APPROXIMATE MVDR AND MMSE BEAMFORMERS EXPLOITING SCALE-INVARIANT RECONSTRUCTION OF SIGNALS ON MICROPHONES

Zbyněk Koldovský[1] *and Francesco Nesta*[2]

[1]Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec,
Studentská 2, 461 17 Liberec, Czech Republic
[2]Conexant System, 1901 Main Street, Irvine, CA (USA)

ABSTRACT

Minimum Variance Distortionless Response (MVDR) and Minimum Mean-Squared Error (MMSE) beamformers are popular array processors for enhancing multichannel recordings of a directional source. We propose their approximate variants having the generalized sidelobe canceler structure whose performances depend purely on the blocking matrix part. No auxiliary methods such as adaptive interference canceler or voice activity detector to estimate the source/noise covariance are needed. Instead, scale-invariant least square estimators are used, which enable to estimate the noise also during speech activity and to recover original spectra of the target source on the microphones. In experiments, we compare signal-to-noise ratio improvements of several variants of the beamformers achieved on six-channel recordings of speech in nonstationary noisy conditions.

Index Terms— Minimum Variance Distortionless Beamformer, Minimum Mean Square Error Beamformer, Blind Source Separation, Scaling Ambiguity

1. INTRODUCTION

Consider a noisy recording of a directional source observed through m microphones. In the frequency domain, the model reads

$$\mathbf{x}(k) = \mathbf{a}(k)s(k) + \mathbf{y}(k), \tag{1}$$

where $\mathbf{x}(k)$ is the $m \times 1$ vector of the signals on microphones, $s(k)$ is the target signal, $\mathbf{a}(k)$ is the steering vector whose elements contain acoustic transfer functions between the source and the microphones, and $\mathbf{y}(k)$ involves all other interfering sources and noise components that are uncorrelated with $s(k)$[1]. From now on we will omit the frequency index k from the notation in order to simplify the exposition.

Popular array processors that extract s from $\mathbf{x}$, thereby reduce noise, enhance or even dereverberate the target signal, are the Minimum Variance Distrotionless Response (MVDR) and Minimum Mean-Squared Error (MMSE) beamformers [3]. They are, respectively, represented by vectors

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{a}} \quad \text{and} \quad \mathbf{w}_{\text{MMSE}} = \mathbf{C}^{-1}\mathbf{c}_{\mathbf{x}s}, \tag{2}$$

where $\mathbf{C} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$, $\mathbf{C}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$, $\mathbf{c}_{\mathbf{x}s} = \mathrm{E}[\mathbf{x}s^*]$, $\mathrm{E}[\cdot]$ stands for the expectation operator, and $\cdot^*$ and $\cdot^H$ denote the conjugate value and the conjugate transpose, respectively. The output $\hat{s} = \mathbf{w}^H\mathbf{x}$ can be either the estimate of s or a scaled variant of it, which is, typically, the response (spatial image) of s on a microphone [1] or a partly dereverberated signal on a microphone [2]. In this paper, we aim at the estimation of the spatial images of s, which does not require any additional knowledge about the original spectrum of s.

Various strategies have been proposed to estimate $\mathbf{C}_{\mathbf{y}}$ or $\mathbf{c}_{\mathbf{x}s}$ from $\mathbf{x}$, many of which try to decompose $\mathbf{C}$ as a sum of signal and noise covariances [4]. $\mathbf{C}_{\mathbf{y}}$ can be observed during noise-only intervals, for which an activity detection of s is needed. Recently, deep neural networks (DNN) trained on clean signals have been popular tools to perform in-depth frequency-dependent voice activity detection (VAD); see, e.g., [5, 6, 7]. The goal of this paper is to avoid the need for any signal activity detection.

A popular implementation of MVDR is Generalized Sidelobe Canceler (GSC) that is comprised of three blocks, called fixed beamformer (FB), blocking matrix (BM), and adaptive interference canceler (AIC) [8]. The FB block acquires the target signal under the distortionless constraint. The BM works in parallel with the FB, where it cancels s and produces noise-only reference signals. The AIC aims to minimize the power of the residual noise in the output of the FB using the reference from the BM.

To realize GSC, three crucial problems have to be solved. (P1) One has to cope with the scaling ambiguity in $\hat{s}$ when $\mathbf{a}$ is not precisely known. For example, $\mathbf{a}$ can be blindly identified up to a scaling factor, e.g., as the principal vector of an estimate of the covariance of $\mathbf{a}s$ [6]. (P2) The BM has to be

This work was supported by The Czech Science Foundation through Project No. 14-11898S and partly by California Community Foundation through Project No. DA-15-114599.

[1] In fact, $\mathbf{a}(k)$ can be defined in a different way depending on the target scaling of the enhanced signal [1]. For example, it can represent early parts of the acoustic transfer functions as considered in [2].

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

found such that the leakage of the target signal into the noise reference is as small as possible [1, 9, 10]. (P3) In dynamic situations (e.g, moving sources), the AIC, typically realized through adaptive normalized least-square algorithm, can perform its task efficiently only during noise-only periods [1].

In this paper, we propose approximate MVDR and MMSE beamformers that depend purely on the solution of (P2) and cope with the problems (P1) and (P3) by using scale-invariant least-square estimators. The scale-invariance property is important, because $\mathbf{a}$ is identified up to its scale as a vector generating the null-space of the BM. Also the output of the BM could be seen as a scaled variant of $\mathbf{y}$ where the scaling factor is unknown. Here, the AIC is realized by the scale-invariant estimate of $\mathbf{y}$ that does not require any additional assumptions or VAD (can be applied during periods of activity of the target signal).

The paper is organized as follows. Section 2 introduces the scale-invariant estimators. In Section 3, the approximate MVDR and MMSE beamformers are derived. Section 4 is devoted to an experiment in which the beamformers are compared with optimum ones (using known target signal and noise). The proposed beamformers are compared in variants where the BM is built up from known and estimated Relative Transfer Functions (RTFs). Section 5 concludes the paper.

2. SCALE-INVARIANT ESTIMATORS

The scaling ambiguity appears in Blind Source Separation (BSS) models, for example, in Independent Component Analysis (ICA) [11]. Let $\mathbf{B}$ be a $d \times m$ matrix that extracts from $\mathbf{x}$ a d-dimensional signal $\mathbf{u} = \mathbf{B}\mathbf{x}$. Due to indeterminacy of the BSS model, $\mathbf{u}$ has random scaling. A popular solution to the scaling ambiguity proposed in [12, 13] assumes that a regular de-mixing matrix that separates $\mathbf{x}$ into individual signals is available. In this respect, $\mathbf{B}$ is a submatrix that consists of the rows of the de-mixing matrix. The inverse of the de-mixing matrix is used to estimate $\mathbf{u}$ as it appears in the mixture $\mathbf{x}$; let the contribution of $\mathbf{u}$ to $\mathbf{x}$ be denoted as $\mathbf{u}_\mathbf{x}$.

In fact, the approach [12, 13] can be applied when $\mathbf{x}$ is a determined mixture of signals, which is true only if $\mathbf{x} = \mathbf{A}\mathbf{v}$ where $\mathbf{A}$ is a regular square matrix. The model (1) is determined only if $\mathbf{y} = \mathbf{H}\mathbf{z}$ where $\mathbf{H}$ is an $m \times (m-1)$ full-column-rank matrix, and $\mathbf{z}$ represents an $(m-1)$-dimensional signal. For example, $\mathbf{x}$ described by (1) corresponds to a regular mixture if $\mathbf{y}$ embodies $m-1$ directional sources.

A more flexible solution is to project $\mathbf{u}$ back to $\mathbf{x}$ using least squares. The projection matrix is

$$\mathbf{V} = \arg \min_{\mathbf{P} \in \mathcal{C}^{m,m}} \mathrm{E}\|\mathbf{x} - \mathbf{P}\mathbf{u}\|^2 = \mathbf{C}\mathbf{B}^H(\mathbf{B}\mathbf{C}\mathbf{B}^H)^{-1}, \quad (3)$$

and the estimate of the projected signal is

$$\widehat{\mathbf{u}}_\mathbf{x} = \mathbf{C}\mathbf{B}^H(\mathbf{B}\mathbf{C}\mathbf{B}^H)^{-1}\mathbf{B}\mathbf{u}. \quad (4)$$

This estimator has advantageous properties. First, it is scale-invariant in the sense that $\mathbf{B}$ can be substituted by $\mathbf{\Lambda}\mathbf{B}$ where $\mathbf{\Lambda}$ is a $d \times d$ regular matrix while $\widehat{\mathbf{u}}_\mathbf{x}$ will not be changed by $\mathbf{\Lambda}$. Second, the estimator is applicable even if the whole de-mixing matrix is not available or if the mixture $\mathbf{x}$ is underdetermined.

Now, we apply (4) as if $\mathbf{x}$ is determined as follows. Let $\mathbf{w}$ and $\mathbf{W}$ be, respectively, a separating vector and a matrix such that $\mathbf{w}^H\mathbf{x}$ is equal to s up to a scalar multiple while $\mathbf{W}^H\mathbf{x}$ is equal to $\mathbf{z}$ up to a multiple by an unknown regular scaling matrix. The estimates of the extracted signals by $\mathbf{w}$ and $\mathbf{W}$ as they appear on the microphones are, respectively,

$$\widehat{\mathbf{a}s} = \frac{\mathbf{C}\mathbf{w}\mathbf{w}^H}{\mathbf{w}^H\mathbf{C}\mathbf{w}}\mathbf{x}, \quad (5)$$

$$\widehat{\mathbf{y}} = \mathbf{C}\mathbf{W}(\mathbf{W}^H\mathbf{C}\mathbf{W})^{-1}\mathbf{W}^H\mathbf{x}. \quad (6)$$

The theoretical properties of the estimators are as follows. In the special case that $\mathbf{x}$ is determined, the estimators are consistent (approach the true signals when the length of data grows). In the fully underdetermined case, which happens e.g. whenever $\mathbf{y}$ involves a diffused source that is active on all microphones, then if $\mathbf{W}^H$ is orthogonal[2] to $\mathbf{a}$, (6) is an optimal estimate of $\mathbf{y}$ in the least-squared error sense. Note that no $\mathbf{w}$ such that separates s from $\mathbf{x}$ exists in the underdetermined scenario. Nevertheless, $\widehat{\mathbf{a}s}$ still appears to be a good approximate of $\mathbf{a}s$ if $\mathbf{w}$ extracts s with an improved signal-to-noise ratio. We therefore use (5) to cope with (P1).

3. PROPOSED BEAMFORMERS

Let $\mathbf{W}^H$ represent the BM part of GSC, that is, be orthogonal to $\mathbf{a}$, and let its rank be equal to $m-1$. We will derive approximate MVDR and MMSE beamformers using the above estimators.

3.1. Approximate MVDR

The strategy is to replace $\mathbf{y}$ and $\mathbf{C}_\mathbf{y}$ in (2), respectively, by (6) and by the covariance of $\widehat{\mathbf{y}}$, which is, according to (6), equal to

$$\mathbf{C}_{\widehat{\mathbf{y}}} = \mathrm{E}[\widehat{\mathbf{y}}\widehat{\mathbf{y}}^H] = \mathbf{C}\mathbf{W}(\mathbf{W}^H\mathbf{C}\mathbf{W})^{-1}\mathbf{W}^H\mathbf{C}. \quad (7)$$

Next, $\mathbf{a}$ can be identified, up to an unknown scale, as a nonzero vector from the null space of $\mathbf{W}^H$. Let us denote such vector as $\mathbf{b}$, which will be used to replace $\mathbf{a}$ in (2).

After the replacements, two problems have to be resolved. First, $\mathbf{C}_{\widehat{\mathbf{y}}}$ has rank $m-1$, so its inverse matrix does not exist. Second, as $\mathbf{b}$ differs from $\mathbf{a}$ by a multiple, so will differ the scale of the extracted signal from the desired one.

We propose two ways to cope with the rank deficiency of $\mathbf{C}_{\widehat{\mathbf{y}}}$. One is to replace $\mathbf{C}_{\widehat{\mathbf{y}}}^{-1}$ by the Moore-Penrose pseudoinverse denoted as $\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}$. By doing so, we arrive at an off-scaled

[2] Note that $\mathbf{W}^H$, such that is orthogonal to $\mathbf{a}$, in fact, satisfies the definition of a blocking matrix.

approximate MVDR beamformer that is represented by

$$\widehat{\mathbf{w}}_{\mathrm{MVDR},1} = \frac{\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}\mathbf{b}}{\mathbf{b}^H\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}\mathbf{b}}. \tag{8}$$

The second way resides in reducing the signal dimension, e.g., by neglecting a selected microphone signal. Generally speaking, let $\mathbf{E}$ be a $d \times m$ full-row-rank matrix where $0 < d < m$ (e.g., the upper submatrix of the $m \times m$ identity matrix). Let also $\mathbf{E}$ be such that $\mathbf{E}\mathbf{C}_{\widehat{\mathbf{y}}}\mathbf{E}^H$ is regular. Then, $\mathbf{E}\mathbf{x}$ is the reduced signal, and $\mathbf{E}\mathbf{C}_{\widehat{\mathbf{y}}}\mathbf{E}^H = \mathbf{C}_{\mathbf{E}\widehat{\mathbf{y}}}$, which is the approximate covariance matrix of $\mathbf{E}\mathbf{y}$, the noisy part of $\mathbf{E}\mathbf{x}$. Finally, $\mathbf{E}\mathbf{b}$ is equal to $\mathbf{E}\mathbf{a}$ up to the unknown scaling.

Hence, the corresponding off-scaled approximate MVDR beamformer that applies to the reduced signal $\mathbf{E}\mathbf{x}$, is defined through

$$\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}} = \frac{(\mathbf{E}\mathbf{C}_{\widehat{\mathbf{y}}}\mathbf{E}^H)^{-1}\mathbf{E}\mathbf{b}}{\mathbf{b}^H\mathbf{E}^H(\mathbf{E}\mathbf{C}_{\widehat{\mathbf{y}}}\mathbf{E}^H)^{-1}\mathbf{E}\mathbf{b}}. \tag{9}$$

The last step is to cope with the unknown scaling of $\widehat{\mathbf{w}}_{\mathrm{MVDR},1}$ and of $\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}}$ due to that of $\mathbf{b}$. We propose to apply (5), thereby re-scale the beamformers so that they estimate $\mathbf{a}s$ and $\mathbf{E}\mathbf{a}s$, respectively. To justify, note that, in the context of BSS, $\widehat{\mathbf{w}}_{\mathrm{MVDR},1}$ and $\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}}$ could be seen as vectors that aim to separate (extract) the target signal s, respectively, from $\mathbf{x}$ and from $\mathbf{E}\mathbf{x}$.

Finally, the beamformers estimate vector signals and are, respectively, represented by rank-1 transform matrices

$$\widehat{\mathbf{W}}_{\mathrm{MVDR},1} = \frac{\mathbf{C}\widehat{\mathbf{w}}_{\mathrm{MVDR},1}(\widehat{\mathbf{w}}_{\mathrm{MVDR},1})^H}{(\widehat{\mathbf{w}}_{\mathrm{MVDR},1})^H\mathbf{C}\widehat{\mathbf{w}}_{\mathrm{MVDR},1}}, \tag{10}$$

$$\widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}} = \frac{\mathbf{E}\mathbf{C}\mathbf{E}^H\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}}(\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}})^H}{(\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}})^H\mathbf{E}\mathbf{C}\mathbf{E}^H\widehat{\mathbf{w}}_{\mathrm{MVDR},\mathbf{E}}}, \tag{11}$$

where their outputs are $\widehat{\mathbf{W}}_{\mathrm{MVDR},1}\mathbf{x}$ and $\widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}}\mathbf{E}\mathbf{x}$, respectively.

3.2. Approximate MMSE

The MMSE beamformer can be implemented as the MVDR beamformer followed by the Wiener postfilter; see page 447 in [3]. Specifically, $\mathbf{w}_{\mathrm{MMSE}}$ can be written as

$$\mathbf{w}_{\mathrm{MMSE}} = \underbrace{\frac{\sigma_s^2}{\sigma_s^2 + (\mathbf{a}^H\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{a})^{-1}}}_{\text{Wiener postfilter}} \cdot \mathbf{w}_{\mathrm{MVDR}}, \tag{12}$$

where $\sigma_s^2 = \mathrm{E}[|s|^2]$. The Wiener postfilter is represented by the scalar fraction where $\Lambda = (\mathbf{a}^H\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{a})^{-1}$ is, in fact, the variance of the residual noise at the MVDR output, that is, $\Lambda = \mathrm{E}[|\mathbf{w}_{\mathrm{MVDR}}^H\mathbf{y}|^2]$. In addition, the denominator $\sigma_s^2 + \Lambda$ is equal to the variance of the whole MVDR output. By incorporating the approximate MVDRs derived above into this MMSE implementation, we derive an approximate MMSE processor that estimates $\mathbf{a}s$ as follows.

The output of (10) and of (11) can be, respectively, approximated as

$$\widehat{\mathbf{W}}_{\mathrm{MVDR},1}\mathbf{x} \approx \mathbf{a}s + \widehat{\mathbf{W}}_{\mathrm{MVDR},1}\mathbf{y}, \tag{13}$$

$$\widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}}\mathbf{x} \approx \mathbf{E}\mathbf{a}s + \widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}}\,\mathbf{E}\mathbf{y}. \tag{14}$$

The last terms of (13) and (14) correspond to the residual noise in the output of the approximate MVDRs. We propose to replace $\mathbf{y}$ by $\widehat{\mathbf{y}}$ defined through (6) to approximate these terms, i.e.,

$$\mathbf{r}_1 = \widehat{\mathbf{W}}_{\mathrm{MVDR},1}\widehat{\mathbf{y}} \approx \widehat{\mathbf{W}}_{\mathrm{MVDR},1}\mathbf{y},$$

$$\mathbf{r}_{\mathbf{E}} = \widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}}\,\mathbf{E}\widehat{\mathbf{y}} \approx \widehat{\mathbf{W}}_{\mathrm{MVDR},\mathbf{E}}\,\mathbf{E}\mathbf{y}.$$

These terms can be used to approximate the Wiener postfilter at each output of the approximate MVDRs as follows.

Let the vector outputs of the approximate MVDRs be, respectively, denoted as $\mathbf{u}_1 = \mathbf{W}_{\mathrm{MVDR},1}\mathbf{x}$ and $\mathbf{u}_{\mathbf{E}} = \mathbf{W}_{\mathrm{MVDR},\mathbf{E}}\,\mathbf{E}\mathbf{x}$. Approximate Wiener postfilters[3] for all input channels, represented by a diagonal matrix, are given through

$$\widehat{\mathbf{W}}_{\mathrm{WF,i}} = \mathrm{diag}\big[\max\{|\mathbf{u}_i|^2 - |\mathbf{r}_i|^2, \epsilon\}\big] \cdot \mathrm{diag}\big[|\mathbf{u}_i|^2 + \epsilon\big]^{-1}, \tag{15}$$

where $\mathrm{i} \in \{1, \mathbf{E}\}$, $\mathrm{diag}[\cdot]$ denotes a diagonal matrix with the argument of its diagonal, and ϵ is a small positive constant that restrains from division by zero. Finally, the transform matrices of the proposed approximate MMSE beamformer are given through

$$\widehat{\mathbf{W}}_{\mathrm{MMSE,i}} = \widehat{\mathbf{W}}_{\mathrm{WF,i}}\widehat{\mathbf{W}}_{\mathrm{MVDR,i}}, \tag{16}$$

where $\mathrm{i} \in \{1, \mathbf{E}\}$.

4. EXPERIMENTAL EVALUATION

The experiments reported here were performed in the off-line regime[4] with simulated six-channel noisy recordings. The data were taken from the third CHiME evaluation campaign [15]. A 7 s female utterance from file `F05_440C0202_BTH.CH0.wav` was convolved with room impulse responses taken from [16]. The speaker's distance corresponded to one meter from a linear array of six microphones at 0°. The reverberation time was $T_{60} = 360$ ms; the microphone spacing was 8 cm; the sampling frequency is 16 kHz.

The microphone images of the utterance were mixed with six-channel recordings of noise from CHiME. There are four different noisy locations: BUS, CAF, PED, and STR. Each experiment was repeated 100 times with randomly selected

[3] It is possible to consider other approximations of the Wiener postfilter or other single-channel post-filtering approaches; see, e.g., [14].

[4] The proposed beamformers can be easily modified for on-line processing using recursive estimators of covariance matrices. We report only the off-line testing in this paper due to limited space.

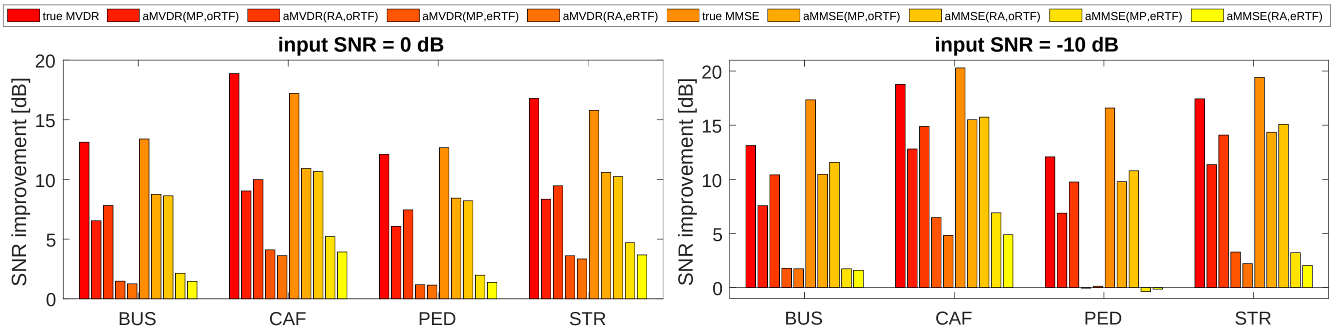

Fig. 1. SNR improvement averaged over 100 trials when input SNR is 0 and -10 dB, respectively. The length of DFT is 1024.

samples of noise. Then, the signals were processed in the short-time Discrete Fourier Transform (STFT) with the length of the DFT L, hop-size $L/2$, and the Hamming window.

Methods applied to the noisy recordings were assessed in terms of the output signal-to-noise ratio (SNR). Both variants of the proposed beamformers were compared, that is, the one using the Moore-Penrose pseudoinverse (MP) and the one based on the reduction approach (RA). In the latter case, $\mathbf{E}$ was taken as the upper $m-1\times m$ part of the $m\times m$ identity matrix ($m=6$).

The BM part of the beamformers was constructed using estimated RTFs between pairs of adjacent microphones. Let $H^i_{\rm RTF}(k)$ denote the RTF between the ith and the $(i+1)$th microphone, $i=1,\dots,5$, k is the frequency index. The blocking matrix is then defined as

$$\mathbf{W}(k)=\begin{pmatrix} H^1_{\rm RTF}(k) & 0 & \dots & 0\\ q(k) & H^2_{\rm RTF}(k) & \dots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & \dots & q(k) & H^5_{\rm RTF}(k)\\ 0 & \dots & 0 & q(k)\end{pmatrix} \tag{17}$$

where $q(k)=-\exp\{\frac{2\pi\sqrt{-1}Dk}{L}\}$ and D is the integer delay parameter introduced due to possible acausality of the RTFs; the default value was $D=20$.

The RTFs were estimated using (1) least-squares (see, e.g., [9]) and noise-free target signal images (available in simulations only) and (2) the frequency-domain estimator by [1] that is fully independent of a priori knowledge. The latter assumes that $\mathbf{y}$ is stationary, which, in fact, is not satisfied in case of the CHiME noise, so a suboptimal accuracy could be expected. The reader is referred, e.g., to [17, 18, 19, 9] for a more advanced RTF estimators based on BSS.

Fig. 1 shows results of the experiment in two scenarios where input SNR is 0 and -10 dB, respectively. Ten approaches are compared. As a reference, “true” MVDR and MMSE beamformers were computed using (2) and known source and noise signals. The SNR improvements achieved by these methods stand for optimum results, which are between 10 to 20 dBs, depending on the type of noise.

Next, the approximate MVDR and MMSE beamformers have the acronyms aMVDR(A,B) and aMMSE(A,B), respectively, where A signifies the variant MP or RA, and B signifies the RTF estimates used to form the blocking matrix (oRTF - oracle estimates, eRTF - estimates by [1]).

With the oracle RTFs, the beamformers achieve high SNR improvement in the range from 8 to 15 dBs, which points to their efficiency (although the SNR is lower than that of the “true” beamformers). For both input SNR situations, aMVDR(RE,oRTF) achieves higher SNR than aMVDR(MP, oRTF), nevertheless, the difference is maximum about 1-2 dB. The same holds for aMMSE, but the improvement is significant only when the input SNR is -10 dB; in the 0 dB case, MMSE with MP is slightly better than with RE.

While true MVDR can achieve higher SNR than true MMSE, especially, when the input SNR is 0 dB, aMVDR appears to be always outperformed by aMMSE. This follows from the aMMSE structure as it is, in fact, a post-processed aMVDR by the approximate Wiener postfilter. The postfilter thus always succeeds in improving the SNR (nevertheless, this is usually achieved at the cost of a small distortion of the target signal).

The performance with eRTF is significantly lower than with oRTF, which points to the importance of the accuracy of the RTF estimates (or blocking matrix, in general). Also, the results with eRTF are better in case of input SNR 0 dB then in the -10 dB case, obviously because the accuracy of the eRTF estimates improves with the input SNR. The SNR improvement also depends on the characteristics of noise; the best results seem to be achieved for the CAF noise.

5. CONCLUSIONS

The results confirm that the performances of the proposed beamformers depend purely on the accuracy of the target signal blocking. A way to improve the latter is thus to employ more robust and accurate RTF estimates. Using BSS estimates is the topic for our future works as well as on-line (adaptive) implementations of the beamformers.

6. REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[2] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, Feb 2015.

[3] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sept 2010.

[5] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, March 2016.

[6] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[7] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi, and Y. Hioka, "Pinpoint extraction of distant sound source based on dnn mapping from multiple beamforming outputs to prior snr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 435–439.

[8] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan 1982.

[9] Z. Koldovský, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug 2015.

[10] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sept 2004.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley, 2004.

[12] K. Matsuoka, "Minimal distortion principle for blind source separation," in *SICE 2002. Proceedings of the 41st SICE Annual Conference*, Aug 2002, vol. 4, pp. 2138–2143.

[13] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, Oct. 2001.

[14] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, 2009.

[15] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 504–511.

[16] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, Sept 2014, pp. 313–317.

[17] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, May 2009.

[18] N. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, Aug 2015, pp. 409–413.

[19] S. Meier and W. Kellermann, "Analysis of the performance and limitations of ica-based relative impulse response identification," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, Aug 2015, pp. 414–418.