Multichannel Acoustic Echo Cancellation Applied to Microphone Leakage Reduction in Meetings

Patrick Meyer, Samy Elshamy, Jan Franzen, Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunchweig
38106 Braunschweig, Germany
Email: {patrick.meyer, s.elshamy, j.franzen, t.fingscheidt}@tu-bs.de

***Abstract*—Microphone leakage occurs in multichannel close-talk audio recordings of a meeting, when speech of an active speaker couples into both the dedicated target microphone and all other microphone channels. For an automatic transcription or analysis of a meeting, the interferer signals in the target microphone channels have to be eliminated. Therefore, we apply a frequency domain adaptive filtering-based multichannel acoustic echo cancellation (MAEC) method, which typically requires clean reference channels. We consider a wide range of different speech-to-interferer ratios and evaluate two cascading schemes for the MAEC, which leads to an improved speech component quality and interferer reduction by up to $0.1\,\mathrm{MOS}$ points and $0.5\,\mathrm{dB}$, respectively. However, the purpose of this work is not to improve the MAEC method, but instead to show that it can be successfully applied to microphone leakage reduction, such as in meetings with headset-equipped participants. Therefore, we analyze and point out why the MAEC method is able to cancel the interferer signals in this scenario even though the reference signals are themselves disturbed by interfering speech portions.**

***Index Terms*—speaker interference reduction, Kalman filter, meeting, social signal processing, crosstalk, microphone leakage**

I. Introduction

Social signal processing denotes the automatic analysis of interpersonal communications, relationships, and behaviors [1–3]. These analyses are commonly based on a meeting of more than two people, since meetings are a natural form of communication. An automatic analysis of such a meeting contains a lot of challenges like spontaneous speech, multi-talk or the audio recordings themselves, which is why ”nearly every problem in spoken language recognition (and understanding) can be explored in the context of meetings” [4].

With the aim of analyzing audio signals, it is common to use a single table-top microphone, a microphone array, or personalized close-talk microphones (e.g., headsets as depicted in Fig. 1) to record a meeting [5, 6]. In this work, we focus on multichannel close-talk recordings to obtain high quality and robust speech even in interactive meetings, in which the participants are free to stand up and use the entire room (e.g., for using the flip-chart). However, the microphone channel of the target speaker is still disturbed by interfering speech from all other participants coupling into the target microphone channel with a non negligible level. This effect is known as *crosstalk* or microphone *leakage* [4, 7–9] and requires a reduction of the interferer speech signals to improve the speech intelligibility and to allow further signal processing, especially in multi-talk situations [10, 11].

Microphone leakage is also well known in live music performances and recordings [12, 13] or in call-centers [14]. Hence, several approaches have been published in the last years, dealing with the enhancement of the target microphone channel in such a scenario: Adaptive filtering in time [12] and frequency domain [9, 13–15] are popular solutions, using additional information like a speaker activity detection or a signal-to-interferer ratio (SIR) in order to control the adaptation. Furthermore, [9, 12, 14, 15] use the crosstalk-resistant adaptive noise canceler (CTRANC) scheme [16]. It describes the cascading of individual filters for each channel in order to obtain crosstalk-free interferer channels for further processing. Opposite to that, in [13] a frequency domain Wiener filter approach is proposed, which is based on power spectral densities to estimate the target and interfering speech portions. Thereby, all enumerated approaches assume that the interfering crosstalk level is lower than the target speech component, which can lead to problems in practice. Further methods are based on nonnegative matrix factorization [17] or machine learning [18]. For the sake of completeness it may be mentioned that the field of blind source separation is also related to our problem formulation [19].

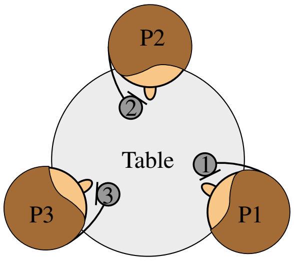

Fig. 1. **Considered meeting scenario** with three persons (P1, P2, P3), all wearing a headset and talking to each other.

Based on our previous work and findings in [20, 21], in this paper, we apply a frequency domain adaptive Kalman filter [22, 23], which was developed for multichannel acoustic echo cancellation (MAEC), directly to our microphone leakage problem. We investigate various crosstalk levels of the interfering speech signals and analyze why the MAEC approach can be applied to this complex scenario, even in conditions with negative SIR $[\mathrm{dB}]$ and without using the CTRANC scheme or similar. Moreover, we compare the performance of the MAEC to an idealized scenario, in which the reference (interferer) channels are not disturbed, and thereby quantify an upper performance bound. Finally, we evaluate two cascading strategies including both MAEC and a CTRANC scheme to improve the performance in our scenario.

The paper is organized as follows: We introduce our considered meeting scenario in Section II. Afterwards, we describe briefly the applied MAEC algorithm and different cascading strategies in Section III. The experimental evaluation is done in Section IV, before we conclude this paper with some remarks in Section V.

II. Scenario Model and Notations

The considered meeting scenario is depicted in Fig. 1, in which three persons (P1, P2, P3) are sitting around a table and talking to

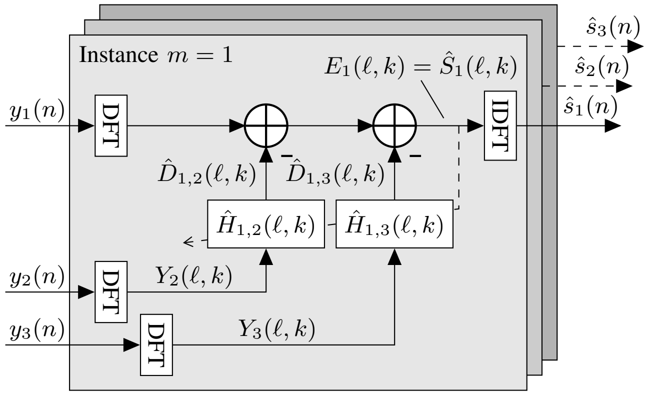

Fig. 2. **MAEC approach** in the meeting scenario for speaker $m = 1$. The RIRS $H_{1,2}(\ell,k)$ and $H_{1,3}(\ell,k)$ of interferer signals $S'_2(\ell,k)$ and $S'_3(\ell,k)$ are estimated ($\hat{H}_{1,2}(\ell,k)$, $\hat{H}_{1,3}(\ell,k)$). Afterwards, microphone signals $Y_2(\ell,k)$ and $Y_3(\ell,k)$ are multiplied with $\hat{H}_{1,2}(\ell,k)$ and $\hat{H}_{1,3}(\ell,k)$, respectively, and are then subtracted from the source 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. In order to obtain $\hat{S}_2(\ell,k)$ and $\hat{S}_3(\ell,k)$, the inputs are changed accordingly.

each other. In order to obtain good and robust close-talk speech quality, all persons are equipped with a headset and are recorded in an individual channel. However, microphone leakage occurs. In the following, we define the considered person as *target* speaker $m \in \mathcal{M} = \{1, 2, ..., M\}$, and $\mu \in \mathcal{I} = \{1, 2, ..., M | \mu \neq m\}$ as *interfering* speaker. The microphone channels $y_m(n)$ of the related target speakers m are modeled as

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

with sample index n and $n_m(n)$ being some noise in channel $y_m(n)$. Furthermore, $s_m(n) = s'_m(n) * h_{m,m}(n)$ defines the convolution of the target speech signal with the room impulse response (RIR) of the acoustic path between the mouth and microphone of target speaker m. On the contrary, $d_{m,\mu}(n) = s'_\mu(n) * h_{m,\mu}(n)$ denotes the interfering signals, which are convolved with the corresponding RIR from the mouth of interferer μ to the microphone of target speaker m.

III. FDAF-based MAEC

In this section, we briefly introduce the MAEC approach in the context of our meeting scenario and propose subsequently two iterative schemes in order to improve the performance of the MAEC.

A. Algorithmic Approach

On the basis of the clean reference signals $s'_\mu(n)$ (loudspeaker signals), the MAEC algorithm estimates the RIRs $h_{m,\mu}(n)$ resulting in $\hat{h}_{m,\mu}(n)$. Thereby, $\hat{d}_{m,\mu}(n) = s'_\mu(n) * \hat{h}_{m,\mu}(n)$ can be calculated and afterwards used in order to enhance $y_m(n)$ to $\hat{s}_m(n)$ by subtracting $\hat{d}_{m,\mu}(n)$ from $y_m(n)$. The algorithm is based on a Kalman filter and estimates $h_{m,\mu}(n)$ in the frequency domain by means of a prediction and a correction step. The ability to detect and distinguish near-end speech from the reference signal is a certain strength of this method. For reasons of space, the authors refer to [22–24] for a detailed mathematical description of the MAEC.

In analogy to the MAEC problem, we interpret the two interfering speaker signals $s'_2(n)$ and $s'_3(n)$ as reference signals in order to cancel them out of our target speaker's microphone signal $y_m(n)$. However, $s'_2(n)$ and $s'_3(n)$ are not available. Therefore, we use the only available signals $y_2(n)$ and $y_3(n)$, which themselves are disturbed by each other, and, even worst, by the target speech, respectively, as the reference signals for the MAEC algorithm as depicted in Fig. 2. Furthermore, each channel $m \in \mathcal{M}$ has to be enhanced and is independently processed by a basic MAEC approach, depicted by the $|\mathcal{M}| = 3$ instances in Fig. 2.

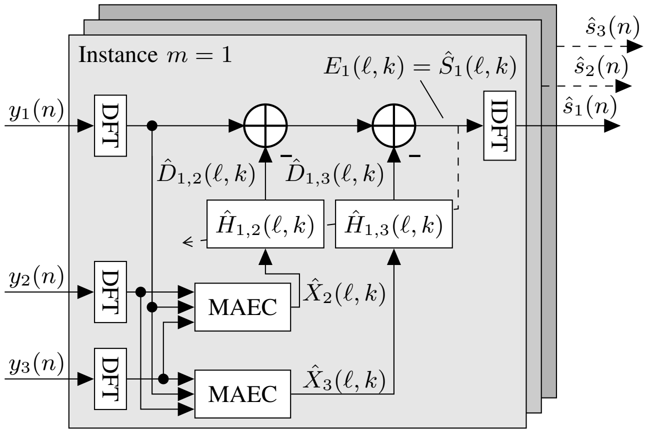

Fig. 3. **MAEC$(\mathbf{1})$/MAEC topology** for speaker $m = 1$. In order to obtain $\hat{s}_2(n)$ and $\hat{s}_3(n)$, the inputs are changed accordingly.

B. Iterative MAEC Topologies

Since the classical MAEC relies on clean reference signals without any distortion (here: interferers), it is obvious that a reduction of the crosstalk level in the reference signals, which contain speech portions of the target and the other interfering speakers, will lead to improved results. Therefore, we apply two iterative cascading schemes:

- **MAEC$(\mathbf{i})$**: All microphone channels $m \in \mathcal{M}$ are successively and iteratively improved, whereby the index $i = \{1, 2, \dots\}$ denotes the number of iterations. Thus, **MAEC$(\mathbf{1})$** is identical to the standard MAEC as depicted in Fig. 2, while **MAEC$(\mathbf{2})$** uses the enhanced output signals $\hat{s}_m(n)$ of **MAEC$(\mathbf{1})$** as input signals.
- **MAEC$(\mathbf{i})$/MAEC**: Only the *interferer* channels $\mu \in \mathcal{I}$ are iteratively improved by the already introduced **MAEC$(\mathbf{i})$** and are then used, along with the original target microphone signal $y_m(n)$ as input of the MAEC (cf. Fig. 3). Thereby, $i \in \{1, 2, \dots\}$ indicates the number of iterations regarding the improvement of the interferer channels. For $i = 1$, this structure is very close to the CTRANC scheme in [16]. However, each additional iteration of the interferer channel improvement is itself based on the original interferer signal $y_\mu(n)$ instead of $\hat{x}_\mu(n)$ in order to prevent too much loss of the speech component quality. The interferer channels $\hat{x}_\mu(n)$ as improved by **MAEC$(\mathbf{i})$** are only used as *reference inputs* for both **MAEC$(\mathbf{i+1})$** and the final MAEC as depicted in Fig. 3.

IV. Experimental Validation

The experimental validation is two-fold: First, we explain why the MAEC is able to deal with microphone leakage of the target speaker in the reference signal, and second, we evaluate the performance of the MAEC with and without cascading schemes.

A. Experimental Setup

For the investigations of the MAEC approach, we recorded a set of RIRs in a common meeting room of size 6.6 m × 5.8 m × 2.5 m (length × width × height) between the microphones of the three considered persons according to [25]. The acoustic source and sink are represented by a `Yamaha HS80M` studio monitor and a

`Beyerdynamic MM1` microphone, respectively, whereby the microphone was placed close to the mouth of a head-and-torso simulator to simulate the considered meeting scenario with close-talk headsets. A more detailed description can be found in [21, 26].

By using the measured RIRs and the NTT multilingual speech database [27], we are able to generate an artificial dialog between three persons according to Fig. 1. In this work, we focus on tripe-talk (all persons speak at the same time), since this is the most challenging case for the applied MAEC. Furthermore, we assume $h_{m,m}(n) = \delta(n)$ due to the considered close-talk scenario. Thus, the target speech signal $s_m(n) = s'_m(n)$, while the interfering signals $s'_\mu(n)$ are convolved with the corresponding RIRs $h_{m,\mu}(n)$. The target speech signals are generated by concatenating 15 NTT speech samples including some speech pauses for each speaker, which are adjusted to an active speech level (ASL) of $-26\,$dBov. We further investigate different interferer levels from $-16\,$dBov to $-46\,$dBov with a step size of $2\,$dB and add a Gaussian noise floor with a level of -75 dBov to each microphone channel $y_m(n)$ simulating some sensor noise. All applied signals are sampled with 16 kHz, level adjustments are done according to ITU-T P.56 [28], and the parameter configuration of the MAEC is in line with [23].

B. Quality Measures

All experiments are evaluated by means of a black-box signal separation method according to [29, 30], which decomposes the enhanced target signals $\hat{s}_m(n)$ into its *components*. Thus, we obtain $\tilde{s}_m(n)$, $\tilde{d}_{m,\mu}(n)$, and $\tilde{n}_m(n)$ referring to the target, interferer, and noise signals from $\hat{s}_m(n)$, respectively. Furthermore, the black-box approach carries out a delay compensation of $\hat{s}_m(n)$ w.r.t. $s_m(n)$, and is applied with a frame length of 1024 samples, a frame shift of 128 samples and a periodic Blackman window [29].

Based on these components, we determine four measures to evaluate the MAEC experiments. Due to the fact that we enhance $|\mathcal{M}| = 3$ microphone channels in the considered meeting scenario, we define the measures w.r.t. one target channel, but report in the evaluation the average over all processed microphone channels $m \in \mathcal{M}$. First, the improvement of the signal-to-interferer ratio (SIR) is defined by

$$\Delta\mathrm{SIR}_m = \mathrm{oSIR}_m - \mathrm{iSIR}_m \text{ [dB]}, \qquad (2)$$

with oSIR_m and iSIR_m being the output SIR after black-box processing and the input SIR for channel m, respectively. Both are measured according to ITU-T P.56 [28]. Second, the segmental interferer attenuation is determined by

$$\mathrm{IA}_m^{\mathrm{seg}} = \frac{1}{|\mathcal{L}^{(d_m)}|} \sum_{\ell \in \mathcal{L}^{(d_m)}} 10 \log_{10} \frac{\sum\limits_{n \in \mathcal{N}_\ell} \sum\limits_{\mu \in \mathcal{I}} d_{m,\mu}^2(n)}{\sum\limits_{n \in \mathcal{N}_\ell} \sum\limits_{\mu \in \mathcal{I}} \tilde{d}_{m,\mu}^2(n)} \text{ [dB]}, \qquad (3)$$

with $\mathcal{L}^{(d_m)}$ being the frame index set, which contains the indices of all speech-active frames of $d_m(n)$. Furthermore, samples n of frame ℓ are indicated by sample index set $\mathcal{N}_\ell$. Third, to evaluate the speech *component* quality of $\hat{s}_m(n)$, we consider a segmental speech-to-speech distortion ratio through

$$\mathrm{SSDR}_m^{\mathrm{seg}} = \frac{1}{|\mathcal{L}^{(s_m)}|} \sum_{\ell \in \mathcal{L}^{(s_m)}} \min\{\max\{\mathrm{SSDR}'_m(\ell), R_{\min}\}, R_{\max}\}, \qquad (4)$$

with $\mathcal{L}^{(s_m)}$ comprising the speech-active frame indices of $s_m(n)$, $R_{\min}$ and $R_{\max}$ being -10 dB and 30 dB, respectively, and

$$\mathrm{SSDR}'_m(\ell) = 10 \log_{10} \frac{\sum\limits_{n \in \mathcal{N}_\ell} s_m^2(n)}{\sum\limits_{n \in \mathcal{N}_\ell} \big(s_m(n) - \tilde{s}_m(n)\big)^2} \text{ [dB]}. \qquad (5)$$

Finally, we analyze speech distortion by measuring the mean opinion score $\mathrm{MOS_{LQO}}$ of the target *speech component* $\tilde{s}_m(n)$ w.r.t. $s_m(n)$ using wideband PESQ according to ITU-T P.862 [31].

C. Results and Discussion

Fig. 4 depicts the results of our different MAEC topologies regarding an iSIR of -10 to $20\,$dB for ΔSIR [dB], $\mathrm{IA^{seg}}$ [dB], $\mathrm{SSDR^{seg}}$ [dB] and $\mathrm{MOS_{LQO}}$. In accordance with most other work in this field, we assume that the target speech component is dominant against the interferer signals in the target microphone channel. Hence, the iSIR range of $0\,$dB and $20\,$dB is most relevant for the investigations of our meeting scenario. However, in order to understand and verify the usefulness of the MAEC approach in the meeting context, we consider, as already mentioned, a wider range.

For each of the four measures in Fig. 4, the common $\mathbf{MAEC(1)}$ performs best for 5 dB $<$ iSIR $<$ 20 dB, but even for iSIR $= -10$ dB, $\mathbf{MAEC(1)}$ can still enhance the target microphone signal with an adequate speech quality. This is interesting, since the interfering signals $d_{m,\mu}(n)$ are in this case dominant in the target's channel $y_m(n)$. The target's speech signal is in turn dominant in it's own reference channels, i.e., the interferers' microphone signals $y_\mu(n)$. Hence, it could be expected that the target speech of the target microphone channels would be eliminated by the MAEC. Furthermore, by comparing $\mathbf{MAEC(1)}$ with an oracle MAEC, which processes the same target microphone channels, but with crosstalk-free reference channels, two things can be observed: First, ΔSIR and $\mathrm{IA^{seg}}$ are substantially higher for the oracle MAEC (up to $7\,$dB and $6\,$dB, respectively). Second, the speech component quality of the enhanced target signal is quite comparable (except for very low iSIR conditions) in terms of $\mathrm{SSDR^{seg}}$, and even equal w.r.t. $\mathrm{MOS_{LQO}}$. Since an increase of the iSIR leads to a decrease of the performance gap between the oracle MAEC and $\mathbf{MAEC(1)}$ w.r.t. all four measures, it is obvious that the crosstalk level in the reference channels has a strong influence on the performance. But since the $\mathbf{MAEC(1)}$ indicates no significant loss of the target's speech component over the whole iSIR range, and moreover, is still working for negative iSIR values, there has to be a second cause besides the interference level, why the MAEC does not dramatically degrade the target speech components in the target microphone channels.

In line with our work in [21] and to comprehend this behavior, we have to analyze the influence of a RIR on the interferer source signals. Therefore, we consider a single-talk scenario with an adjusted ASL of $-26\,$dBov for each speaker and convolve the speaker signals $s'_m(n)$ with different fundamental components of a RIR, before coupling as interferer signals into the non dedicated microphone channels $y_\mu(n)$. For this purpose, we first break down a RIR to the fundamental characteristics with an impulse $\alpha \cdot \delta(n - n_0)$:

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

With the aid of these impulses, any discrete RIR can be modeled by superposition and concatenation including reverberation, which corresponds to a sequence $h(n)$, with $n \in \mathbb{N}_0$ and $h(n) \in \mathbb{R}$. We thus define five RIRs $h_{m,\mu}(n)$ as acoustic path from the mouth

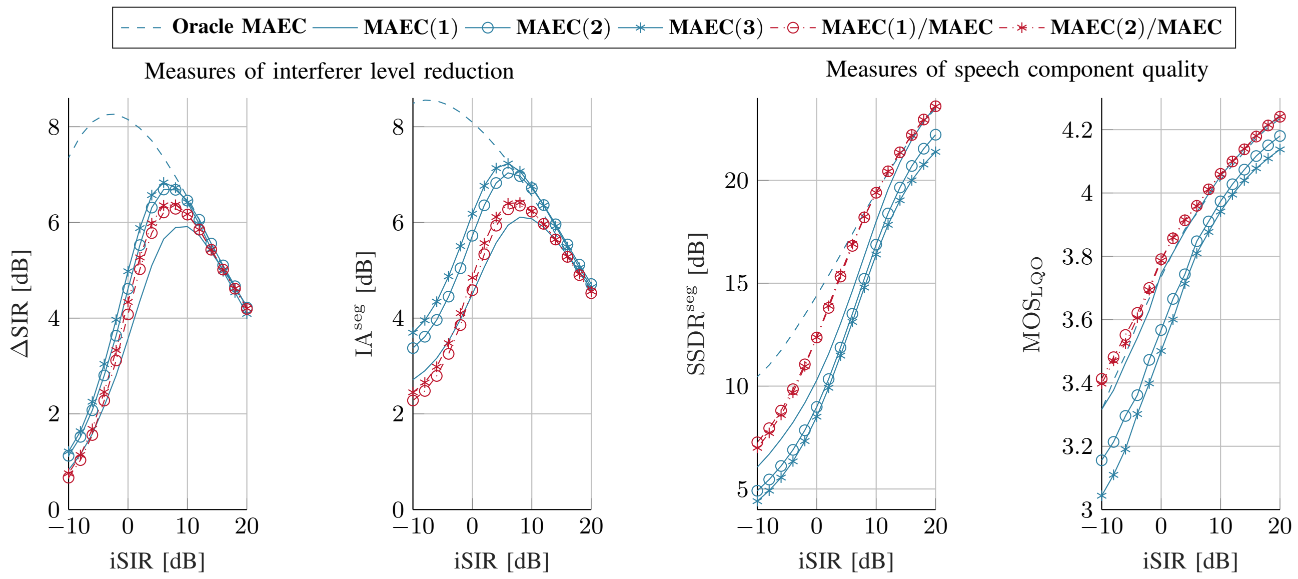

Fig. 4. **Performance evaluation of the various MAEC topologies**, x-axes denoting the iSIR [dB], and y-axes showing from left to right $\Delta\mathrm{SIR}$ [dB], $\mathrm{IA}^{\mathrm{seg}}$ [dB], $\mathrm{SSDR}^{\mathrm{seg}}$ [dB] and $\mathrm{MOS}_{\mathrm{LQO}}$. Circle and asterisk markers indicate overall two and three iterations, respectively.

of interferer speaker μ to the microphone of target speaker m: $h_{m,\mu}(n) = \delta(n)$ (no RIR), an attenuating and amplifying RIR of -5 and $5\,$dB, respectively, a delaying RIR with $n_0 = 100$ samples, and a reverberating RIR with $T_{60} = 5$ ms. The latter is truncated after 6.25 ms (i.e., after 100 samples) and simulated with a random sequence, which is shaped by an exponentially decreasing function. Moreover, we still assume $h_{m,m}(n) = \delta(n)$.

Tab. I shows the MAEC results for a single-talk scenario with the five different RIRs $h_{m,\mu}(n)$ w.r.t. the four objective measures (c.f. Sec. IV-B). By means of the $\mathrm{SSDR}^{\mathrm{seg}}$ and $\mathrm{MOS}_{\mathrm{LQO}}$ measures, it can be seen that the target speech signal is attacked and almost eliminated by the MAEC if no RIR or only an amplification or attenuation is applied to the interfering signals. Thus, we can exclude that these RIR characteristics are the reason why the MAEC can be applied to our meeting scenario without eliminating the target speech signal. Nevertheless, we already know from the results of Fig. 4, that the interferer level has an effect on the MAEC performance during multi-talk, which corresponds to the results in Tab. I, where the attenuating RIR achieves a better result than both no RIR and the amplifying RIR. However, the main reason why the MAEC works seems to be *the delay*, which is evident in Tab. I by the obtained positive $\Delta\mathrm{SIR}$ result and the achieved high speech quality of $10.77\,$dB and $3.94\,$MOS points for $\mathrm{SSDR}^{\mathrm{seg}}$ and $\mathrm{MOS}_{\mathrm{LQO}}$, respectively. The obtained performance is also comparable to the MAEC result in Fig. 4 for $\mathrm{iSIR} = 0$ dB. This is due to the fact that the interfering (crosstalk) signals, which are emanated from target speaker m, are delayed in the reference channels for target microphone channel m. Thus, the MAEC would have to estimate a RIR with negative delay to eliminate the target speech portions in the target microphone channel m, which is physically not possible. As a result, the MAEC interprets the target speech as *near-end speech* (as mentioned in Sec. III-A) and therefore leaves it untouched. For that reason, *the MAEC can be successfully applied to microphone leakage reduction in multichannel close-talk recordings*, which is confirmed by the results in Fig. 4. Furthermore, it is consistent that the MAEC obtains also adequate results for the RIR inducing reverberation, since it consists of a combination of delay and amplification/attenuation.

TABLE I
MAEC performance results IN CASE OF A SINGLE-TALK SCENARIO WITH $|\mathcal{M}| = 3$ SPEAKERS WITH VARIOUS RIRS.

RIR	ΔSIR [dB]	$\mathrm{IA}^{\mathrm{seg}}$ [dB]	$\mathrm{SSDR}^{\mathrm{seg}}$ [dB]	$\mathrm{MOS}_{\mathrm{LQO}}$
No RIR	-0.73	28.03	0.26	1.69
Amplification	-19.90	27.49	0.08	1.61
Attenuation	15.25	27.43	1.86	2.12
Delay	5.82	21.05	10.77	3.94
Reverberation	5.90	21.04	5.13	3.73

This insight leads to the conclusion that the dedication of the microphones to the sound sources is based on the shortest distance between source and microphone, instead of the highest signal energy level of a source in the microphone channel (even though it has some influence on the overall performance). Thus, both (wireless) headsets and lapel microphones, which depict the typical equipment for close-talk recordings, should work robustly in a meeting scenario, in which the participants can move freely (e.g., to use the flip chart), w.r.t. a crosstalk reduction based on the MAEC. In addition, the MAEC should also operate well in typical multichannel audio (live) performances and recordings of multiple instruments, since it is common to apply close-talk in this field. In comparison to state-of-the-art methods such as the multichannel Wiener filter approach of Kokkinis et al. [13], we showed in [20, 21] that the MAEC can achieve quite similar results w.r.t. the crosstalk reduction and the remaining target speech component quality over a wide iSIR range in a meeting scenario.

We now focus on the comparison of the various MAEC topologies in Fig. 4. $\mathbf{MAEC(1)}$ delivers in the iSIR-range of 0 to 20 dB suitable $\Delta\mathrm{SIR}$ results of about 5 dB on average and for $\mathrm{iSIR} \geq 15$ dB even equal to the performance of the oracle MAEC regarding all reported measures. This is consistent, since the reference channels of the $\mathbf{MAEC(1)}$ and oracle MAEC converge with an increasing iSIR. The performance of the interferer reduction can be further improved by the iterative $\mathbf{MAEC(i > 1)}$ scheme, which is illustrated in Fig. 4 with a blue solid line. Two and three iterations of the MAEC are shown with circle and asterisk markers, respectively. While the interferer reduction is getting better with a higher number of iterations, the speech component quality of the enhanced target signal is getting poorer with each iteration. This is because the target microphone channel of $\mathbf{MAEC(i>1)}$ is already enhanced and has thus already a little bit of speech distortion, which cannot be reversed. In contrast to this, $\mathbf{MAEC(i)/MAEC}$, which is plotted in dashed-dotted red lines, improves both interference reduction and speech quality of all our applied measures over the whole relevant iSIR range by up to $0.5\,\mathrm{dB}$ and $0.1\,\mathrm{MOS}$ points, respectively. Moreover, a higher number of iterations has a slightly positive effect on the interferer reduction, while maintaining a comparable speech component quality.

V. Conclusions

In this work, we used an approach for multichannel acoustic echo cancellation (MAEC) to reduce the speaker interference in multichannel close-talk audio recordings of a meeting. Thereby, we improved the performance of the MAEC w.r.t. both interference reduction and speech component quality by using an iterative cascading scheme. Moreover, we investigated why the MAEC is able to eliminate crosstalk in a meeting scenario, in which the reference channels are disturbed by the target speech. Thereby, we showed that the delay of the interfering signals, which is induced by the room impulse responses, is the reason why the MAEC can be applied to a microphone leakage reduction in meetings.

References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an Emerging Domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] D. Gatica-Perez, "Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[3] A. Vinciarelli, M. Pantic, D. Heylen, D. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, Jan-Mar 2012.

[4] 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, and C. Wooters, "Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations," in *Proc. of ICASSP*, Hong Kong, China, April 2003, pp. 740–743.

[5] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proc. of ICASSP*, Hong Kong, China, April 2003, pp. 364–367.

[6] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic Analysis of Multimodal Group Actions in Meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, 2005.

[7] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," in *Proc. of ASRU*, Madonna di Campiglio, Italy, Dec. 2001, pp. 107–110.

[8] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multichannel Audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, Jan. 2005.

[9] T. Prätzlich, R. M. Bittner, A. Liutkus, and M. Müller, "Kernel Additive Modeling for Interference Reduction in Multi-Channel Music Recordings," in *Proc. of ICASSP*, Brisbane, Australia, April 2015, pp. 584–588.

[10] E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Proc. of Interspeech (EUROSPEECH)*, Aalborg, Denmark, Sept. 2001, pp. 1359–1362.

[11] Ö. Cetin and E. Shriberg, "Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. 357–360.

[12] A. Clifford and J. Reiss, "Microphone Interference Reduction in Live Sound," in *Proc. of Int. Conf. on Digital Audio Effects (DAFx-11)*, Paris, France, Sept. 2011, pp. 2–9.

[13] E. K. Kokkinis, J. D. Reiss, and J. Mourjopoulos, "A Wiener Filter Approach to Microphone Leakage Reduction in Close-Microphone Applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 20, no. 3, pp. 767–779, March 2012.

[14] A. Lombard and W. Kellermann, "Multichannel Cross-Talk Cancellation in a Call-Center Scenario Using Frequency-Domain Adaptive Filtering," in *Proc. of IWAENC*, Seattle, WA, USA, Sept. 2008, pp. 14–17.

[15] T. Matheja, M. Buck, and T. Fingscheidt, "A Dynamic Multi-Channel Speech Enhancement System for Distributed Microphones in a Car Environment," *EURASIP Journal on Advances in Signal Processing*, vol. 2013 (191), Dec. 2013.

[16] G. Mirchandani, R. L. Zinser, and J. B. Evans, "A New Adaptive Noise Cancellation Scheme in the Presence of Crosstalk (Speech Signals)," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 10, pp. 681–694, Oct. 1992.

[17] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative Signal Factorization With Learnt Instrument Models for Sound Source Separation in Close-Microphone Recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 184, pp. 1–16, Dec. 2013.

[18] S. Amari and A. Cichocki, "Adaptive Blind Signal Processing-Neural Network Approaches," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.

[19] H. Buchner and W. Kellermann, "A Fundamental Relation Between Blind and Supervised Adaptive Filtering Illustrated for Blind Source Separation and Acoustic Echo Cancellation," in *Proc. of HSCMA*, Trento, Italy, May 2008, pp. 17–20.

[20] P. Meyer, S. Elshamy, and T. Fingscheidt, "A Multichannel Kalman-Based Wiener Filter Approach for Speaker Interference Reduction in Meetings," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 451–455.

[21] ——, "Multichannel Speaker Interference Reduction Using Frequency Domain Adaptive Filtering," *submitted to EURASIP Journal on Audio, Speech and Music Processing*.

[22] S. Malik and G. Enzner, "Recursive Bayesian Control of Multichannel Acoustic Echo Cancellation," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 619–622, Nov. 2011.

[23] M.-A. Jung, S. Elshamy, and T. Fingscheidt, "An Automotive Wideband Stereo Acoustic Echo Canceler Using Frequency-Domain Adaptive Filtering," in *Proc. of EUSIPCO*, Lisbon, Portugal, Sept. 2014, pp. 1452–1456.

[24] S. Malik and G. Enzner, "Online Maximum-Likelihood Learning of Time-Varying Dynamical Models in Block-Frequency-Domain," in *Proc. of ICASSP*, Dallas, TX, USA, March 2010, pp. 3822–3825.

[25] S. Müller and P. Massarani, "Transfer-Function Measurement with Sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, Jun. 2001.

[26] P. Meyer, R. Jongebloed, and T. Fingscheidt, "Multichannel Speaker Activity Detection for Meetings," in *Proc. of ICASSP*, Calgary, AB, Canada, April 2018, pp. 5539–5543.

[27] NTT, *Multi-Lingual Speech Database for Telephonometry*, NTT Advanced Technology Corporation, 1994. [Online]. Available: http://www.ntt-at.com/product/speech/

[28] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[29] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in *Proc. of INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 818–821.

[30] ITU, *Rec. P.1110: Wideband Hands-Free Communication in Motor Vehicles*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Jan. 2015.

[31] ——, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ)*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Feb. 2001.