2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3315-3745-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/WASPAA66052.2025.11230992

Microphone Occlusion Mitigation for Own-Voice Enhancement in Head-Worn Microphone Arrays Using Switching-Adaptive Beamforming

Wiebke Middelberg[1,2*], *Jung-Suk Lee*[1], *Saeed Bagheri Sereshki*[1], *Ali Aroudi*[1], *Vladimir Tourbabin*[1], *Daniel D. E. Wong*[1]

[1]Meta Reality Labs, Redmond, WA, USA [2]Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

***Abstract*—Enhancing the user's own-voice for head-worn microphone arrays is an important task in noisy environments to allow for easier speech communication and user-device interaction. However, a rarely addressed challenge is the change of the microphones' transfer functions when one or more of the microphones gets occluded by skin, clothes or hair. The underlying problem for beamforming-based speech enhancement is the (potentially rapidly) changing transfer functions of both the own-voice and the noise component that have to be accounted for to achieve optimal performance. In this paper, we address the problem of an occluded microphone in a head-worn microphone array. We investigate three alternative mitigation approaches by means of (i) conventional adaptive beamforming, (ii) switching between a-priori estimates of the beamformer coefficients for the occluded and unoccluded state, and (iii) a hybrid approach using a switching-adaptive beamformer. In an evaluation with real-world recordings and simulated occlusion, we demonstrate the advantages of the different approaches in terms of noise reduction, own-voice distortion and robustness against voice activity detection errors.**

1. INTRODUCTION

Head-worn microphone arrays like hearing aids, most modern headphones, and smart or virtual reality glasses can be effective at capturing the user's own-voice due to the vicinity to the source. This is beneficial for speech communication, e.g., for telephony or human-machine interaction using voice commands [1]–[3]. Despite the fact that even in noisy environments the user's own-voice might already be captured at a relatively high signal-to-noise ratio (SNR), noise reduction algorithms, such as beamformers steering towards the user's mouth, can further help to improve speech intelligibility [1]–[6].

A common problem for body-worn microphone arrays is susceptibility to user movement, i.e., inducing changes in the microphones' relative transfer functions due to deformation of the array [7] or quick movements relative to the acoustic scene [8]. Especially fixed spatial filters, typically relying on known microphone array geometries and microphone transfer functions, can experience a degradation of performance due to changes in the array characteristics [9]–[12], e.g., due to deformation. A re-calibration procedure for changing transfer functions used in a generalized sidelobe canceler was proposed in [13]. Another problem of body- or head-worn microphone arrays, also causing potentially rapid changes in the transfer functions, is the risk of certain microphones being occluded by skin, hair or clothing, often resulting in a muffled, i.e., low-pass filtered, sound, which can affect the performance of speech enhancement system. The problem of occluded microphones for array processing was considered in [14] where a deformable array with partial occlusion was addressed. In the context of dereverberation, the problem of rapidly changing transfer functions/filter vectors was considered in [15], where a switching version of the adaptive weighted prediction error algorithms was proposed. Switching beamformers with adaptive noise covariance matrices per filter were proposed in [16] to address the problem of interferer reduction in underdetermined situations. Examples for dictionary-based approaches for rapid dynamics are e.g. wind noise suppression [17] or automatic speech recognition with varying source directions [18].

*This work was done during an internship at Meta Reality Labs.

In this paper, we investigate the problem of own-voice enhancement for a head-worn microphone array, where a particular microphone is prone to being sporadically and dynamically occluded. As we regard the detection of the occlusion as a separate problem [19], [20], we base our work on the assumption that a reliable occlusion detector is available. We propose the following processing strategies to mitigate the effect of the occlusion: (i) a standard implementation of an adaptive beamformer, (ii) a switching mechanism between a-priori estimates of the filter vectors for the occluded and unoccluded state, i.e., similar to a dictionary-based approach, and (iii) a hybrid switching-adaptive beamformer which adapts two sets of covariance matrices depending on the occlusion state.

The proposed algorithms are evaluated using real-world recordings of own-voice in noise with dynamically changing occlusions. The evaluation is performed on multiple speech and noise samples with different dynamics in the occlusion pattern and multiple SNRs. We demonstrate the advantages of the different processing strategies in terms of noise reduction, own-voice distortion and robustness against a non-optimal voice activity detection (VAD). The results show the potential of the proposed switching-adaptive beamformer, which exhibits lower own-voice distortions for highly dynamic changes in the occlusion state than a conventional adaptive beamformer, while clearly outperforming the purely switching beamformer in terms of SNR improvement if a good VAD is available.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We consider a microphone configuration of a head-worn microphone array with M microphones, which capture the user's own-voice (considered as the target signal) and farfield noise. The noisy microphone signal in the m-th microphone in the discrete Fourier transform (DFT) domain can be written as

$$Y_m(k,t) = X_m(k,t) + N_m(k,t), \quad m \in \{1,...,M\}, \tag{1}$$

where k and t denote the frequency bin and time frame index respectively, and $X_m(k,t)$ and $N_m(k,t)$ are the speech and farfield noise captured by the m-th microphone, respectively. As all frequency bins and time frames are assumed to be independent and are processed as such, we will neglect k and t in the remainder of the paper wherever possible. The signal model in (1) can be written in terms of the M-dimensional signal vector $\mathbf{y} = [Y_1, Y_2, ..., Y_M]^T$ containing all microphone signals, where $\{\cdot\}^T$ denotes the transpose operator, i.e.,

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \tag{2}$$

where the speech and noise vector $\mathbf{x}$ and $\mathbf{n}$ are defined similarly to $\mathbf{y}$. For the speech component, a multiplicative transfer function is assumed [21], allowing to write the speech vector as

$$\mathbf{x} = \mathbf{h} X_{\mathrm{ref}}, \tag{3}$$

where X_{ref} denotes the speech component in the reference microphone, and the RTF vector $\mathbf{h}$ contains the ratios of acoustic transfer functions (ATFs) to all microphones $(A_1, ..., A_M)$ relative

to a reference microphone, such that the entry of the reference microphone is 1 by definition, i.e.,

$$\mathbf{h} = [A_1/A_{\text{ref}}, A_2/A_{\text{ref}}, ..., 1, ..., A_M/A_{\text{ref}}]^T. \tag{4}$$

Assuming statistical independence of the speech and noise component, the noisy covariance matrix can be written as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_x + \mathbf{R}_n, \tag{5}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, $\{\cdot\}^H$ denotes the Hermitian transpose operator, and $\mathbf{R}_x$ and $\mathbf{R}_n$ are the speech and noise covariance matrix, respectively. Using (3), $\mathbf{R}_x$ can be written as a rank-1 matrix spanned by the RTF vector i.e.,

$$\mathbf{R}_x = \phi_x \mathbf{h}\mathbf{h}^H, \tag{6}$$

which is scaled by the speech power spectral density in the reference microphone $\phi_x = \mathcal{E}\{|X_{\text{ref}}|^2\}$. For the noise component, we assume the covariance matrix $\mathbf{R}_n$ to be full-rank.

To include the potential occlusion of a certain microphone into the signal model, we define the first microphone to be the potentially occluded one, without loss of generality. Furthermore, we define the two ATFs $A_1 = A_{\text{ø}}$ for the unoccluded default state in (4) and A_{o} for the occluded state (and similar for the unoccluded and occluded RTFs $H_{\text{ø}}$ and H_{o}), respectively. The relation between these two states of the first microphone can be defined as

$$X_{1,\text{o}} = B_{\text{o}} X_{1,\text{ø}}, \quad N_{1,\text{o}} = G_{\text{o}} N_{1,\text{ø}}, \tag{7}$$

where $X_{1,\text{o}}$, $X_{1,\text{ø}}$, $N_{1,\text{o}}$ and $N_{1,\text{ø}}$ are the occluded and unoccluded speech and noise component in the first microphone, respectively. B_{o} and G_{o} are the occlusion transfer functions, i.e., the RTFs between the occluded and unoccluded state of the first microphone, where the occlusion transfer function for the speech component can be written as $B_{\text{o}} = A_{\text{o}}/A_{\text{ø}} = H_{\text{o}}/H_{\text{ø}}$, for which it should be noted that the definition in terms of the RTFs H_{o} and $H_{\text{ø}}$ only holds if the first microphone is not selected as the reference microphone. Also note that even though we cannot write the noise component $\mathbf{n}$ in terms of ATFs (or an RTF vector), we can still define the transfer function G_{o} between the unoccluded and occluded state of the first microphone. Using (7), the occluded RTF vector $\mathbf{h}_{\text{o}}$ can be written as

$$\mathbf{h}_{\text{o}} = \mathbf{B}\mathbf{h}_{\text{ø}}, \tag{8}$$

where $\mathbf{h}_{\text{ø}}$ is the unoccluded RTF vector and the transformation matrix $\mathbf{B}$ is defined as

$$\mathbf{B} = \text{diag}([B_{\text{o}}, \mathbf{1}_{M-1}^T]), \tag{9}$$

where $\text{diag}(\cdot)$ creates a diagonal matrix out of a vector and $\mathbf{1}_{M-1}$ is the $(M-1)$-dimensional vector of ones. Using (9), the occluded speech covariance matrix is given by

$$\mathbf{R}_{x,\text{o}} = \mathbf{B}\mathbf{R}_{x,\text{ø}}\mathbf{B}^H = \phi_x \mathbf{B}\mathbf{h}_{\text{ø}}\mathbf{h}_{\text{ø}}^H\mathbf{B}^H. \tag{10}$$

The transformation matrix $\mathbf{G}$ and the occluded noise covariance matrix $\mathbf{R}_{n,\text{o}}$ are defined similarly to (9) and (10). Generally, the two occlusion transfer functions B_{o} and G_{o} for the speech and noise component do not have to be the same. Figure 1 depicts the power of the occluded speech and noise transfer function respectively for our specific case[1], clearly showing different transfer characteristics

[1]The RTFs B_{o} and G_{o} between the occluded and unoccluded state were obtained from recordings of clean speech and noise on multiple users (data from 16 users contributed to the extraction of the transfer functions). For the speech component, the occluded RTF was obtained from the occluded and unoccluded RTF vector, while for the noise component, only a relative gain was extracted from the occluded and unoccluded noise covariance matrix.

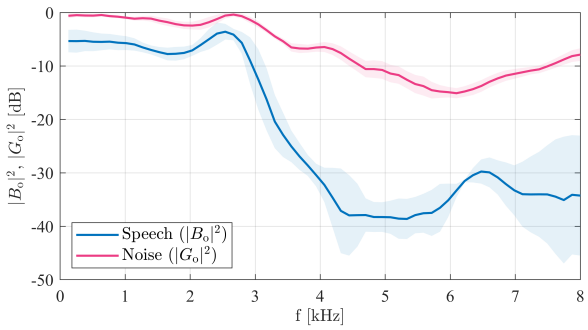

Fig. 1: Nearfield (own-voice) and farfield (noise) transfer function between occluded and unoccluded state averaged over multiple users and sound fields (solid lines) and their standard deviations (shaded areas).

for the speech and the noise component. Such differences in the effect of occlusion on different signal components can, for example, be caused by the directionality of the occluded microphone.

The output signal of the multi-channel filter is defined as $Z = \mathbf{w}^H\mathbf{y}$ with the M-dimensional filter vector $\mathbf{w}$ for which we will use the minimum-variance distortionless response (MVDR) beamformer [5], [6], [9], i.e.,

$$\mathbf{w}_\nu = \frac{\hat{\mathbf{R}}_{n,\nu}^{-1}\hat{\mathbf{h}}_\nu}{\hat{\mathbf{h}}_\nu^H\hat{\mathbf{R}}_{n,\nu}^{-1}\hat{\mathbf{h}}_\nu}, \quad \nu \in \{\text{ø},\text{o}\}, \tag{11}$$

where a hat denotes an estimate of a quantity and the beamformer depends on the occlusion state ν.

3. OCCLUSION MITIGATING PROCESSING

In this section, we discuss possible approaches on how to mitigate the influence of the occlusion on an MVDR beamformer. In Section 3.1, we present the state-of-the-art processing for dynamic acoustic scenarios, i.e., adaptive covariance matrix and RTF vector estimation. In Section 3.2, we introduce a switching mechanism for the occluded and unoccluded state, making use of a-priori known transfer functions. In Section 3.3, we propose a combination that adapts different sets of covariance matrices for the occluded and unoccluded state, respectively.

3.1. Adaptive Beamforming

In many speech enhancement algorithms for dynamic scenes, the noisy and noise covariance matrix are estimated by means of recursive smoothing [22]–[25], i.e.,

$$\hat{\mathbf{R}}_y(t) = \begin{cases} \alpha_y\hat{\mathbf{R}}_y(t-1) + (1-\alpha_y)\mathbf{y}(t)\mathbf{y}(t)^H, & \text{if VAD}(t) = 1 \\ \hat{\mathbf{R}}_y(t-1), & \text{if VAD}(t) = 0, \end{cases} \tag{12}$$

$$\hat{\mathbf{R}}_n(t) = \begin{cases} \alpha_n\hat{\mathbf{R}}_n(t-1) + (1-\alpha_n)\mathbf{y}(t)\mathbf{y}(t)^H, & \text{if VAD}(t) = 0 \\ \hat{\mathbf{R}}_n(t-1), & \text{if VAD}(t) = 1, \end{cases} \tag{13}$$

with the smoothing constants α_y and α_n, where the updates depend on the (binary) voice activity detection (VAD).

For the adaptive estimation of the RTF vector $\hat{\mathbf{h}}$, a power method implementation [26] of the generalized eigenvalue decomposition (GEVD)-based RTF estimation [27]–[30] is employed. The RTF vector is estimated as the (normalized and de-whitened) principal eigenvector of the matrix pencil $(\hat{\mathbf{R}}_n, \hat{\mathbf{R}}_y)$.

Assuming that a fast adaptation is sufficient to capture the potentially highly time-varying dynamics of occlusion, the adaptive beamforming techniques described above should be able to account for occlusion effects after a short adaptation period. However, it should also be noted that only the noisy or the noise covariance matrix can be updated at a

time, meaning that if occlusion occurs e.g. while speech is active, the noise covariance matrix cannot adapt to the occlusion. This further implies that frequent and rapid changes might not fully be captured and hence the beamformer might be suboptimal in a sense that the estimated covariance matrices do not reflect the true signal characteristics. Furthermore, note that adaptive processing as described above relies on a (good) VAD, while not depending on an occlusion detection.

3.2. Switching Beamforming

Another approach to handling occlusion makes use of the a-priori knowledge of both occluded and unoccluded RTF vectors and occluded and unoccluded noise covariance matrices, which are switched, as indicated in (11), depending on the detected occlusion state.

Hence, we assume that the occluded and unoccluded relative transfer functions, i.e., an a-priori estimate, denoted by a tilde, of the (unoccluded) RTF vector $\tilde{\mathbf{h}}_{\text{ø}}$ (cf. (3)) and the occluded transfer functions for speech and noise $\tilde{B}_{\text{o}}$ and $\tilde{G}_{\text{o}}$ in (7), respectively, are available. Furthermore, the unoccluded noise covariance matrix is modeled as diffuse as the a-priori estimate $\tilde{\mathbf{R}}_{n,\text{ø}}$. Using the transformation in (10) allows for computing the occluded RTF vector and noise covariance matrix. These assumptions are realistic to make for a known, fixed array where the user's own-voice is the signal of interest, which comes from an well known position and where prior measurements of transfer functions can be performed. Effectively, the used beamformer $\mathbf{w}$ is either of the two fixed filter vectors $\tilde{\mathbf{w}}_{\text{ø}}$ or $\tilde{\mathbf{w}}_{\text{o}}$ (cf. (11), computed based on the a-priori estimates), depending on the binary occlusion detection (OD), i.e.,

$$\mathbf{w}(t) = \begin{cases} \tilde{\mathbf{w}}_{\text{ø}}, & \text{if OD}(t) = 0 \\ \tilde{\mathbf{w}}_{\text{o}}, & \text{if OD}(t) = 1. \end{cases} \tag{14}$$

As this processing entirely relies on a-priori knowledge and does not adapt to the acoustic scenario (only to occlusion), the used estimates might not fit the data ideally, and might hence not lead to optimal performance. This means that e.g. user variability, model mismatches or changes in the acoustic scene cannot be accounted for. However, this processing also comes with a certain robustness against VAD errors, as it does not rely on a VAD.

3.3. Hybrid Switching-Adaptive Beamforming

To overcome the limitations of the two approaches in Section 3.1 and Section 3.2, we propose a combination, depending on both the VAD and the OD. Instead of adapting the covariance matrices independent of the occlusion state as in (12) and (13), we propose to adapt two different sets of covariance matrices depending on the occlusion state. The update rule in (12) for the noisy covariance matrix can hence be formulated as

$$\hat{\mathbf{R}}_{y,\nu}(t) = \begin{cases} \alpha_y \hat{\mathbf{R}}_{y,\nu}(t_\nu) + (1-\alpha_y)\mathbf{y}(t)\mathbf{y}(t)^H, & \text{if VAD}(t) = 1 \\ \hat{\mathbf{R}}_{y,\nu}(t_\nu), & \text{if VAD}(t) = 0, \end{cases} \tag{15}$$

where ν corresponds to the currently detected occlusion state and t_ν is the frame where the respective occlusion state was detected last. A similar update rule can be formulated for the noise covariance matrix $\hat{\mathbf{R}}_{n,\nu}(t)$ as in (13). The processing is summarized in Algorithm 1.

It should be noted that a practical occlusion detection and the second set of covariance matrices increase the computational complexity and memory consumption compared to the adaptive beamformer which is a caveat for resource-constrained on-device applications.

Algorithm 1 Switching-adaptive covariance estimation

Inputs:
$\tilde{\mathbf{h}}_{\text{ø}}$, $\tilde{\mathbf{R}}_{n,\text{ø}}$, $\tilde{\mathbf{B}}$, $\tilde{\mathbf{G}}$
Initialize:
$t_{\text{ø}} \leftarrow 0$, $t_{\text{o}} \leftarrow 0$
$\hat{\mathbf{R}}_{y,\text{ø}}(0) \leftarrow \tilde{\mathbf{h}}_{\text{ø}}\tilde{\mathbf{h}}_{\text{ø}}^H$, $\hat{\mathbf{R}}_{y,\text{o}}(0) \leftarrow \tilde{\mathbf{B}}\hat{\mathbf{R}}_{y,\text{ø}}(0)\tilde{\mathbf{B}}^H$
$\hat{\mathbf{R}}_{n,\text{ø}}(0) \leftarrow \tilde{\mathbf{R}}_{n,\text{ø}}$, $\hat{\mathbf{R}}_{n,\text{o}}(0) \leftarrow \tilde{\mathbf{G}}\hat{\mathbf{R}}_{n,\text{ø}}(0)\tilde{\mathbf{G}}^H$
for $t = 1$ to T **do**
$\nu \leftarrow \begin{cases} \text{ø}, \text{ if OD}(t) = 0 \\ \text{o}, \text{ if OD}(t) = 1 \end{cases}$
$\hat{\mathbf{R}}_{y,\nu}(t) \leftarrow \begin{cases} \alpha_y \hat{\mathbf{R}}_{y,\nu}(t_\nu) + (1-\alpha_y)\mathbf{y}(t)\mathbf{y}(t)^H, & \text{if VAD}(t) = 1 \\ \hat{\mathbf{R}}_{y,\nu}(t_\nu), & \text{if VAD}(t) = 0 \end{cases}$
$\hat{\mathbf{R}}_{n,\nu}(t) \leftarrow \begin{cases} \alpha_y \hat{\mathbf{R}}_{n,\nu}(t_\nu) + (1-\alpha_y)\mathbf{y}(t)\mathbf{y}(t)^H, & \text{if VAD}(t) = 0 \\ \hat{\mathbf{R}}_{n,\nu}(t_\nu), & \text{if VAD}(t) = 1 \end{cases}$
$\hat{\mathbf{h}}_\nu \leftarrow \mathcal{P}\{\hat{\mathbf{R}}_{n,\nu}(t)^{-1}\hat{\mathbf{R}}_{y,\nu}(t)\}$ ▷ GEVD-based RTF est.
$\mathbf{w}(t) \leftarrow \frac{\hat{\mathbf{R}}_{n,\nu}(t)^{-1}\hat{\mathbf{h}}_\nu}{\hat{\mathbf{h}}_\nu^H \hat{\mathbf{R}}_{n,\nu}(t)^{-1}\hat{\mathbf{h}}_\nu}$
$t_\nu \leftarrow t$
end for

4. EVALUATION

4.1. Simulation Framework and Conditions

The evaluation of the proposed algorithms was conducted using real-world recordings of speech and diffuse-like noise, recorded separately in an acoustically treated room at a sampling rate of 16 kHz. A total of six noisy signals were used for the evaluation, consisting of three different speech signals and two different noise signals, each with a duration of about 13 s. A five-channel head-mounted array in the form factor of glasses was used, with one microphone on the nose pad affected by occlusion. Speech and noise were mixed at input SNRs of 0, 5, and 10 dB in the reference microphones close to the user's ears. To simulate authentic occlusion patterns, the occlusion transfer functions (see Fig. 1) were imposed on the unoccluded microphone signals before mixing. Random occlusion patterns with varying numbers of switches (2, 8, 24, and 48) per utterance were generated to investigate the effect of different dynamics of occlusion.

The processing was performed in the DFT domain using a weighted overlap add framework with an effective frame length of 16 ms and an overlap of 75%, analysis and synthesis used custom windows to reduce leakage effects. The smoothing constants α_y and α_n for the noisy and noise covariance matrix corresponded to forgetting times of 0.3 s and 0.5 s, respectively. An oracle occlusion detection (as the detection of occlusion is beyond the scope of this paper) and VAD were used, with the latter also tested with 5% artificially induced false negatives to evaluate robustness against VAD errors.

The performance of the algorithms was evaluated in terms of binaural SNR improvement and own-voice distortion (OVD) relative to the reference channels. For the OVD, we employ the negative scale-invariant signal-to-distortion ratio [31], i.e.,

$$\text{OVD} = -20\log_{10}\left(\frac{||cx_{\text{ref}}||_2}{||cx_{\text{ref}} - x_{\text{out}}||_2}\right), \tag{16}$$

with $c = x_{\text{out}}^T x_{\text{ref}} / ||x_{\text{ref}}||_2^2$ and the time domain sequences x_{ref} and x_{out} as the reference input signal and the filtered output signal. Both objective measures are computed on the time domain signals during speech activity, and are averaged over the left and right side.

4.2. Results

The evaluation results are shown in Fig. 2, where the panels on the left depict the SNR improvement (where higher is better), and the

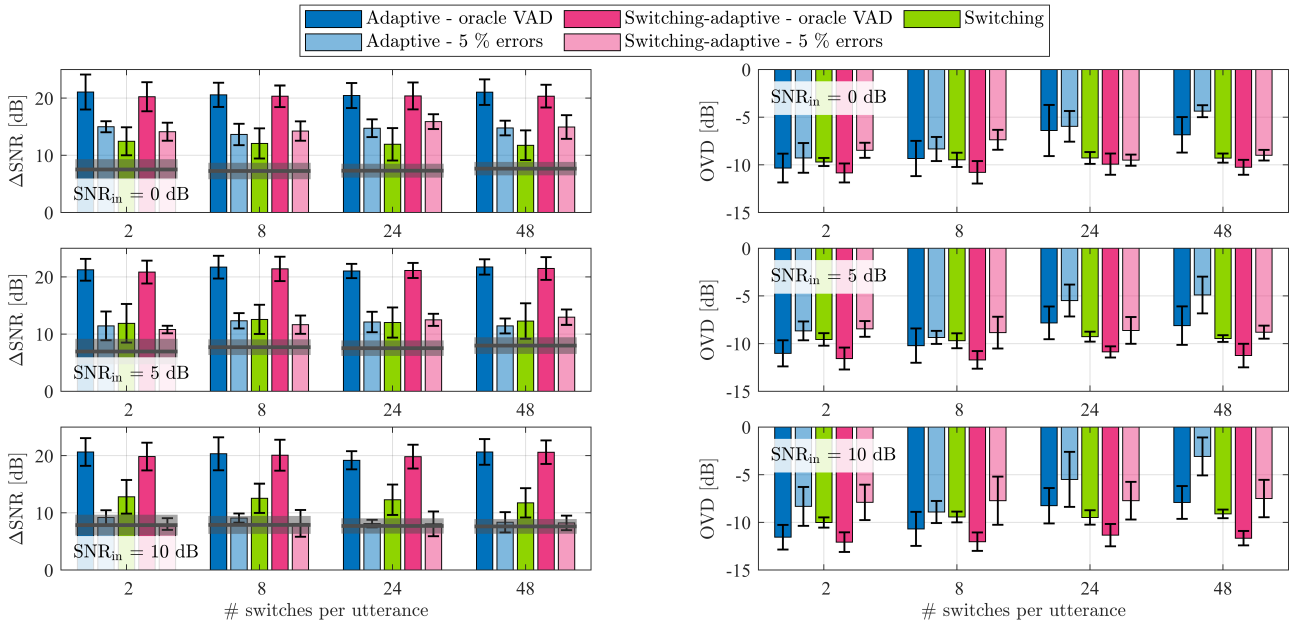

Fig. 2: Results for evaluation with two different VADs (oracle and with 5% false negatives) for different input SNRs and numbers of switches in the occlusion state per utterance. Left panels: SNR improvement with occluded microphone as reference line. Right panels: own-voice distortions.

panels on the right depict the OVDs (where lower is better). The performance is plotted over the numbers of switches in the occlusion state per utterance. The different bars show the mean results over all utterances for the different algorithms (adaptive, switching and switching-adaptive), where the solid colors represent the results for an oracle VAD, while the lighter colors represent the VAD with 5% false negatives. As the switching processing does not depend on the VAD, it is only shown for one case. The error bars represent the standard deviation over all utterances. For the SNR improvement, the gray line and shaded area depict the mean SNR improvement of the nose pad microphone relative to the reference microphone and its standard deviation, since it has the highest input SNR.

The results for the SNR improvement on the left side of Fig. 2 show that all beamformers yield an improvement compared to the nose pad microphone, where the adaptive and switching-adaptive beamformer perform similarly with an improvement of more than 10 dB compared to the best microphone for the oracle VAD, while the purely switching beamformer only yields an improvement of about 4-5 dB. In terms of SNR improvement, the performance of all algorithms is rather constant over the number of switches in the occlusion state and input SNR. For a VAD with 5% false negatives, the performance of the adaptive and switching-adaptive beamformer clearly decrease compared to the oracle VAD. While at a low input SNR of 0 dB the two adaptive beamformers still outperform the purely switching beamformer, at high SNRs the performance drop due to an erroneous VAD is more severe and the performance of the adaptive and switching-adaptive beamformer decreases to the input SNR in the best microphone. As already discussed in Section 3, the results show the robustness of the purely switching beamformer against VAD errors, while also indicating the potential of adaptation when a good VAD is available.

In terms of own-voice distortions, it can be observed that the purely switching beamformer yields rather constant low distortions (around -10 dB) for all input SNRs and numbers of switches. The switching-adaptive beamformer with an oracle VAD (solid pink) constantly leads to even slightly lower distortions. The adaptive beamformer (blue) performs similarly to the switching-adaptive beamformer in terms of OVD for a low number of switches, while inducing more distortions if the occlusion state switches often (24 and 48 switches per utterance). Overall, the distortions for the two adaptive beamformers are slightly lower at a higher input SNR. This result can be interpreted such that the purely adaptive beamformer is not capable of tracking the switches in the occlusion state, which is particularly noticeable if many switches occur at a high rate. There it seems beneficial to adapt two different sets of covariance matrices for the respective occlusion state to account for the fast dynamics in the occlusion pattern.

For an erroneous VAD, the observation for the OVD is similar to the SNR improvement: Overall, the performance decreases for the two adaptive beamformers, i.e., larger distortions are induced, where again this effect becomes more pronounced for high input SNRs. The above described trend of the switching-adaptive beamformer leading to lower distortions than the purely adaptive beamformer for highly dynamic occlusion patterns can also be observed here. The results for the OVDs further underline the robustness of the switching beamformer against VAD errors and the potential benefit of the adaptive beamformers if a good VAD is available.

5. CONCLUSIONS

In this paper, we investigated the influence of microphone occlusion and proposed different methods based on adaptive and switching beamformers, and a hybrid switching-adaptive beamformer. The proposed methods depend on an occlusion detection and/or a voice activity detection. The evaluation showed the advantages in terms of robustness for the purely switching beamformer, while also showing the potential benefits in terms of SNR improvement and own-voice distortions of adaptive beamforming if a good VAD is available. The advantage of the hybrid switching-adaptive beamformer could be demonstrated for fast dynamics in the occlusion pattern where less distortions were observed than for the adaptive beamformer, while performing similarly for relatively static occlusion patterns. Since the switching-adaptive beamformer comes at the cost of higher computational complexity and memory consumption, in practical use cases, this trade-off should further be considered along with the occurring dynamics in the occlusion pattern.

REFERENCES

[1] D. Y. Levin, E. A. Habets, and S. Gannot, "Near-field signal acquisition for smartglasses using two acoustic vector-sensors," *Speech Communication*, vol. 83, pp. 42–53, 2016.

[2] P. Hoang, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022.

[3] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Modeling of speech-dependent own voice transfer characteristics for hearables with an in-ear microphone," *Acta Acustica*, vol. 8, p. 28, 2024.

[4] J. Benesty, M. M. Sondhi, Y. Huang *et al.*, *Springer handbook of speech processing*. Springer, 2008, vol. 1.

[5] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, Apr. 2017.

[7] R. M. Corey and A. C. Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, 2019, pp. 115–119.

[8] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "Nice-beam: Neural integrated covariance estimators for time-varying beamformers," *arXiv preprint arXiv:2112.04613*, 2021.

[9] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[10] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Dec. 2010, pp. 416–420.

[11] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, Nov. 2023.

[12] A. Mannanova, K. Tesch, J.-M. Lemercier, and T. Gerkmann, "Meta-learning for variable array configurations in end-to-end few-shot multichannel speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark, 2024, pp. 200–204.

[13] P. Oak and W. Kellermann, "A calibration algorithm for robust generalized sidelobe cancelling beamformers," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Eindhoven, Netherlands: Citeseer, 2005, pp. 97–100.

[14] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Variational bayesian multi-channel robust NMF for human-voice enhancement with a deformable and partially-occluded microphone array," in *Proc. European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: IEEE, 2016, pp. 1018–1022.

[15] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Processing Letters*, vol. 28, pp. 1580–1584, 2021.

[16] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, Apr. 2019, pp. 7908–7912.

[17] M. Tammen, X. Li, S. Doclo, and L. Theverapperuma, "Dictionary-based fusion of contact and acoustic microphones for wind noise reduction," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, 2022, pp. 1–5.

[18] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 681–685.

[19] N. Madhu and R. Martin, "Low-complexity, robust algorithm for sensor anomaly detection and self-calibration of microphone arrays," *IET signal processing*, vol. 5, no. 1, pp. 97–103, 2011.

[20] M. D. Gaal, A. R. Berkovich, and E. D. Prins, "Signal Processing for Microphone Blockage Detection," U.S. Patent Application US20190014429A1, Jan. 2019.

[21] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

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

[23] N. Gößling and S. Doclo, "RTF-Steered Binaural MVDR Beamforming Incorporating an External Microphone for Dynamic Acoustic Scenarios," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 416–420.

[24] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, 2016.

[25] J. Donley, V. Tourbabin, B. Rafaely, and R. Mehra, "Adaptive multi-channel signal enhancement based on multi-source contribution estimation," in *Proc. European Signal Processing Conference (EUSIPCO)*. Dublin, Ireland: IEEE, 2021, pp. 276–280.

[26] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[27] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[28] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, Apr. 2011.

[29] R. Serizel, M. Moonen, B. van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[30] M. X. Cohen, "A tutorial on generalized eigendecomposition for denoising, contrast enhancement, and dimension reduction in multichannel electrophysiology," *Neuroimage*, vol. 247, p. 118809, 2022.

[31] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 626–630.