude estimator in Section 4.1 assumed a chi PDF with $\mu = 0.5$, 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 used for minima tracking which has been set to 3 s. In Section 4.3, the parameters used to estimate the speech PSD have been set as in [12] while in Section 4.4, T_d has been set to 80 ms. The evaluation has been run for 1 channel and for 8 channels, with the single-channel scenario referring to only applying the proposed single-channel enhancement scheme to the first microphone signal, $y_1(n)$.

6. RESULTS

6.1. Signal-based quality assessment

The different performance for each condition, as well as the mean performance over all conditions are presented in Table 2. The performance of the proposed system has been evaluated using the signal-based measures defined in [7], i.e., the signal to reverberant modulation ratio (SRMR) [24], the frequency-weighted segmental SNR (FWSSNR) [25],

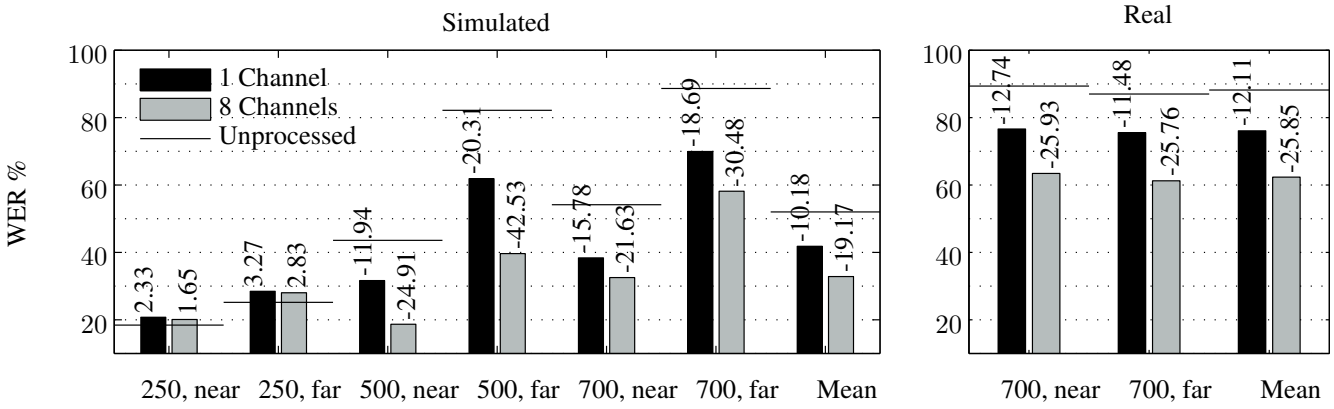

Fig. 3. WER obtained using the baseline recognizer of the REVERB challenge trained on clean data. Numbers indicate the difference with the WER obtained on unprocessed data.

the cepstral distance (CD) [25], the log-likelihood ratio (LLR) [25] and the perceptual evaluation of speech quality (PESQ) [26]. Among these 5 measures, the SRMR is the only non-intrusive measure and is hence the only measure that can be used to evaluate the performance for real data, for which no reference signal is available. The other measures use the clean speech $s(n)$ as reference signal.

The increase in SRMR for all considered conditions shows that reverberation is reduced for both the single- and multi-channel scenarios, with the results being more significant for higher reverberation times as expected. While the SRMR increase is typically higher for the multi-channel scenario, for the condition of $T_{60} = 500$ ms, the single-channel scenario seems to achieve a higher dereverberation performance. The reason behind this performance difference might lie in the fact that the statistical model of the RIR used in Section 4.4 may not hold for the output of the MVDR-beamformer. However, further investigations are needed to derive a sound explanation.

Furthermore, the presented FWSSNR values depict a significant increase in comparison to the unprocessed microphone signal, illustrating the noise reduction capabilities of the proposed system. The difference in the FWSSNR values between the single- and multi-channel scenario further illustrates the benefit of using an MVDR beamformer aiming at noise reduction in the first stage. Finally, the improvement in the overall perceptual quality of the processed signal is illustrated in the average PESQ score increase of 0.5 and 0.2 for the multi- and single-channel scenarios, respectively.

6.2. Word error rate

In order to evaluate the potential benefit of the proposed scheme on the performance of an ASR system, the processed data have been used as the input for the baseline speech recognition system provided by the REVERB challenge [7]. This system is based on the hidden Markov model toolkit (HTK) [27] and uses mel-frequency cepstral coefficients, including Deltas and double-Deltas, as features and acoustic models using tied-state hidden Markov models with 10 Gaussian components per state. In this contribution, the models have been trained on clean data containing 7861 sentences uttered by 92 speakers for a total of approximately 17.5 hours. The achieved performance is measured in terms of word error rate (WER) as depicted in Fig. 3.

Compared to the scores obtained using the unprocessed signals, the absolute WER improvement on simulated data is of 19.17 % and 10.18 % for the multi- and single-channel scenarios, respectively. Greater absolute WER improvement is obtained on real data, i.e. 25.85 % for the multi-channel and 12.11 % for the single-channel scenario. On the other hand, the WER increases slightly for the conditions with the lowest reverberation time, i.e. $T_{60} = 250$ ms. This suggests 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. This drawback could be avoided by training the acoustic models on processed signals.

7. CONCLUSION

This contribution proposes to achieve joint dereverberation and noise reduction a combination of an MVDR-beamformer and a single-channel speech enhancement scheme. In the MVDR-beamformer the noise coherence matrix is estimated on-line using a VAD and the DOA of the target speaker, which is required to compute the steering vector, is obtained using the MUSIC algorithm.

The output of this beamformer is processed using a speech-enhancement scheme combining statistical estimators of the speech, noise and reverberant PSDs and aiming at joint dereverberation and residual noise suppression. The evaluation of the proposed system, carried out using signal-based quality measures and a speech recognizer trained on clean speech, illustrates the benefit of the proposed scheme.

8. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[3] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 31–34.

[4] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[5] H. W. Löllmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3989–3992.

[6] E. A. P. Habets, I. Cohen, S. Gannot, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1433–1451, Nov. 2008.

[7] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E.A.P. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, N.Y., U.S.A., Oct. 2013.

[8] J. Bitzer and K.U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 19–38. Springer Berlin Heidelberg, 2001.

[9] J. Ramırez, J. C Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[10] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[11] N. Madhu, *Acoustic Source Localization: Algorithms, Applications and Extensions to Source Separation*, Ph.D. Thesis, Ruhr-Universität Bochum, May 2009.

[12] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, U.S.A, Apr. 2008, pp. 4037–4040.

[13] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, July 2001.

[14] J. D. Polack, "Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 235–244, 1993.

[15] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, U.S.A, Apr. 2008, pp. 4897 –4900.

[16] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[17] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[18] J. Eaton, N.D. Gaubitch, and P.A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[21] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[22] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: a british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, U.S.A, May 1995, pp. 81–84.

[23] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multichannel Wall Street Journal audiovisual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognition and Understanding (ASRU)*, Cancún, Mexico, Dec. 2005, pp. 357–362.

[24] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[25] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[26] 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," Feb. 2001.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Dept, 3.4.1 edition, Dec. 2009.