Meyer *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:14
https://doi.org/10.1186/s13636-020-00180-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Multichannel speaker interference reduction using frequency domain adaptive filtering

Patrick Meyer*, Samy Elshamy and Tim Fingscheidt

Abstract

Microphone leakage or crosstalk is a common problem in multichannel close-talk audio recordings (e.g., meetings or live music performances), which occurs when a target signal does not only couple into its dedicated microphone, but also in all other microphone channels. For further signal processing such as automatic transcription of a meeting, a multichannel speaker interference reduction is required in order to eliminate the interfering speech signals in the microphone channels. The contribution of this paper is twofold: First, we consider multichannel close-talk recordings of a three-person meeting scenario with various different crosstalk levels. In order to eliminate the crosstalk in the target microphone channel, we extend a multichannel Wiener filter approach, which considers all individual microphone channels. Therefore, we integrate an adaptive filter method, which was originally proposed for acoustic echo cancellation (AEC), in order to obtain a well-performing interferer (noise) component estimation. This results in an improved speech-to-interferer ratio by up to 2.7 dB at constant or even better speech component quality. Second, since an AEC method requires typically clean reference channels, we investigate and report findings why the AEC algorithm is able to successfully estimate the interfering signals and the room impulse responses between the microphones of the interferer and the target speakers even though the reference signals are themselves disturbed by crosstalk in the considered meeting scenario.

Keywords: Acoustic echo cancellation, Multichannel interference reduction, Meetings, Social signal processing

1 Introduction

Meetings, here considered as a face-to-face conversation in a meeting room of at least two persons (c.f. Fig. 1), belong to the most natural ways of humans to communicate with each other. Investigations of social behavior or interaction forms in such a meeting have a long research history in psychology [1, 2] and have also become a vital research topic in Computer Science, which is caused by two aspects: First, an automatic meeting analysis facilitates psychological studies which rely on time-consuming transcription and annotation work. Second, human-computer interaction still lacks in interpreting social signals such as irony or emotions. Therefore, the field of social signal processing aims to teach computers how to understand and interpret human social behavior [3–6]. A typical application is a meeting of at least two persons, in which different relevant aspects such as "what is being said?", "who speaks to whom?", "how is spoken?" (emotions), "why is spoken?" (motivation), and others (e.g., the relationship of the participants) are automatically tracked [7]. Answering these questions with an automatic analysis system deals with multiple challenges: spontaneous speech, gestures, multi-talk, or the audio and video recordings themselves, which is why Morgan et al.

*Correspondence: patrick.meyer@tu-bs.de
Institute for Communications Technology, Technische Universität Braunschweig, Schnleinitzstraße 22, 38106, Braunschweig, Germany

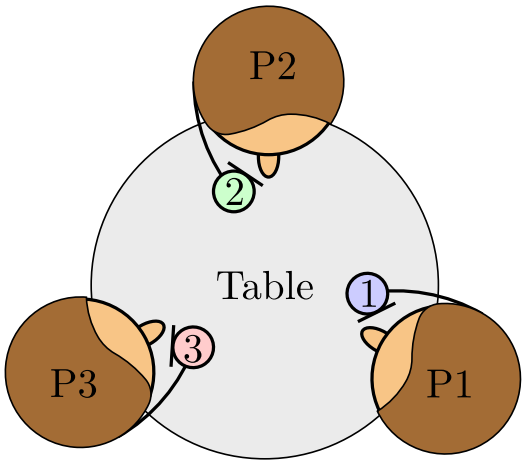

Fig. 1 Investigated meeting scenario with three participants (P1, P2, P3), all wearing a (wireless) headset. For simulation purposes, all persons are represented by a head-and-torso simulator (HATS) with a close-talk microphone placed in front of the mouth

[8] take the view that "nearly every problem in spoken language recognition (and understanding) can be explored in the context of meetings."

Putting the focus on the automatic analysis of audio and speech signals, a lot of research, based on audio recordings of meetings, has already been carried out in the last two decades [9–13]. Thereby, there are three typical methods for data acquisition: a single table-top microphone, a microphone array, or, as illustrated in Fig. 1, personalized close-talk microphones (e.g., headsets or lapel microphones) [13, 14]. In order to process and analyze recorded interactions, first of all, isolated speech of all participants is needed. For this purpose, it is obvious that recordings with a single microphone or a microphone array require blind source separation (BSS) approaches, which have received great attention in a variety of applications over many years [15, 16].

Since we deal with interactive meetings and project work, in which the participants use the entire room including workstations, flip charts, and a meeting desk such as in [17], a single table-top microphone or a microphone array would not be able to deliver sufficient speech quality. Therefore, our research is based on multichannel close-talk recordings, which offer, especially in this case, the best audio quality by recording the individual speech signals robustly with a suitable signal level. Furthermore, distorting room characteristics are mostly negligible. However, even in this case, the *target* speaker channel is disturbed by speech portions of the *interfering* speakers which couple into the target microphone with a non-negligible level. This effect is known as *crosstalk* [8, 18–20] or microphone *leakage* [20, 21] and requires a multichannel speaker interference reduction (MSIR) in order to obtain the desired isolated speech of each person. This issue is getting worse when considering the rate of multi-talk (e.g., double-talk) situations, which occur up to 14% of the time in a professional meeting [22, 23] and can easily exceed 20% in an informal get-together [24]. Since these statistical values are too high to being ignored, and moreover, multi-talk can dramatically decrease the performance of later applications such as automatic speech recognition systems [22, 23], eliminating or rather separating multi-talk situations is one of the key challenges in signal preprocessing for analyzing a meeting.

For the purpose of eliminating crosstalk signals in a microphone channel of interest, we have only access to the microphone channels of all other persons, which, however, are disturbed by crosstalk as well. Even worse, the unknown room impulse responses (RIRs) from the interfering speakers to the target speaker's microphone affect the interfering signals, so that the crosstalk signals, recorded by the target speaker's microphone, differ to quite some extent from the recordings of the interfering persons' microphones. One possible solution for this issue is again BSS approaches, and there are decades of research dealing with multichannel recordings of dedicated microphones [25–29], but the main focus is not on close-talk scenarios.

Another field which is very familiar with close-talk recordings and the effect of microphone leakage is signal processing for (live) music performances [21, 30], where different instruments, each recorded with at least one microphone, are played at the same time. It is a big advantage for the final mix, if the sound engineer has access to the undisturbed microphone signals of each instrument in order to apply reverberation or equalization. Furthermore, microphone leakage can lead to unwanted artifacts such as the *comb filtering* effect, which occurs when mixing a signal and a delayed version of the same signal together [31]. Using directional microphones can decrease the leakage effect, but does not completely eliminate it and, even worse, can cause other artifacts such as the *proximity effect* [32], which describes an energy increase at low frequencies. They are in addition sensitive to microphone orientation, which is why omnidirectional microphones are still a good choice for robust and clean recordings.

In order to reduce the interfering signals in each microphone channel, several approaches have been published in the last years: adaptive filtering in time and frequency domain [30, 33–35] are popular solutions to estimate the room impulse response from the interferer source microphone to the target microphone. These methods often use additional information such as a speaker activity detection or a signal-to-interferer ratio (SIR) for controlling the adaptation process. Other approaches propose nonnegative signal factorization [36], kernel additive modeling [20], or a Gaussian probabilistic framework [37].

Thereby, [20, 30, 34, 35] use iterative or cascading schemes of individual filters for each channel to obtain crosstalk-free interferer channels for further processing.

Kokkinis et al. [21] tackled this problem by a multichannel Wiener filter (MWF), thereby interpreting the problem as a noise suppression task, thus taking a big step forward by ignoring the explicit room characteristics and focusing on the energy of the interfering signals by a simple gain factor. This is primarily motivated by the two facts that music productions commonly use sampling rates of at least 44.1 kHz and deal with reverberation times, especially in live sounds, that can easily exceed 1 s. This in combination can lead to an extremely high number of filter coefficients in order to estimate the RIRs, which is synonymous to a slow convergence behavior and high computational costs. However, calculation power increases steadily, and w.r.t. a meeting, we typically deal with 16 kHz sampling rate and RIR lengths of around 250 ms in a common-sized meeting room [28, 38].

We already proposed a multichannel Kalman-based Wiener filter (MKWF) method [39], which is an extension of the MWF approach of Kokkinis et al. [21], taking into account the characteristics of the RIRs between the microphones of the persons in our meeting scenario. Thereby, we improved the SIR at constant speech component quality during triple-talk. Since Buchner et al. [40] understand acoustic echo compensation (AEC) with clean reference signals as a special case of BSS, we applied an adaptive filter to estimate the RIRs, similar to [30, 33]. We used a multichannel AEC (MAEC) method from Enzner et al. [41], which was developed for a hands-free teleconferencing system to estimate the RIRs from the loudspeakers to the microphone on the basis of clean loudspeaker reference signals. Surprisingly, the MAEC RIR estimation performed well without any preprocessing steps to enhance the disturbed microphone channels. Considering the fact that the MAEC is based on the assumption that the reference signals are clean, some questions remained open.

In this paper, we further enhance the MKWF and precisely analyze why the MAEC RIR estimation method of Enzner et al. [41] is performing well in the considered meeting scenario with crosstalk-disturbed reference channels. First, we briefly recap and extend our proposed MKWF method [39] by a control strategy for the case of interferer speech pauses (ISPs), in which the RIR estimation is lost due to a missing excitation signal. Afterwards, we compare the improved MKWF with the MWF [21] and the MAEC [41] in a more realistic and challenging meeting scenario with multiple RIR changes. Subsequently, we investigate the performance of the MAEC RIR estimation compared to an oracle MAEC, which has access to clean reference signals and points out the main performance differences. Finally, we elaborate why the MAEC RIR estimation can be successfully applied in a meeting scenario with leaky microphone channels.

The outline of this work is as follows: We introduce the considered meeting scenario with some important notations and the problem formulation in Section 2. While Section 3 contains the algorithmic descriptions of the MWF and MAEC baseline approaches, we define our extended MKWF method in Section 4. A comparison of the baselines and the MKWF is carried out in Section 5, which is followed by a more detailed analysis of the MAEC with focus on RIR estimation in Section 6 to answer the question why the MAEC is able to work despite leaky reference channels. The paper is concluded with some remarks in Section 7.

2 Scenario model and data acquisition

First of all, we formulate the problem regarding the considered meeting scenario with some notations and describe the data acquisition process and preparation for the later experiments.

2.1 Considered meeting scenario and problem formulation

We consider a meeting scenario of three persons (P1, P2, P3) sitting at a table and talking to each other as depicted in Fig. 1. Note that most interesting algorithmic aspects can already be investigated with three persons.

In order to analyze or transcribe the course of a meeting's conversation, all persons are equipped with a (wireless) headset for two reasons: on the one hand, a microphone channel with good close-talk audio quality for each person is obtained; on the other hand, the participants are free to stand up, go to the flip chart and workstations, or even walk around, still allowing for high-quality sound acquisition, which is in this case hardly possible with a single table-top microphone or a microphone array. However, for simulations, we consider only fixed positions of the persons at the table, which is already a challenging scenario in terms of interfering speaker levels. Furthermore, we assume the headsets to have an omnidirectional microphone characteristic, supporting robustness w.r.t. the acquisition of the target speaker's voice. However, not only the speech of the target speaker is recorded, but also the speech portions from all other persons. Depending on the position and the loudness level of the other interfering speakers, low- and high-level crosstalk may occur. This undesired effect is well known from audio recordings or the mixing of live sounds and is designated as microphone *leakage*.

In the following, we denote $m \in \mathcal{M} = \{1, \ldots, M\}$ as one specific *target* speaker being currently focused on, and $\mu \in \mathcal{I} = \{1, \ldots, M \,|\, \mu \neq m\}$ as an *interfering* speaker. As shown in Fig. 2, each microphone channel $y_m(n)$ of the associated target speaker m is modeled as

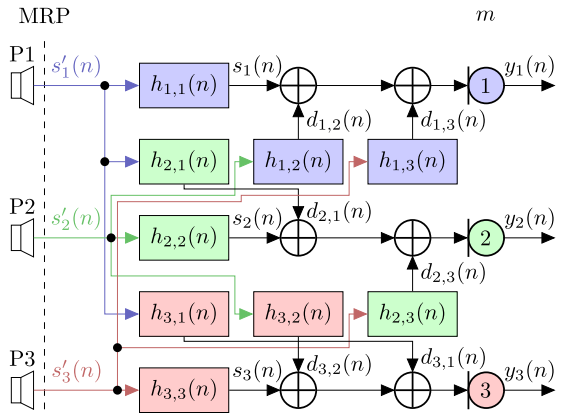

Fig. 2 The acoustic channel model of the considered meeting scenario. Each speaker's voice couples into the close-talk microphones of all other participants. Thereby, the source signal $s'_m(n)$, $m \in \mathcal{M}$, is filtered by different room impulse responses $h_{m,\mu}(n)$ with $m, \mu \in \mathcal{M}$

$$y_m(n) = s_m(n) + \sum_{\mu \in \mathcal{I}} d_{m,\mu}(n), \tag{1}$$

with discrete-time sample index n. Thereby, $s_m(n) = s'_m(n) * h_{m,m}(n)$ denotes the convolution of the source signal $s'_m(n)$ of target speaker m with the corresponding room impulse response (RIR) $h_{m,m}(n)$, describing the path from mouth to microphone. In contrast, the interfering signals are defined by $d_{m,\mu}(n) = s'_\mu(n) * h_{m,\mu}(n)$, with the interferer source signal $s'_\mu(n)$, being convolved with the corresponding RIR $h_{m,\mu}(n)$, $\mu \in \mathcal{I}$, reflecting the acoustic path from the mouth of interferer μ to the microphone of target speaker m.

2.2 Data acquisition and signal processing

A detailed evaluation of applied MSIR methods requires the use of objective measures. Therefore, it is necessary to have access to each individual signal component in (1), more specifically $s_m(n) = s'_m(n) * h_{m,m}(n)$ and $d_{m,\mu}(n) = s'_\mu(n) * h_{m,\mu}(n)$. In order to be able to generate realistic microphone signals of a meeting scenario under these conditions, in line with ITU-T P.1110 [42] and P.1130 [43], we record real RIRs in a typical meeting room which are then used to generate the individual signal components of microphone channel $y_m(n)$ by means of various speech samples.

To allow the acquisition of RIRs containing the typical characteristics of the direct path and the early reflections in a realistic meeting scenario, the participants of the considered meeting are represented by head-and-torso simulators (HATSs), which are equipped with a headset and placed around a table. Thus, we can measure the RIRs from the mouth reference point (MRP) [44] (cf. Fig. 2) of each HATS to the headset microphone of all HATSs.

Due to the considered close-talk scenario, the MRP and the headset microphone of the target speaker are almost located at the same place. Hence, we assume $h_{m,m}(n) = \delta(n)$ in order to simplify the acoustic channel model w.r.t. the investigated conversational group interaction, requiring only the RIRs from the target speaker to all other persons' microphones to be measured. For this purpose, one `Yamaha HS80M` studio monitor loudspeaker, representing the MRP, and two `HEAD acoustics HMS II.6` HATSs were employed as acoustic source and sinks, respectively. While later on in practice the participants shall wear a wireless headset, for data recording and the purpose of this work, we acquired the audio signals of each speaker (HATS) by a wired omnidirectional close-talk `Beyerdynamic MM1` measurement microphone placed at the position of a typical headset microphone in order to exclude transmission effects.

We recorded two sets of RIRs, later denoted as *RIR set I* and *RIR set II*, in the already described three-person scenario in a meeting room of size 6.6 m × 5.75 m × 2.5 m (length × width × height) according to [45]. As excitation, a 48 kHz and 32 bit linear sweep signal from 0.01 Hz to 24 kHz with a length of 10 s was used. The excitation was played back with the studio monitor placed at the position of each speaker and recorded at each of the other speaker's headset position by means of the `Beyerdynamic MM1` microphone. Afterwards, the RIRs were determined with the aid of a recorded electrical reference signal, which was recorded once for all measurements, by a linear deconvolution in the frequency domain [45] and finally downsampled to 16 kHz. The T60 times of the measured RIRs are on average 0.24 s, which is in line with [28, 38] regarding a common meeting room. As a result, the RIR signals were cut off after 4000 samples (0.25 s) for our experiments.

By using the measured RIRs, we are able to simulate any desired dialog between the considered three persons as defined in (1). The simulation diagram structure of all signals for the later experiments according to the acoustic channel model in Fig. 2 is illustrated in detail in Fig. 3. Due to the assumption that $h_{m,m}(n) = \delta(n)$, the speech signal $s'_m(n)$ of the target speaker of microphone channel $y_m(n)$ is not convolved with any RIR. In order to obtain $s_m(n)$, the active speech level (ASL) is scaled with α_{s_m} to −26 dBov in accordance with ITU-T P.56 [46]. In addition, two interferer signals $s'_\mu(n)$ are convolved with $h_{m,\mu}(n)$ and also adjusted to −26 dBov ASL by α_{s_μ}. Afterwards, the two interferer signals are superimposed and jointly scaled with α_d to the desired crosstalk level. Finally, $s_m(n)$ and $d_m(n)$ are superimposed and a white Gaussian noise floor $n'_m(n)$, adjusted to −75 dBov (using α_n), is added to the microphone channel, to simulate some sensor noise. The explicit signal mixtures of the respective experiments follow in the corresponding sections.

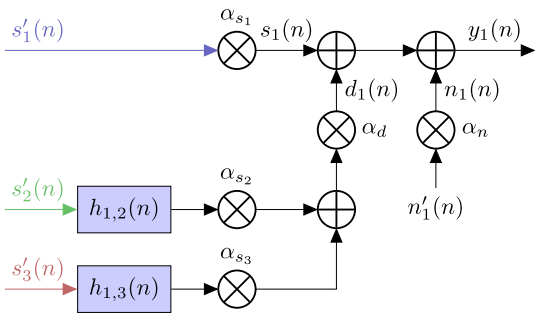

Fig. 3 Simulation diagram of the close-talk microphone signal $y_1(n)$, employing input speech data, measured room impulse responses (RIRs), and assuming $h_{m,m}(n) = \delta(n)$. For computation of microphone signals $y_2(n)$ and $y_3(n)$, the inputs are to be changed accordingly

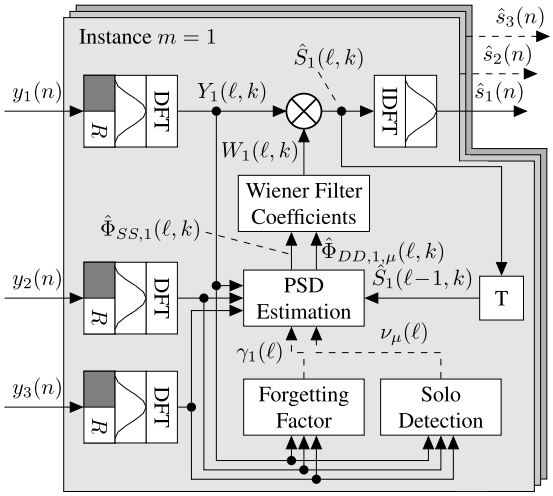

Fig. 4 MWF approach for speaker $m = 1$. The box ■ denotes the buffering of the previous $K-R$ samples. With the aid of a PSD estimation, a Wiener filter $W_1(\ell,k)$ is adapted to reduce the interferer signals in microphone channel $y_1(n)$. As indicated by the three layers, microphone channels $y_2(n)$ and $y_3(n)$ are enhanced by changing the input signals accordingly

3 Baseline approaches

In this section, we present the multichannel Wiener filter (MWF) approach according to [21] and the multichannel acoustic echo cancellation (MAEC) method adopted from [41], applied to the considered meeting scenario. Some mathematical detail is important here to understand, since later on we will refer to it with our proposed approach.

3.1 Multichannel Wiener filter (MWF)

Kokkinis et al. [21] consider microphone leakage effects in music recordings with several instruments, each assigned to a microphone being close by the instrument, resulting in (1). Under the assumption that each microphone captures primarily the audio signal of the assigned audio source and only to some lower extent the interferer sources, a Wiener filter $W_m(\ell,k)$ considering all interfering signals is applied in the discrete Fourier transform (DFT) domain to each microphone channel m in order to reduce the interferer signals according to

$$\hat{S}_m(\ell,k) = W_m(\ell,k) \cdot Y_m(\ell,k). \tag{2}$$

Here, $Y_m(\ell,k)$ is the short-time Fourier transform (STFT) of the target microphone channel $y_m(n)$ and $\hat{S}_m(\ell,k)$ is the estimated STFT of target speech signal $s_m(n)$. Furthermore, $k \in \mathcal{K} = \{0,1,\ldots,K-1\}$ is the discrete frequency bin index. Assuming statistical independence of target speech and interferer signals, the Wiener filter is modeled by

$$W_m(\ell,k) = \frac{\hat{\Phi}_{SS,m}(\ell,k)}{\hat{\Phi}_{SS,m}(\ell,k) + \sum_{\mu\in\mathcal{I}} \hat{\Phi}_{DD,m,\mu}(\ell,k)}, \tag{3}$$

with the estimated power spectral densities (PSDs) $\hat{\Phi}_{SS,m}(\ell,k)$ and $\hat{\Phi}_{DD,m,\mu}(\ell,k)$ of the speech signal of the target speaker and the interferer's, respectively. Since these signals are not accessible, they have to be estimated.

An overview of the MWF method is illustrated in Fig. 4. The input signals $y_m(n)$ constitute the continuation of the output signals of Fig. 2 and are transformed into the DFT domain by using an overlap-add (OLA) structure. The PSD estimation block delivers the update of the Wiener filter coefficients (3) and is utilized for the calculation of both target speaker PSD and interferer PSDs, whereby each has its own control unit in form of the forgetting factor and the solo detection block, respectively, which will be explained later in this section. In the following, the estimations of both the target speaker PSD and the interferer PSDs are described in accordance with [21].

3.1.1 *Estimation of the target speaker PSD $\hat{\Phi}_{SS,m}$*

It is assumed that PSD bins k without an influence of an interferer signal show almost equivalence between the PSDs of the input microphone signal $\hat{\Phi}_{YY,m}(\ell,k)$ and the enhanced output signal $\hat{\Phi}_{\hat{S}\hat{S},m}(\ell,k)$ [21], which are both obtained by squaring the absolute value of $Y_m(\ell,k)$ and $\hat{S}_m(\ell-1,k)$, respectively. We obtain these *dominant* frequency bins $k \in \mathcal{K}^{\mathrm{dom}}(\ell)$ by comparing their *active* frequency bins $k \in \mathcal{K}^{\mathrm{act}}(\ell)$, which are calculated for channel m by

$$\begin{aligned} \mathcal{K}^{\mathrm{act}}_{Y,m}(\ell) &= \{k \in \mathcal{K} \,|\, \hat{\Phi}_{YY,m}(\ell,k) \geq \mathrm{E}_{Y,m}(\ell)\} \\ \mathcal{K}^{\mathrm{act}}_{\hat{S},m}(\ell) &= \{k \in \mathcal{K} \,|\, \hat{\Phi}_{\hat{S}\hat{S},m}(\ell-1,k) \geq \mathrm{E}_{\hat{S},m}(\ell-1)\}, \end{aligned} \tag{4}$$

with $\mathrm{E}_{Y,m}(\ell)$ and $\mathrm{E}_{\hat{S},m}(\ell-1)$ being the root-mean-squared amplitudes of $\hat{\Phi}_{YY,m}(\ell,k)$ and $\hat{\Phi}_{\hat{S}\hat{S},m}(\ell-1,k)$, respectively. Afterwards, the desired dominant frequency bins $\mathcal{K}^{\mathrm{dom}}(\ell)$

are identified as those, which are active in both PSDs according to

$$\mathcal{K}_m^{\text{dom}}(\ell) = \mathcal{K}_{Y,m}^{\text{act}}(\ell) \cap \mathcal{K}_{\hat{S},m}^{\text{act}}(\ell). \tag{5}$$

By means of $\mathcal{K}_m^{\text{dom}}(\ell)$, a binary mask

$$\mathcal{X}_m^{\text{dom}}(\ell,k) = \begin{cases} 1, & k \in \mathcal{K}_m^{\text{dom}}(\ell) \\ 0, & k \notin \mathcal{K}_m^{\text{dom}}(\ell) \end{cases} \tag{6}$$

can be defined, and thus, both a *dominant* PSD component

$$\hat{\Phi}_m^{\text{dom}}(\ell,k) = \mathcal{X}_m^{\text{dom}}(\ell,k) \cdot \delta^{\text{dom}} \cdot \hat{\Phi}_{YY,m}(\ell,k) \tag{7}$$

and a *residual* PSD component

$$\begin{aligned} \hat{\Phi}_m^{\text{res}}(\ell,k) = (1 - \mathcal{X}_m^{\text{dom}}(\ell,k) \cdot \Big(\delta_Y^{\text{res}} \cdot \hat{\Phi}_{YY,m}(\ell,k) \\ + \delta_{\hat{S}}^{\text{res}} \cdot \hat{\Phi}_{\hat{S}\hat{S},m}(\ell-1,k)\Big) \end{aligned} \tag{8}$$

are determined. The parameters δ^{dom}, δ_Y^{res}, and $\delta_{\hat{S}}^{\text{res}}$ take on values between 0 and 1 with $\delta^{\text{dom}} + \delta_Y^{\text{res}} + \delta_{\hat{S}}^{\text{res}} = 1$. Finally, the PSD of the target speaker results in

$$\begin{aligned} \hat{\Phi}_{SS,m}(\ell,k) = \gamma_m(\ell) \cdot \hat{\Phi}_{SS,m}(\ell-1,k) \\ + (1 - \gamma_m(\ell)) \cdot \Big(\hat{\Phi}_m^{\text{dom}}(\ell,k) + \hat{\Phi}_m^{\text{res}}(\ell,k)\Big), \end{aligned} \tag{9}$$

with $\gamma_m(\ell)$ being an energy-adaptive forgetting factor considering the time-varying energy of all microphone channels in order to ensure that the PSD estimation only proceeds, if the target microphone channel has more energy than the others, and hence, there is a high confidence of hardly interfering signals (Fig. 4 block Forgetting Factor).

3.1.2 Estimation of the interferer PSDs $\hat{\Phi}_{DD,m}$

The estimation of $\hat{\Phi}_{DD,m,\mu}(\ell,k)$ is formulated on the basis of the interferer source $s'_\mu(n)$, $\mu \in \mathcal{I}$, which is once convolved with $h_{\mu,\mu}(n)$ to provide $s_\mu(n)$, and once with all $h_{m,\mu}(n)$ to provide the interferer signals $d_{m,\mu}(n)$ (cf. Fig. 2). Hence, by ignoring further interferences, it is assumed for reasons of simplification that the corresponding PSDs $\hat{\Phi}_{SS,\mu}(\ell,k)$ and $\hat{\Phi}_{DD,m,\mu}(\ell,k)$ only differ by a time-variant but full-band factor $\alpha_{m,\mu}(\ell)$, which is obtained by

$$\alpha_{m,\mu}(\ell) = \frac{E_{Y,m}(\ell)}{E_{Y,\mu}(\ell)}. \tag{10}$$

Thereby, $\alpha_{m,\mu}(\ell)$ is only updated in solo intervals (single-talk) of the considered interfering speaker, yielding $E_{D,m}(\ell) = E_{Y,m}(\ell)$ and $E_{S,\mu}(\ell) = E_{Y,\mu}(\ell)$. The solo parts (Fig. 4 block Solo Detection), depicted by $\nu_\mu(\ell)$, are detected by an energy function based on a sigmoid function as well as $E_{Y,m}(\ell)$ and $E_{Y,\mu}(\ell)$. For further details, please refer to [21]. With the aid of $\alpha_{m,\mu}(\ell)$, the interferer PSD is estimated as

$$\hat{\Phi}_{DD,m,\mu}(\ell,k) = \alpha_{m,\mu}(\ell) \cdot \hat{\Phi}_{SS,\mu}(\ell,k). \tag{11}$$

3.2 Multichannel acoustic echo cancellation (MAEC)

The frequency domain adaptive filtering-based MAEC approach by Malik and Enzner [41] has been originally intended for full-duplex hands-free telephony. The basic idea of the MAEC directly applied to our speaker interference scenario is depicted in Fig. 5. Note that Fig. 5 can be seen as a continuation of Fig. 2 with the output microphone signals $y_m(n)$ of Fig. 2 being the input signals to Fig. 5. Note further that different to a classical MAEC scenario as assumed in [41], the reference signals $y_2(n)$ and $y_3(n)$ in our scenario are themselves distorted by each other and, even worse, by the target speech $s'_1(n)$ coupling into $y_2(n)$ and $y_3(n)$ as depicted in Fig. 3. Furthermore, each channel $m \in \mathcal{M}$ has to be enhanced and is independently processed by a basic MAEC approach, depicted by the $M = 3$ instances in Fig. 5.

In the following, we adapt the basic MAEC approach to our meeting scenario in accordance with [47, 48]. The MAEC is implemented in an overlap-save (OLS) structure and is based on a Kalman filter consisting of an alternating prediction and correction step. After windowing of frame length K with frame shift R, we obtain frames of each interfering microphone channel $y_\mu(n)$ (i.e., as reference signal), frame-wise packed in $K \times 1$ vectors $\mathbf{y}_\mu(\ell)$, with frame index ℓ. The first frame is headed by $K -$

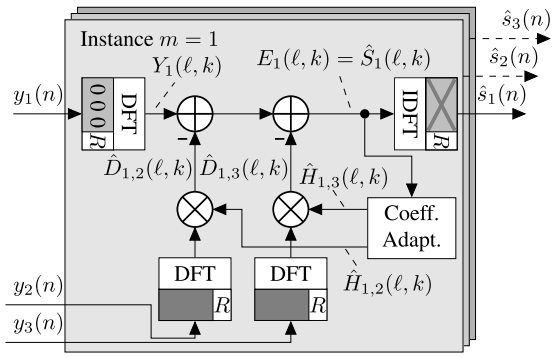

Fig. 5 MAEC approach in the meeting scenario for speaker $m=1$. Thereby, the box ■ denotes again the buffering of the previous $K-R$ samples, while ⊠ shows the elimination of $K-R$ samples w.r.t. the OLS constraint. Additionally, 0 depicts the replacement of $K-R$ samples by zeros. The RIRs $H_{1,2}(\ell,k)$ and $H_{1,3}(\ell,k)$ of interferer signals $S'_2(\ell,k)$ and $S'_3(\ell,k)$ are estimated ($\hat{H}_{1,2}(\ell,k)$, $\hat{H}_{1,3}(\ell,k)$). Afterwards, microphone signals $Y_2(\ell,k)$ and $Y_3(\ell,k)$ are multiplied with $\hat{H}_{1,2}(\ell,k)$ and $\hat{H}_{1,3}(\ell,k)$, respectively, and are then subtracted from the microphone signal $Y_1(\ell,k)$, resulting in $\hat{S}_1(\ell,k)$. By means of the error signal $E_1(\ell,k)$, the estimation of the RIRs is adapted. To obtain $\hat{S}_2(\ell,k)$ and $\hat{S}_3(\ell,k)$, the inputs are changed accordingly

R zeros, followed by the first R samples of the respective microphone signal. Each frame is transformed into the frequency domain and shaped into a main diagonal matrix, to allow for a convenient notation, by

$$\underline{\mathbf{Y}}_{\mu}(\ell) = \mathrm{diag}\{\underline{\mathbf{F}}_{K\times K}\cdot\mathbf{y}_{\mu}(\ell)\}, \tag{12}$$

with $\underline{\mathbf{F}}_{K\times K}$ being the K-point DFT matrix and $\mu\in\mathcal{I}$. In contrast, the $R\times 1$ target microphone channel $\mathbf{y}_m(\ell)$ is processed with the $K\times R$ overlap-save projection matrix $\underline{\mathbf{Q}} = \left(\underline{\mathbf{0}}_{R\times(K-R)}\ \underline{\mathbf{I}}_{R\times R}\right)^T$ and DFT matrix $\underline{\mathbf{F}}_{K\times K}$ resulting in a $K\times 1$ vector as

$$\mathbf{Y}_m(\ell) = \underline{\mathbf{F}}_{K\times K}\underline{\mathbf{Q}}\cdot\mathbf{y}_m(\ell), \tag{13}$$

whereby $\underline{\mathbf{0}}$ and $\underline{\mathbf{I}}$ denote a zero and unity matrix, respectively. The subsequent steps are done for each speaker $m\in\mathcal{M}$, i.e., M instances of the MAEC are in operation.

3.2.1 Prediction step

The prediction of the current $K\times 1$ MAEC filter coefficient state vector $\hat{\mathbf{H}}^{+}_{m,\mu}(\ell)$ of interferer channel $\mu\in\mathcal{I}$ w.r.t. the target microphone channel $m\in\mathcal{M}$ is obtained by

$$\hat{\mathbf{H}}^{+}_{m,\mu}(\ell) = A_{m,\mu}\hat{\mathbf{H}}_{m,\mu}(\ell-1), \tag{14}$$

whereby $(\cdot)^{+}$ indicates a prediction and $A_{m,\mu}$ is a first-order Markov model prediction coefficient (also denoted as forgetting factor) of interferer μ in instance m, initialized with $A_{m,\mu} = 0.998$. Furthermore, by means of an overestimation factor λ, the $K\times K$ covariance matrices of the state error for both intra-channels ($\mu=\nu$) and cross-channels ($\mu\neq\nu$) are predicted with $\mu,\nu\in\mathcal{I}$ by

$$\begin{aligned}\underline{\mathbf{P}}^{+}_{m,\mu,\nu}(\ell) &= A_{m,\mu}A_{m,\nu}\underline{\mathbf{P}}_{m,\mu,\nu}(\ell-1)\\ &\quad+\lambda\underline{\Psi}^{\Delta}_{m,\mu,\nu}(\ell-1).\end{aligned} \tag{15}$$

The same-size covariance submatrices $\underline{\Psi}^{\Delta}_{m,\mu,\nu}$ denote the process noise and are determined by

$$\begin{aligned}\underline{\Psi}^{\Delta}_{m,\mu,\nu}(\ell-1) = (1-A^2_{m,\mu})\Big(&\hat{\mathbf{H}}_{m,\mu}(\ell-1)\hat{\mathbf{H}}^{H}_{m,\nu}(\ell-1)\\ &+\underline{\mathbf{P}}_{m,\mu,\nu}(\ell-1)\Big),\end{aligned} \tag{16}$$

initialized with $\underline{\mathbf{P}}_{m,\mu,\nu}(\ell=0) = \underline{\mathbf{I}}_{K\times K}$, $\underline{\Psi}^{\Delta}_{m,\mu,\nu}(\ell=0) = \underline{\mathbf{0}}_{K\times K}$ for $\mu\neq\nu$, and $\underline{\Psi}^{\Delta}_{m,\mu,\nu}(\ell=0) = \underline{\mathbf{I}}_{K\times K}$ for $\mu=\nu$. Note that in accordance with [47], only the intra-channels ($\mu=\nu$) are computed in (16), whereby $(\cdot)^{H}$ designates the Hermitian transpose.

3.2.2 Correction step

With the aid of the $K\times K$ overlap-save constraint matrix $\underline{\mathbf{G}} = \underline{\mathbf{F}}_{K\times K}\underline{\mathbf{Q}}\cdot\underline{\mathbf{Q}}^{T}\underline{\mathbf{F}}^{-1}_{K\times K}$, the *preliminary error vector* [48] (DFT coefficient vector of length K) is obtained as

$$\tilde{\mathbf{E}}_m(\ell) = \mathbf{Y}_m(\ell) - \sum_{\mu\in\mathcal{I}}\underline{\mathbf{G}}\cdot\underline{\mathbf{Y}}_{\mu}(\ell)\hat{\mathbf{H}}^{+}_{m,\mu}(\ell). \tag{17}$$

Subsequently, the preliminary error signal, weighted with the Kalman gain diagonal $K\times K$ matrix $\underline{\mathbf{K}}_{m,\mu}(\ell)$, is used to correct (update) the predicted MAEC filter coefficient states to obtain

$$\hat{\mathbf{H}}_{m,\mu}(\ell) = \hat{\mathbf{H}}^{+}_{m,\mu}(\ell) + \underline{\mathbf{K}}_{m,\mu}(\ell)\tilde{\mathbf{E}}_m(\ell), \tag{18}$$

with the initialization of $\hat{\mathbf{H}}_{m,\mu}(\ell=0) = \underline{\mathbf{0}}_{K\times 1}$. Besides, the state error covariance prediction matrix is updated as well by using the Kalman gain as

$$\underline{\mathbf{P}}_{m,\mu,\nu}(\ell) = \underline{\mathbf{P}}^{+}_{m,\mu,\nu}(\ell) - \frac{R}{K}\underline{\mathbf{K}}_{m,\mu}(\ell)\sum_{\kappa\in\mathcal{I}}\underline{\mathbf{Y}}_{\kappa}(\ell)\underline{\mathbf{P}}^{+}_{m,\kappa,\nu}(\ell). \tag{19}$$

Thereby, the Kalman gain diagonal matrix is defined by

$$\underline{\mathbf{K}}_{m,\mu}(\ell) = \sum_{\nu\in\mathcal{I}}\underline{\boldsymbol{\mu}}_{m,\mu,\nu}(\ell)\underline{\mathbf{Y}}^{H}_{\nu}(\ell), \tag{20}$$

with the $K\times K$ diagonal matrix

$$\underline{\boldsymbol{\mu}}_{m,\mu,\nu}(\ell) = \frac{R}{K}\underline{\mathbf{P}}^{+}_{m,\mu,\nu}(\ell)\underline{\mathbf{B}}^{-1}_{m}(\ell), \tag{21}$$

being the step-size for the Kalman gain. Furthermore, the diagonal matrix $\underline{\mathbf{B}}_m(\ell)$ of the target microphone channel m results in

$$\underline{\mathbf{B}}_m(\ell) = \frac{R}{K}\sum_{\mu\in\mathcal{I}}\sum_{\nu\in\mathcal{I}}\underline{\mathbf{Y}}_{\mu}(\ell)\underline{\mathbf{P}}^{+}_{m,\mu,\nu}(\ell)\underline{\mathbf{Y}}^{H}_{\nu}(\ell) + \underline{\Psi}^{S}_{m}(\ell) \tag{22}$$

and includes the covariance diagonal matrix $\underline{\Psi}^{S}_{m}(\ell)$ of the measurement noise, which indicates the presence of *near-end speech* and is determined by a temporal smoothing as

$$\begin{aligned}\underline{\Psi}^{S}_{m}(\ell) = \beta\cdot\underline{\Psi}^{S}_{m}(\ell-1) + (1-\beta)\cdot\Big(&\tilde{\mathbf{E}}_m(\ell)\tilde{\mathbf{E}}^{H}_{m}(\ell)\\ &+\frac{R}{K}\sum_{\mu\in\mathcal{I}}\sum_{\nu\in\mathcal{I}}\underline{\mathbf{Y}}_{\mu}(\ell)\underline{\mathbf{P}}^{+}_{m,\mu,\nu}(\ell)\underline{\mathbf{Y}}^{H}_{\nu}(\ell)\Big),\end{aligned} \tag{23}$$

initialized with $\underline{\Psi}^{S}_{m}(\ell=0) = \underline{\mathbf{0}}_{K\times K}$. Finally, using (18), the error signal, namely the estimated target speaker's signal, can be computed and the time-domain signal $\hat{s}_m(n)$ is recovered by overlap-save synthesis based on

$$\hat{\mathbf{S}}_m(\ell) = \mathbf{E}_m(\ell) = \mathbf{Y}_m(\ell) - \sum_{\mu\in\mathcal{I}}\underline{\mathbf{G}}\cdot\underline{\mathbf{Y}}_{\mu}(\ell)\hat{\mathbf{H}}_{m,\mu}(\ell). \tag{24}$$

4 Multichannel Kalman-based Wiener filter (MKWF)

Investigations on the MWF and MAEC (as presented in Section 5.3) result in two observations: first, the MWF has a bigger potential for a high interferer reduction compared to the MAEC, and second, the MAEC achieves a better and more homogeneous quality of the remaining target speech signal over a wide range of different SIRs. Based on these observations, we start in this section with the MKWF approach according to [39], which is an extension

of the MWF. Thereby, the MKWF considers the influence of the real RIRs on the interferer signals to obtain a better interferer PSD estimation. Thus, the quality of the remaining target speech is improved, since this allows a more precise filtering (over all frequencies) of the leaky microphone channels, instead of just using a single fullband gain factor (c.f. (10)). For this purpose, we replace the MWF-interferer PSD estimation (c.f. Section 3.1.2) by applying the Kalman filter of the MAEC to estimate the interferer (noise) PSD $\hat{\Phi}_{DD,m}(\ell,k)$. We further improve the estimation of the target PSD $\hat{\Phi}_{SS,m}(\ell,k)$ by using the output signals of the MAEC and integrate a new extended control strategy for the RIR update of the MAEC to be able to deal with interferer speech pauses. In the following, we describe the proposed MKWF, which is depicted in Fig. 6.

In line with the MWF, the Wiener filter is modeled by

$$W_m(\ell,k) = \frac{\hat{\Phi}_{SS,m}(\ell,k)}{\hat{\Phi}_{SS,m}(\ell,k) + \sum_{\mu\in\mathcal{I}} \hat{\Phi}_{DD,m,\mu}(\ell,k)}. \tag{25}$$

To obtain $\hat{\Phi}_{DD,m,\mu}(\ell,k)$, the determination of the interferer signals in channel m by means of the MAEC is already defined in (24) as

$$\hat{\mathbf{D}}_{m,\mu}(\ell) = \underline{\mathbf{G}} \cdot \underline{\mathbf{Y}}_{\mu}(\ell)\hat{\mathbf{H}}_{m,\mu}(\ell). \tag{26}$$

Since the MAEC uses in contrast to the MWF an OLS structure, we first have to adapt the MAEC output. Therefore, we calculate

$$\hat{\mathbf{d}}_m(\ell) = \underline{\mathbf{F}}_{K\times K}^{-1} \cdot \sum_{\mu\in\mathcal{I}} \hat{\mathbf{D}}_{m,\mu}(\ell) \tag{27}$$

and retain only the last R samples of $\hat{\mathbf{d}}_m(\ell)$ due to the OLS constraint, yielding $\hat{d}_m(n)$ (cf. Fig. 6). Afterwards, $\hat{d}_m(n)$ is transformed into the frequency domain, delivering $\hat{D}_m^{\mathrm{WF}}(\ell,k)$, by applying an OLA structure with a Hann window, frame shift R, and frame length $K_{WF} = 2R$. Subsequently, the interferer signal PSD determination is done by

$$\hat{\Phi}_{\hat{D}\hat{D},m}(\ell,k) = \left|\hat{D}_m^{\mathrm{WF}}(\ell,k)\right|^2. \tag{28}$$

The estimation of the target speaker PSD $\hat{\Phi}_{SS,m}(\ell,k)$ follows the MWF approach. In order to make the estimation of $\hat{\Phi}_{SS,m}(\ell,k)$ more robust to low-SIR input signals, we subtract the estimated interferer signals from the target microphone signal, before calculating the PSD of $Y_m^{\mathrm{WF}}(\ell,k)$ according to

$$\hat{\Phi}_{YY,m}(\ell,k) = \left|Y_m^{\mathrm{WF}}(\ell,k) - \hat{D}_m^{\mathrm{WF}}(\ell,k)\right|^2. \tag{29}$$

Following Section 3.1, (4) to (8) and neglecting the forgetting factor results in

$$\hat{\Phi}_{SS,m}(\ell,k) = \hat{\Phi}_m^{\mathrm{dom}}(\ell,k) + \hat{\Phi}_m^{\mathrm{res}}(\ell,k). \tag{30}$$

4.1 New extended RIR update control strategy

The MAEC algorithm contains an intelligent RIR update function, which is mainly based on $\underline{\Psi}_m^S(\ell)$ (23) and $\underline{\mathbf{P}}_{m,\mu,\nu}^{+}(\ell)$ (15), both indirectly controlling the step-size $\underline{\boldsymbol{\mu}}_{m,\mu,\nu}(\ell)$ (21). Thus, the adaptation of the RIR filter coefficients depends on the presence of target speech and the state error. Nevertheless, in case of interferer speech pauses, there is no excitation for the MAEC to estimate

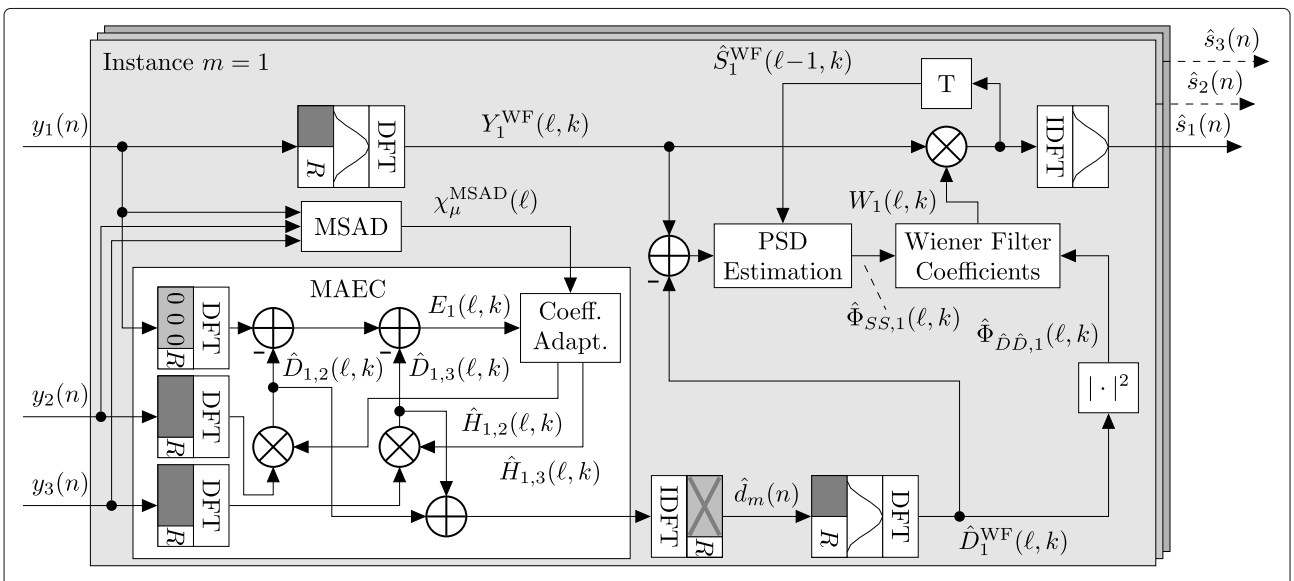

Fig. 6 Proposed multichannel Kalman-based Wiener filter (MKWF) for interferer reduction in channel $m = 1$. The box ■ denotes the previous $K-R$ and $K_{\mathrm{WF}}-R$ samples in the buffer of the MAEC and WF, respectively; ⊠ the elimination of $K-R$ samples w.r.t. the OLS constraint, and 0 the replacement of $K-R$ samples by zeros. The inputs are changed accordingly to obtain $\hat{s}_2(n)$ and $\hat{s}_3(n)$

the RIRs. In such a case, the current RIR estimate is lost and converges to zero due to the integrated update function aiming to protect the target speech signal. This in consequence leads to a permanent reconvergence of the estimation process in a meeting if a person does not speak continuously, and thus, only a suboptimal result is obtained during the beginning of a new utterance.

In order to prevent this behavior, we detect the speech activity of the interferer signals and store the corresponding latest filter coefficients of $\hat{\mathbf{H}}_{m,\mu}(\ell)$ (18) as well as $\underline{\mathbf{P}}_{m,\mu,\nu}(\ell)$ (19) during active speech of interferer μ. After a speech pause of at least 0.5 s of interferer μ, the stored filter coefficients are restored if the interferer starts to speak again. Thereby, the internal RIR update function of the MAEC is not interrupted during interferer speech pauses to prevent the target speech signal. Since RIRs can change rapidly over time, the restored RIRs may be incorrect to some extent, but as we will show in the experimental evaluation, it still reduces the required time for the MAEC to reconverge in a common meeting scenario.

To detect the speech activity in a channel, we apply a multichannel speaker activity detection (MSAD), which is inspired by [49] and briefly described in the following. The MSAD is based on a comparison of the PSDs of all considered microphone channels $m \in \mathcal{M}$. Frames are obtained with a Hann window, a frame shift R, and a frame length $K^{\text{MSAD}} = 2R$. We determine the PSD comparison for channel m by

$$\text{SPR}_m(\ell,k) = 10\log_{10}\left(\frac{\hat{\Phi}_{\Sigma\Sigma,m}(\ell,k)}{\max\limits_{\mu\in\mathcal{I}}\left\{\hat{\Phi}_{\Sigma\Sigma,\mu}(\ell,k)\right\}}\right), \quad (31)$$

with $\hat{\Phi}_{\Sigma\Sigma,m}(\ell,k) = \hat{\Phi}_{YY,m}(\ell,k) - \hat{\Phi}_{NN,m}(\ell,k)$, whereby $\hat{\Phi}_{NN,m}(\ell,k)$ being the noise signal PSD of channel m. For more details, please refer to [49]. We further calculate a signal-to-noise ratio (SNR) as

$$\hat{\xi}_m(\ell,k) = \frac{\max\left\{\min\left\{\hat{\Phi}_{YY,m}(\ell,k), |Y_m(\ell,k)|^2\right\} - \hat{\Phi}'_{NN,m}(\ell,k), 0\right\}}{\hat{\Phi}'_{NN,m}(\ell,k)}, \quad (32)$$

with $\hat{\Phi}'_{NN,m}(\ell,k) = \lambda^{\text{SNR}}\hat{\Phi}_{NN,m}(\ell,k)$ and $\lambda^{\text{SNR}} = 4$ being an overestimation factor to be more robust during speech pauses. By means of $\text{SPR}_m(\ell,k)$ and $\hat{\xi}_m(\ell,k)$, we determine all relevant frequency bins in channel m by

$$\mathcal{K}^+_m(\ell) = \{k \in \mathcal{K} \,|\, \text{SPR}_m(\ell,k) > 0,\ \hat{\xi}_m(\ell,k) \geq \vartheta^{\text{SNR}}\}, \quad (33)$$

with $\vartheta^{\text{SNR}} = 0.25$. Thus, we obtain a soft full-band MSAD by

$$0 \leq \chi_m^{\text{MSAD}'}(\ell) = G^{(B)}_{\min,m}(\ell)\cdot\kappa^+_m(\ell) \leq 1, \quad (34)$$

with $\kappa^+_m(\ell) = |\mathcal{K}^+_m(\ell)|/K^{\text{MSAD}}$ and

$$G^{(B)}_{\min,m}(\ell) = \min\left\{\alpha\cdot\hat{\xi}^{(B)}_m(\ell), 1\right\}, \quad (35)$$

being an SNR-dependent weighting function with $\alpha = 0.1$. Thereby,

$$\hat{\xi}^{(B)}_m(\ell) = \max_{b\in\mathcal{B}}\left\{\frac{1}{|\mathcal{K}_b|}\cdot\sum_{k\in\mathcal{K}_b}\hat{\xi}_m(\ell,k)\right\}, \quad (36)$$

which depicts the maximum SNR value of $B = 10$ averaged frequency bands with index $b \in \mathcal{B} = \{1,2,...,B\}$ and $\mathcal{K}_b$ being the set of frequency bin indices k in band b. Finally, we obtain a MSAD decision for channel m by

$$\chi_m^{\text{MSAD}}(\ell) = \begin{cases} 1, \text{ if } \chi_m^{\text{MSAD}'}(\ell) > \theta^{\text{MSAD}} \\ 0, \text{ otherwise,} \end{cases} \quad (37)$$

with $\theta^{\text{MSAD}} = 0.2$. The values of λ^{SNR}, ϑ^{SNR}, α and θ^{MSAD} were determined empirically.

5 Experiments and discussion

We first introduce the applied evaluation metrics for all upcoming experiments and define the experimental setup, before we discuss the performance comparison between the MWF, MAEC, and MKWF methods.

5.1 Quality measures

With the aid of the PEASS toolbox[1] according to [50], each estimated target signal $\hat{s}_m(n)$ is decomposed into a *target distortion* $e_m^{\text{target}}(n)$, an *interference* $e_m^{\text{interf}}(n)$, and an *artifact* $e_m^{\text{artif}}(n)$ component, defined by

$$\hat{s}_m(n) - s_m(n) = e_m^{\text{target}}(n) + e_m^{\text{interf}}(n) + e_m^{\text{artif}}(n). \quad (38)$$

Based on these signals, four source separation evaluation metrics are calculated by the PEASS toolbox, which are briefly introduced in the following. The measures are defined w.r.t. target channel m. First, the signal-to-interferer ratio (SIR) is determined by

$$\text{oSIR}_m = 10\log_{10}\frac{\sum\limits_{n\in\mathcal{N}}|s_m(n)+e_m^{\text{target}}(n)|^2}{\sum\limits_{n\in\mathcal{N}}|e_m^{\text{interf}}(n)|^2}, \quad (39)$$

with oSIR_m being the output SIR after applying PEASS for channel m and sample index $n \in \mathcal{N} = \{1,\ldots,N\}$. We then define the improvement of the SIR as

$$\Delta\text{SIR}_m = \text{oSIR}_m - \text{iSIR}_m, \quad (40)$$

whereby iSIR_m is the input SIR of channel m, measured according to ITU-T P.56 [46]. The further measures of the

[1] http://bass-db.gforge.inria.fr/peass/

PEASS toolbox are the signal-to-distortion ratio (SDR), the source image to spatial distortion ratio (ISR), and the signal-to-artifact ratio (SAR) [50], which are defined as

$$\mathrm{SDR}_m = 10\log_{10}\frac{\sum_{n\in\mathcal{N}}|s_m(n)|^2}{\sum_{n\in\mathcal{N}}|\hat{s}_m(n)-s_m(n)|^2}, \tag{41}$$

$$\mathrm{ISR}_m = 10\log_{10}\frac{\sum_{n\in\mathcal{N}}|s_m(n)|^2}{\sum_{n\in\mathcal{N}}|e_m^{\mathrm{target}}(n)|^2}, \tag{42}$$

$$\mathrm{SAR}_m = 10\log_{10}\frac{\sum_{n\in\mathcal{N}}|s_m(n)+e_m^{\mathrm{target}}(n)+e_m^{\mathrm{interf}}(n)|^2}{\sum_{n\in\mathcal{N}}|e_m^{\mathrm{artif}}(n)|^2}, \tag{43}$$

respectively. Please note, since a subjective perception of the enhanced target signals is not the primary evaluation criterion for the purpose of an automatic meeting analysis, we do not consider further perceptual measures from the PEASS toolbox for the evaluation.

5.2 Experimental setup

During a group interaction with three persons, at any time, one out of four different states may occur: *silence* (nobody speaks), *single-talk* (only one person speaks), *double-talk* (two persons speak), and *triple-talk*. Thereby, the elimination of crosstalk during multi-talk situations is obviously the most challenging task, since both the interfering and the target speaker are talking at the same time. This requires precise filtering of the recorded microphone signals and is quite similar to the case of common (live) music performances, in which most instruments are playing at the same time. Especially, short interruptions occur very often in a meeting and depict a challenging double-talk task for MSIR methods due to the short adaptation time. For these reasons, our main focus is on multi-talk scenarios.

In order to generate a challenging meeting scenario that is as real as possible and has a focus on multi-talk situations, we created a conversation for target speaker $m = 1$ with explicit single-, double-, and triple-talk parts as well as short speech pauses and multiple RIR changes. We used the speech signals from the ITU-T Recommendation P.501 [51] for the implementation, which are designed for challenging double-talk scenarios in the field of echo compensation in telephonometry.

The composition of the $m = 3$ source signals is as follows: the target signal consists of the 10 s long *female short conditioning sequence*, a *speech pause* of around 6 s, and the *single-talk sequence* with a length of 35 s. Interferer signal $\mu = 2$ begins with the *male short conditioning sequence* of 10 s, followed by the *female short conditioning sequence*, which was cut off after 6 s, and ends with the 35 s long *double-talk sequence*. The second interferer signal $\mu = 3$ is generated with 10 s of *speech pause*, the *female short conditioning sequence*, where the first 4 s were cut off, again *speech pause* of 12 s, and the *female long conditioning sequence* with a duration of around 23 s (c.f. Fig. 8).

The microphone signals are obtained in accordance with Section 2.2 (c.f. Fig. 3) by means of the recorded RIRs. We insert multiple RIR changes at different points in time for each speaker by changing between RIR *set I* and RIR *set II* of the corresponding crosstalk signals of the respective speaker (for a visualization of the crosstalk dependencies w.r.t. the microphone channels, please refer to Fig. 2). The RIRs from target speaker $m = 1$ to channel $\mu = 2$ and $\mu = 3$ are changed after 7, 18, 40, and 47 s; RIRs from speaker $\mu = 2$ to channel $m = 1$ and $\mu = 3$ are changed after 3, 12, 26, and 47 s; and the RIRs from speaker $\mu = 3$ to channel $m = 1$ and $\mu = 2$ are changed after 28, 30, and 47 s. For a better overview, all changing points are marked in Fig. 8 with the corresponding speech source color. The last changing point after 47 s is colored in black, because at this point all applied RIRs are changed simultaneously. Microphone channels are mixed as described in Section 2.2 with the recorded RIR sets; we level the target speaker signals to −26 dBov and the sensor noise to −75 dBov. Furthermore, we investigate crosstalk levels between −26 and −46 dBov with a step-size of 2 dB. All applied signals are sampled at 16 kHz.

The parameter configuration of the MAEC and the MWF is in line with [48] and [21], respectively. Due to the reduced sampling frequency, we only adapt the frame length of the MWF to 512 samples and the frame shift to 256 samples. Moreover, the applied MSAD obtains an accuracy between 76.3% for iSIR = 0 dB and 85.0% for iSIR = 20 dB, with a maximum value of 86.2% for iSIR = 12 dB w.r.t the interferer channels $\mu \in \{2, 3\}$. We also integrated the MSAD-based extended RIR update control mechanism to the MAEC for better comparability. In contrast, replacing the energy-based solo detection of the MWF (c.f. Fig. 4) by the MSAD leads to a significant drop of the ΔSIR performance, so that we did not apply the MSAD to the MWF approach. To ensure fair comparison, we further apply the same parameter values for all common parameters of our MKWF method and the two other approaches. An overview of all parameters for each method is given in Table 1.

5.3 Results and discussion

Figure 7 illustrates the performance comparison of the MWF (red dash-dotted line), MAEC (blue dashed line), and MKWF (green solid line) for target channel $m = 1$ in the considered scenario regarding the four evaluation

Table 1 Parameter configuration of the MWF, MAEC, and MKWF methods

Method	K (K_{WF})	R	λ	β	δ^{dom}	δ_Y^{res}	δ_S^{res}
MWF	512	256	–	–	0.6	0.25	0.15
MAEC	1024	256	1.5	0.5	–	–	–
MKWF	1024 (512)	256	1.5	0.5	0.6	0.25	0.15

metrics from the PEASS toolbox as a function of a wide range of crosstalk levels (iSIR) from 0 to 20 dB. Please note, the discussion of the oracle MAEC (skyblue dotted line) will be the topic of Section 6.1 in the context of a deeper analysis of the RIR estimation by means of the MAEC.

Comparing the MAEC and the MWF, it is evident that the MAEC reaches better results of the ΔSIR for the iSIR range of 0 to 8 dB as well as 17 dB and higher, while the MWF outperforms the interferer reduction of the MAEC by up to 1.2 dB for 8 dB < iSIR < 17 dB and obtains a higher maximum interferer reduction than the MAEC. Furthermore, the MAEC outperforms the MWF significantly regarding SDR, ISR, and SAR for iSIR > 12 dB by up to 6.7 dB, 13.7 dB, and 4.9 dB, respectively. This is due to the fact that the MWF operates very aggressively in high iSIR conditions, whereby it also affects the target speech component negatively. Interestingly, the MWF reaches better ISR and SAR results for low iSIR conditions (iSIR < 12 dB) compared to the MAEC, which is, however, mainly due to the significantly stronger drop of the ΔSIR performance for this iSIR range. To conclude, the MAEC achieves better speech quality of the remaining target speech, while the MWF has more potential for a higher interferer reduction.

The MKWF outperforms both the MAEC and the MWF in almost all concerns. It achieves better results over the whole considered iSIR range regarding the ΔSIR and SDR, whereby it improves the baselines by up to 2.7 dB and 2.3 dB for ΔSIR and SDR, respectively. An exception is the performance of the MWF regarding the ISR and the SAR for iSIR < 8 dB, in which the MKWF obtains a somewhat poorer performance than the MWF, while achieving an improved ΔSIR performance of up to 4 dB at the same time. However, the MKWF outperforms the MWF significantly as well as the MAEC regarding all considered measures for iSIR > 8 dB, which depicts the most relevant range for a common meeting scenario. In addition, shifting the optimal operating point towards a lower iSIR is, besides the significant ΔSIR increase while maintaining approximately equal or even better speech quality (SDR, ISR), the main advantage of the proposed MKWF method compared to the MAEC and the MWF.

The effect of the extended RIR update control strategy (c.f. Section 4.1) w.r.t. the (oracle) MAEC and the proposed MKWF approach is depicted in Table 2. All results are averaged over the iSIR conditions from 0 to 20 dB of the considered scenario, and the results are additionally compared with the MWF method. It is evident that the use of the extended RIR update control significantly improves the performance of the MAEC, MKWF, and oracle MAEC regarding both the ΔSIR and the SDR measure. Thereby, the MKWF achieves the biggest improvement of around

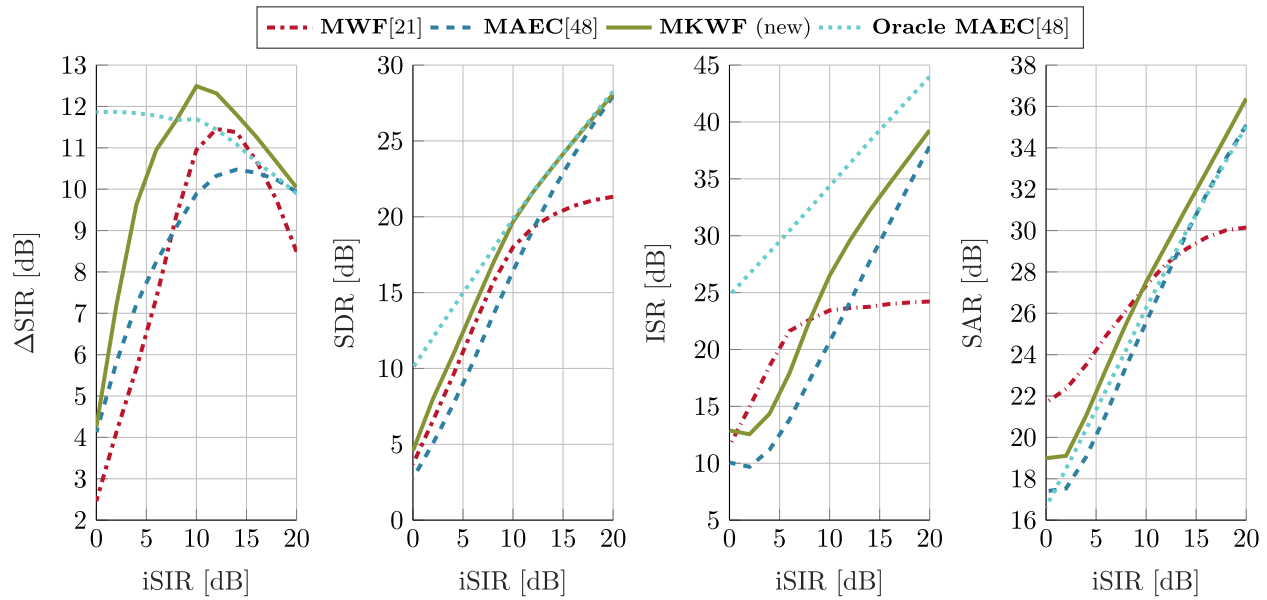

Fig. 7 Performance evaluation of the MWF, MAEC, new MKWF, and oracle MAEC in the considered meeting scenario. The abscissa denotes the iSIR [dB], while the ordinates show from left to right ΔSIR [dB], SDR [dB], ISR [dB], and SAR [dB], respectively. Figure best seen on the screen

Table 2 Performance evaluation of the extended RIR update control strategy w.r.t. the MAEC, MKWF, and oracle MAEC approach. MWF results are provided for better comparison. All results are averaged over all iSIR conditions in the considered meeting scenario. First and second ranked results are bold face and italicized, respectively (oracle results excluded)

Approach	Extended RIR update control	$\overline{\Delta\text{SIR}}$ [dB]	$\overline{\text{SDR}}$ [dB]	$\overline{\text{ISR}}$ [dB]	$\overline{\text{SAR}}$ [dB]
MWF	–	8.3	15.3	21.1	26.7
MAEC	✓	*8.8*	15.9	21.5	25.6
	–	6.2	15.1	24.9	26.8
MKWF	✓	**10.3**	**18.0**	*25.2*	*27.2*
	–	7.2	*16.3*	**28.9**	**28.3**
Oracle MAEC	✓	11.3	19.5	34.4	26.0
	–	10.3	18.7	34.1	25.7

3.1 dB and 1.7 dB w.r.t. the ΔSIR and SDR, respectively. This improvement is obtained, since the implemented control strategy prevents the reconvergence process after an interferer speech pause by restoring the filter coefficients of the last active interferer speech frame, as already mentioned in Section 4.1. An in-depth analysis of this issue can be found in Section 6.1. As expected, due to the stronger interference reduction by means of the extended RIR update control strategy, the performance decreases w.r.t. the ISR and SAR measure for both the MAEC and the MKWF. Interestingly, the ISR and SAR performance of the oracle MAEC is improved by the extended RIR update control. However, the averaged MKWF results with $\overline{\text{ISR}} = 25.2$ dB and $\overline{\text{SAR}} = 27.2$ dB are still the best compared to the results of the MWF and the MAEC (with and without using the extended RIR update control).

6 Analysis of the MAEC RIR estimation with leaky reference channels

Section 5.3 has shown that using the MAEC RIR estimation improves the MWF approach and, also, that the overall performance of the MAEC is quite good in the considered meeting scenario, which is somehow surprising, since the MAEC was not developed to deal with leaky reference channels. In order to understand these results and verify the usability of the MAEC RIR estimation for the proposed MKWF, we analyze the MAEC in this section in more detail and answer two main questions—first: *How does the MAEC actually operate with leaky instead of clean reference channels?* (addressed in Section 6.1); second: *Why does the target speech remain undistorted even though it is present in the reference channels?* (addressed in Section 6.2).

6.1 Influence of leaky reference channels on the MAEC

To answer the first question, we compare the applied MAEC with an oracle MAEC, which processes the same target microphone channel, but has access to crosstalk-free reference channels. Thus, we obtain performance results of our target channel for an echo cancellation scenario, which, considering the fact that the MAEC was developed for echo cancellation, depicts a kind of upper quality limit for the MAEC in the depicted meeting scenario.

Comparing the MAEC to the oracle MAEC (c.f. Fig. 7), two main observations are evident: First, the performance of the MAEC for iSIR $>$ 16 dB is approximately equal to the oracle MAEC, which is conclusive, since the input signals of the MAEC are converging towards the oracle signals due to the decreasing distortion. Only the ISR measure shows quite a gap due to the target speech occurring as crosstalk in the reference channels, which has a very slight effect on the target speech in the target microphone channel, especially when the system is still in the initial convergence process. Second, the ΔSIR of the oracle MAEC for iSIR $<$ 14 dB is significantly higher than that of the MAEC, where the performance achieves a maximum of about 11.9 dB for iSIR $=$ 0 dB. An iSIR of 0 dB means that the source signal and the crosstalk signals share the same active speech level. Thus, the algorithm must not estimate the compensation of level mismatches with the RIR but can focus on the room characteristics without paying attention to a global attenuation or gain factor. Even though the MAEC is theoretically able to determine the gain factor, these types of calibration are common in practice to achieve the best possible result. In contrast, the MAEC cannot benefit from this balanced signal levels in the considered meeting scenario, since the accompanying distortion level is too high for a good performance. However, an iSIR $<$ 5 dB is already a very challenging task and depicts a lower limit for a meeting scenario in practice. Nevertheless, even if the ΔSIR result of the MAEC decreases with an increasing crosstalk level (distortion of reference channels), it still achieves adequate results in this range compared to the MWF.

In order to get a deeper insight into the MAEC RIR estimation, we consider two MAEC metrics. The system distance in [dB], which is defined by

$$d^{\text{sys}}_{m,\mu}(\ell) = 10\log_{10}\frac{||\mathbf{h}_{m,\mu}(\ell) - \hat{\mathbf{h}}_{m,\mu}(\ell)||^2}{||\mathbf{h}_{m,\mu}(\ell)||^2}, \tag{44}$$

as well as the averaged absolute value of the measurement noise, which is averaged over all frequency bins (main diagonal of (23)) and determined by

$$\overline{|\Psi|^S_m}(\ell) = \frac{1}{K}\cdot\text{tr}\left(\left|\underline{\Psi}^S_m(\ell)\right|\right), \tag{45}$$

with tr(·) being the trace of the matrix. Figure 8 depicts the time course of $d^{\text{sys}}_{m=1,\mu}(\ell)$ and $\overline{|\Psi|^S_{m=1}}(\ell)$ for both the MAEC (solid line) and the oracle MAEC (dotted line)

for target channel $m = 1$ of the considered meeting scenario at iSIR = 10 dB. Thereby, $\overline{|\Psi|^S_{m=1}}(\ell)$ is smoothed over time for better illustration. Furthermore, to verify that the MAEC RIR estimation is able to deal with time-variant RIRs, we consider multiple RIR changes of different crosstalk signals in Fig. 8 (c.f. Section 5.2 and Fig. 2), each carried out in a specific meeting situation. The RIR changes are illustrated in the lower plot of Fig. 8 by abbreviations in colored circles. Thereby, circle colors mark the changed RIRs w.r.t. the corresponding speech source color and the abbreviations for single-talk (ST), double-talk (DT), triple-talk (TT), and interferer speech pause (ISP) specify the respective meeting situation (e.g., ST depicts single-talk and a change of the RIRs $h_{1,2}(n)$ and $h_{3,2}(n)$). The black colored circle at 47 s defines a simultaneous change of all RIRs.

It can be seen that the plots of $\overline{|\Psi|^S_{m=1}}(\ell)$, which indicate the presence of target speech, are very similar for the MAEC and the oracle MAEC. The system distances of the MAEC are also pretty good compared to the oracle MAEC, the latter of course obtaining still better results, which is in line with Fig. 7. By looking at the system distance, we can get a picture of some general characteristic behaviors of the (oracle) MAEC.

First, RIR changes emanating from the target signal $m = 1$ ($h_{3,1}(n)$ and $h_{2,1}(n)$) seem to have no influence on the RIR estimation of $h_{m=1,\mu}(n)$. This is indicated by the three RIR changes ST at 7, 18, and 40 s, since there is no effect on the system distance $d^{\text{sys}}_{m=1,\mu}(\ell)$, $\mu \in \{2, 3\}$ for the (oracle) MAEC.

Second, the best results w.r.t. $d^{\text{sys}}_{m=1,\mu}(\ell)$ are obtained during (single) interferer-only talk (e.g., in the intervals 0 s ... 5 s or 10 s ... 16 s). This is shown for the MAEC by the fast reconvergence process from −4 dB back to −11 dB in less than 2 s w.r.t $d^{\text{sys}}_{1,2}(\ell)$ for RIR change ST at 3 s. Similar results are obtained for RIR change DT at 12 s, where $d^{\text{sys}}_{1,2}(\ell)$ of the MAEC decreases again fast and is close to the performance of the oracle MAEC. In contrast, estimating $h_{1,3}(n)$ after DT takes a little longer for the MAEC, which is mainly due to the fact that this is the initial convergence process that takes place during DT.

Third, the system distance for both the MAEC and the oracle MAEC decreases slower or remains approximately equal during large periods of target speech (e.g., in the intervals 30 s ... 32 s, 42 s ... 44.5 s, and 47 s ... 51 s). Thus, the reconvergence process is significantly slower after the RIR changes DT and TT at 30 and 47 s, respectively, compared to ST and DT. This is due to the fact that target speech represents already a distortion to the RIR estimation process of the (oracle) MAEC. These sections are typically marked by huge values of the measurement noise (indicating the presence of target speech), which leads to a small value of the step-size for the Kalman gains (c.f. (21)). Nevertheless, the system distance of the (oracle) MAEC still decreases in all considered cases, so that the functionality of the RIR estimation is ensured.

Finally, the system distance increases during speech pauses of the interferer (e.g., in the intervals 5 s ... 10 s, 15 s ... 19 s, or 31.5 s ... 42 s). This is consistent, since in this case the algorithm has no excitation to estimate the RIRs, which is the main reason why we integrated the extended RIR update control strategy (c.f. Section 4.1) into both the MKWF and the (oracle) MAEC. The positive effect is illustrated in the lowest plot of Fig. 8. It can be seen that the restored RIR leads to a very fast reconvergence rate after ISPs w.r.t. $d^{\text{sys}}_{1,2}(\ell)$ for the estimate of $h_{1,2}(n)$ at 10 s and in the interval 19 s ... 25 s. Even though the RIR change ISP at 26 s makes the stored RIR $\hat{h}_{1,2}(n)$ obsolete, restoring $\hat{h}_{1,2}(n)$ at 26.5 s (when interferer $\mu = 2$ starts talking again) still decreases $d^{\text{sys}}_{1,2}(\ell)$ by around 0.6 dB and has no disadvantages compared to the basic MAEC method that does not use the extended RIR update control. On the contrary, it is quite obvious that restoring the RIR $\hat{h}_{1,3}(n)$ at 30 s, which is obsolete after RIR change ISP at 29 s, leads clearly to a better RIR estimation by 1.7 dB w.r.t. $d^{\text{sys}}_{1,3}(\ell)$ compared to the basic MAEC without using the extended RIR update control. This indicates a certain dependency between different RIRs inside the same environment (meeting room), so that it is a benefit to store the latest RIR during active speech instead of starting the complete initialization process of the MAEC again.

We can conclude from Fig. 8 that the MAEC RIR estimation is still completely operational during a crosstalk level of an iSIR = 10 dB, even if the distortion by the crosstalk leads to some minor performance limitations compared to the oracle MAEC (c.f. Fig. 7). In summary, since in a meeting scenario we typically deal with iSIR ≥ 5 dB, by considering the results of Figs. 7 and 8, we can assume that the MAEC RIR estimation is suitable for the proposed MKWF and for this kind of application.

6.2 Preservation of the target speech by the MAEC

So far, we know how the microphone leakage of the reference channels influences the adaptation process and thus also the performance of the MAEC RIR estimation. The remaining question "*how does the MAEC in general distinguish between the target and the interferer signals?*" is being answered in the following.

In order to understand this behavior, we have to investigate the influence of a RIR on our source signals.

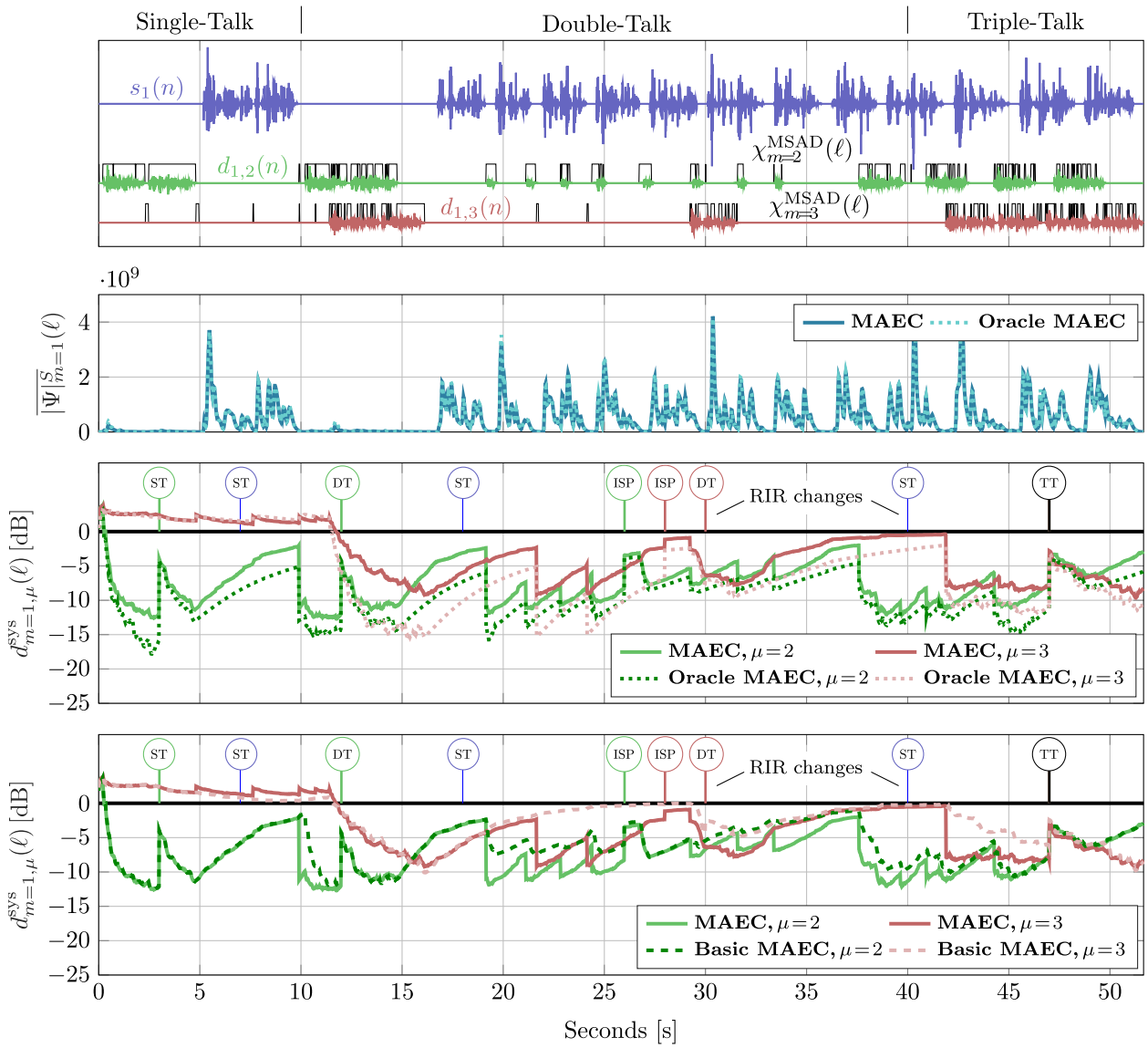

Fig. 8 Comparison of the averaged and time-smoothed measurement noise $\overline{|\Psi|^S_{m=1}}(\ell)$ and the system distance $d^{\mathrm{sys}}_{m=1,\mu}(\ell)$ of the MAEC and the oracle MAEC, respectively, for one specific dialog at iSIR = 10 dB. From top to bottom the target speech signal $s_1(n)$, both interferer speech signals $d_{1,\mu=2}(n)$ and $d_{1,\mu=3}(n)$ are visualized with the corresponding MSAD output $\chi^{\mathrm{MSAD}}_m(\ell) \in \{0,1\}$. The measurement noise component $\overline{|\Psi|^S_{m=1}}(\ell)$ of the MAEC and oracle MAEC are denoted with a blue solid line and a dotted skyblue line, respectively. Finally, the associated system distance $d^{\mathrm{sys}}_{m=1,\mu}(\ell)$ of both interferer signals $\mu \in \{2,3\}$ are colored according to the interfering signals, whereby dotted lines represent the results of the oracle MAEC, dashed lines the results of the basic MAEC without using the extended RIR update control strategy, and solid lines the results of the MAEC using the extended RIR update control strategy. Figure best seen on the screen

Therefore, we create a single-talk scenario based on speech samples of the NTT multi-lingual speech database [52] and use a very much simplified RIR, which is represented by an impulse $\alpha \cdot \delta(n - n_0)$:

- Attenuation: $\alpha < 1, n_0 = 0$
- Amplification: $\alpha > 1, n_0 = 0$
- Delay: $\alpha = 1, n_0 > 0$

By superposition and concatenation of such impulses, we can model any discrete RIR. This includes also some reverberation, which corresponds to a sequence $h(n)$, with $n \in \mathbb{N}_0$ and $h(n) \in \mathbb{R}$. Table 3 depicts the performance of the MAEC for this single-talk experiment, in which all speaker signals $s'_m(n)$ are convolved with the described simplified RIRs (ASL of all speakers is adjusted to −26 dBov before the convolution) by coupling

Table 3 MAEC performance results in case of a single-talk scenario with $|\mathcal{M}| = 3$ speakers with various RIRs

Room impulse response		ΔSIR [dB]	SDR [dB]	ISR [dB]	SAR [dB]
None	$\alpha = 1, n_0 = 0$	−4.96	0.22	0.30	9.78
Amplification	$\alpha > 1, n_0 = 0$	−20.39	−0.61	0.07	13.20
Attenuation	$\alpha < 1, n_0 = 0$	11.27	1.72	1.77	14.08
Delay	$\alpha = 1, n_0 > 0$	6.72	5.33	10.85	20.99
Reverberation		5.20	2.37	5.36	20.53

as interferer signals into the non-dedicated microphone channels $y_\mu(n)$ (c.f. Section 2.2). Thereby, we either *attenuate* or *amplify* the interfering signals by −5 and +5 dB, respectively, or choose a *delay* of $n_0 = 100$ samples (which is below the frame shift $R = 256$). Additionally, we use a random sequence to simulate *reverberation*, which is shaped by an exponentially decreasing function to impose an energy decay with a reverberation time of $T_{60} = 5$ ms, and truncate it after 6.25 ms (i.e., after 100 samples). The RIRs from the target speakers to their dedicated microphone channels $y_m(n)$ remain with $h_{m,m}(n) = \delta(n)$.

Even though we already know that the crosstalk level has an effect on the performance of the MAEC during multi-talk (cf. Fig. 7), it is clear to see in Table 3 that this is not the reason why the MAEC is able to work with leaky reference signals. To be more specific, in the case of single-talk, no RIR or an RIR only attenuating or amplifying the crosstalk signals leads to almost an elimination of the target speech signal, which is evident for these three types of RIRs in Table 3 from the poor performances regarding the SDR and the ISR measures. Nevertheless, the attenuating RIR obtains the best results of these three RIRs.

However, the key factor for the MAEC with leaky reference signals seems to be *the delay of the interfering signals*, as is evident from the best-performing result w.r.t. both SDR and ISR measures in Table 3. This can be explained as follows: All interfering signals (especially the crosstalk of the target speech in the reference channels) are delayed w.r.t. their source signals. In consequence, the MAEC would have to estimate an RIR with negative delay in order to affect the target speech component in the desired channel m. But this is physically not possible and causes the MAEC to treat the target speech component as *near-end speech* (as mentioned in Section 3.2), which might be the reason why the MAEC (RIR estimation) works fine without degrading the target speech. Since the reverberation can be seen as a combination of delay and level adjustment, it is obvious that this has also a positive effect on the MAEC (RIR estimation) in our scenario.

We can conclude that the MAEC (RIR estimation) can be applied to close-talk multichannel recordings with crosstalk-disturbed reference channels, if the microphones are closest to their dedicated person.

7 Conclusions

In this work, we investigated the applicability of a multichannel acoustic echo cancellation (MAEC) approach for speaker interference reduction in a close-talk (wireless) headset meeting scenario, which deals with crosstalk and thus reference channels disturbed by both target and interferer speech. We further show that the characteristics of the room impulse response (RIR), especially the delay, and during multi-talk to some extent also the attenuation affecting the energy level of the crosstalk, are the reasons why the MAEC is able to operate successfully with crosstalk-disturbed reference signals in this specific scenario. Moreover, by means of the MAEC RIR estimation, we propose a multichannel Kalman-based Wiener filter (MKWF) method, which is an extension of a multichannel Wiener filter (MWF) approach by considering the RIRs between the microphones of the interferer and the target speakers. Thus, the MKWF estimates the interfering signals more precisely, leading to an increase of up to 2.7 dB signal-to-interferer ratio, while the obtained speech quality remains equal or is even better compared to the MWF and the MAEC.

Abbreviations

AEC: Acoustic echo cancellation; ASL: Active speech level; BSS: Blind source separation; DFT: Discrete Fourier transform; DT: Double-talk; HATS: Head-and-torso simulator; ISP: Interferer speech pause; ISR: Source image to spatial distortion ratio; MAEC: Multichannel acoustic echo cancellation; MKWF: Multichannel Kalman-based Wiener filter; MRP: Mouth reference point; MSAD: Multichannel speaker activity detection; MSIR: Multichannel speaker interference reduction; MWF: Multichannel Wiener filter; OLA: Overlap-add; OLS: Overlap-save; PSD: Power spectral density; RIR: Room impulse response; SAR: Signal-to-artifact ratio; SDR: Signal-to-distortion ratio; SIR: Signal-to-interferer ratio; ST: Single-talk; STFT: Short-time Fourier transform; TT: Triple-talk

Acknowledgements

The authors would like to thank Elias K. Kokkinis for providing the source code of the MWF approach.

Authors' contributions

The major contributions to the manuscript are from PM. SE and TF supervised the experimental work and polished the structure as well as the text of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

P.501 test signals for use in telephonometry [51] are available from the ITU-T Recommendation P.501. The NTT multi-lingual speech database [52] may be available from NTT-AT upon request. Recorded room impulse responses are available from the authors upon request.

Competing interests

The authors declare that they have no competing interests.

Received: 27 February 2020 Accepted: 9 September 2020
Published online: 04 November 2020

References

1. R. F. Bales, *Interaction Process Analysis: A method for the study of small groups*. (Addison-Wesley, Cambridge, 1950)
2. J. E. McGrath, *Groups: Interaction and Performance*. (Prentice-Hall, Englewood Cliff, 1984)
3. J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, A. Vinciarelli, *Social signal processing*. (Cambridge University Press, Cambridge, 2017)
4. A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
5. A. Vinciarelli, M. Pantic, D. Heylen, D. Pelachaud, I. Poggi, F. D'Errico, M. Schröder, Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans. Affect. Comput. **3**(1), 69–87 (2012)
6. D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis. Comput. **27**(12), 1775–1787 (2009)
7. A. Waibel, H. Yu, M. Westphal, H. Soltau, T. Schultz, Y. Pan, F. Metze, M. Bett, in *Proc. of Int. Conf. on Human Language Technology Research*. Advances in meeting recognition (Association for Computational Linguistics (ACL), San Diego, 2001), pp. 1–3
8. N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, in *Proc. of ICASSP*. Meetings about meetings: research at ICSI on speech in multiparty conversations (IEEE, Hong Kong, 2003), pp. 740–743
9. N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, First DIHARD challenge evaluation plan. Tech. Rep. (2018). https://zenodo.org/record/1199638. Accessed May 2019
10. X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, N. Sebe, SALSA: a novel dataset for multimodal group behavior analysis. IEEE Trans. Pattern. Anal. Mach. Intell. **38**(8), 1707–1720 (2016)
11. S. Renals, T. Hain, H. Bourlard, in *Proc. of ASRU*. Recognition and understanding of meetings: the AMI and AMIDA projects, (Kyoto, 2007), pp. 238–247. https://doi.org/10.1109/asru.2007.4430116
12. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, P. Wellner, in *Proc. of Int. Workshop on Machine Learning for Multimodal Interaction*. The AMI meeting corpus: a pre-announcement, (Edinburgh, 2005), pp. 28–39. https://doi.org/10.1007/11677482_3
13. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, in *Proc. of ICASSP*. The ICSI meeting corpus, (Hong Kong, 2003), pp. 364–367. https://doi.org/10.1109/icassp.2003.1198793
14. I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, D. Zhang, Automatic analysis of multimodal group actions in meetings. IEEE Trans. Pattern. Anal. Mach. Intell. **27**(3), 305–317 (2005)
15. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
16. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
17. A.-K. Bavendiek, L. Thiele, P. Meyer, T. Vietor, S. Kauffeld, T. Fingscheidt, in *Proc. of ICED*. Meetings in the product development process: applying design methods to improve team interaction and meeting outcomes (Design Society, Milano, 2015), pp. 319–328
18. T. Pfau, D. P. W. Ellis, A. Stolcke, in *Proc. of ASRU*. Multispeaker speech activity detection for the ICSI meeting recorder (IEEE, Madonna di Campiglio, 2001), pp. 107–110
19. S. N. Wrigley, G. J. Brown, V. Wan, S. Renals, Speech and crosstalk detection in multichannel audio. IEEE Trans. Speech Audio Process. **13**(1), 84–91 (2005)
20. T. Prätzlich, R. M. Bittner, A. Liutkus, M. Müller, in *Proc. of ICASSP*. Kernel additive modeling for interference reduction in multi-channel music recordings (IEEE, Brisbane, 2015), pp. 584–588
21. E. K. Kokkinis, J. D. Reiss, J. Mourjopoulos, A Wiener filter approach to microphone leakage reduction in close-microphone applications. IEEE Trans. Audio Speech Lang. Process. **20**(3), 767–779 (2012)
22. E. Shriberg, A. Stolcke, D. Baron, in *Proc. of Eurospeech*. Observations on overlap: findings and implications for automatic processing of multi-party conversation (ISCA, Aalborg, 2001), pp. 1359–1362
23. Ö. Cetin, E. Shriberg, in *Proc. of ICASSP*. Speaker overlaps and ASR errors in meetings: effects before, during, and after the overlap (IEEE, Toulouse, 2006), pp. 357–360
24. T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, in *Proc of ICASSP*. All-neural Online Source Separation, Counting, and Diarization for Meeting Analysis, (Brighton, 2019), pp. 91–95
25. E. Weinstein, M. Feder, A. V. Oppenheim, Multi-channel signal separation by decorrelation. IEEE Trans. Speech Audio Process. **1**(4), 405–413 (1993)
26. T.-W. Lee, A. J. Bell, R. H. Lambert, in *Proc. of NIPS*. Blind separation of delayed and convolved sources (NIPS Foundation, Denver, 1996), pp. 758–764
27. L. Parra, C. Spence, Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Process. **8**(3), 320–327 (2000)
28. J. P. Dmochowski, Z. Liu, P. A. Chou, in *Proc. of ICASSP*. Blind source separation in a distributed microphone meeting environment for improved teleconferencing (IEEE, Las Vegas, 2008), pp. 89–92
29. K. Ochi, N. Ono, S. Miyabe, S. Makino, in *Proc. of Interspeech*. Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage (ISCA, San Francisco, 2016), pp. 3369–3373
30. A. Clifford, J. Reiss, in *Proc. of Int. Conf. on Digital Audio Effects (DAFx-11)*. Microphone interference reduction in live sound, (Paris, 2011), pp. 2–9
31. A. Clifford, J. Reiss, in *Proc of AES Convention*. Calculating time delays of multiple active sources in live sound (Audio Engineering Society (AES), San Francisco, 2010), pp. 1–8
32. A. Clifford, J. Reiss, in *AES Convention*. Proximity effect detection for directional microphones (Audio Engineering Society (AES), New York City, 2011)
33. C. Uhle, J. Reiss, in *Proc. of AES Convention*. Determined source separation for microphone recordings using IIR filters (Audio Engineering Society (AES), San Francisco, 2010), pp. 1–14
34. A. Lombard, W. Kellermann, in *Proc. of IWAENC*. Multichannel cross-talk cancellation in a call-center scenario using frequency-domain adaptive filtering (IEEE, Seattle, 2008), pp. 14–17
35. T. Matheja, M. Buck, T. Fingscheidt, A dynamic multi-channel speech enhancement system for distributed microphones in a car environment. EURASIP J. Adv. Signal Process. **191**, 1–21 (2013)
36. J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, F. J. Rodríguez-Serrano, Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings. EURASIP J. Adv. Signal Process. **184**, 1–16 (2013)
37. D. D. Carlo, K. Déguernel, A. Liutkus, in *Proc. of AES International Conference on Semantic Audio*. Gaussian framework for interference reduction in live recordings (Audio Engineering Society (AES), Erlangen, 2017), pp. 1–8
38. M. Jeub, M. Schäfer, P. Vary, in *Proc. of Int. Conf. on Digital Signal Processing*. A binaural room impulse response database for the evaluation of dereverberation algorithms (IEEE, Santorini-Hellas, 2009), pp. 1–5
39. P. Meyer, S. Elshamy, T. Fingscheidt, in *Proc. of ICASSP*. A multichannel Kalman-based Wiener filter approach for speaker interference reduction in meetings (IEEE, Barcelona, 2020), pp. 451–455
40. H. Buchner, W. Kellermann, in *Proc. of HSCMA*. A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation (IEEE, Trento, 2008), pp. 17–20
41. G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones. Signal Process. **86**(6), 1140–1156 (2006)
42. ITU, Rec. P.1110: Wideband hands-free communication in motor vehicles. International telecommunication union, telecommunication standardization sector (ITU-T) (2015). International Telecommunication Union, Telecommunication Standardization Sector (ITU-T)
43. ITU, Rec. P.1130: Subsystem requirements for automotive speech services. International telecommunication union, telecommunication standardization sector (ITU-T) (2015). International telecommunication union, telecommunication standardization sector (ITU-T)
44. ITU, Rec. P.58: Head and torso simulator for telephonometry. International telecommunication union, telecommunication standardization sector (ITU-T) (2013). International telecommunication union, telecommunication standardization sector (ITU-T)
45. S. Müller, P. Massarani, Transfer-function measurement with sweeps. J. Audio Eng. Soc. **49**(6), 443–471 (2001)

46. ITU, Rec. P.56: Objective measurement of active speech level. International telecommunication union, telecommunication standardization sector (ITU-T) (2011). International telecommunication union, telecommunication standardization sector (ITU-T)
47. S. Malik, G. Enzner, Recursive Bayesian control of multichannel acoustic echo cancellation. IEEE Signal Process. Lett. **18**(11), 619–622 (2011)
48. M.-A. Jung, S. Elshamy, T. Fingscheidt, in *Proc. of EUSIPCO*. An automotive wideband stereo acoustic echo canceler using frequency-domain adaptive filtering (IEEE, Lisbon, 2014), pp. 1452–1456
49. P. Meyer, R. Jongebloed, T. Fingscheidt, in *Proc. of ICASSP*. Multichannel speaker activity detection for meetings, (Calgary, 2018), pp. 5539–5543
50. V. Emiya, E. Vincent, N. Harlander, V. Hohmann, Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2046–2057 (2011)
51. ITU, Rec. P.501: Test signals for use in telephonometry. International telecommunication union, telecommunication standardization sector (ITU-T) (2017). International telecommunication union, telecommunication standardization sector (ITU-T)
52. NTT, Multi-lingual speech database for telephonometry. NTT Advanced Technology Corporation (1994). NTT Advanced Technology Corporation

Publisher's Note