Adaptive Multi-Channel Signal Enhancement Based on Multi-Source Contribution Estimation

Jacob Donley, Vladimir Tourbabin, Boaz Rafaely and Ravish Mehra
Facebook Reality Labs Research, USA

***Abstract*—Automated solutions to multi-channel signal enhancement for improving speech communication in noisy environments has become a popular goal among the research community. Many proposed approaches focus on adapting to speech signals based on their temporal characteristics but these methods are primarily limited to specific types of desired and undesired sound sources. This paper outlines a new method to adapt to desired and undesired signals using their spatial statistics, independent of their temporal characteristics. The method uses a linearly constrained minimum variance (LCMV) beamformer to estimate the relative source contribution of each source in a mixture, which is then used to weight statistical estimates of the spatial characteristics of each source used for final separation. The proposed method allows for instantaneous desired and undesired source selection, a useful ability for the enhancement of conversations. The simulated results show that the method can adapt to the targeted source in noisy mixture signals and that under realistic conditions it is also capable of reaching ideal MVDR performance.**

***Index Terms*—Adaptive beam-forming, signal enhancement, microphone array, multi-channel processing, parameter estimation**

I. Introduction

The ability to extract clear speech from noisy environments caused by interfering speakers, reverberation and/or ambient noise, using spatial processing algorithms, has been highly sought after in recent years [1]. Common use cases that benefit from extracting only the desired signals of interest include automatic speech recognition [2]–[4], hearing-aid signal enhancement [5]–[7], on-line voice chat, video conferencing [8]–[10] as well as numerous other applications, such as real-time enhancement of voices in noisy restaurants.

Many data-independent methods have been considered as robust solutions to reducing the noise in signals. For example, in the area of spatial signal enhancement, which focuses solely on leveraging signal characteristics in the spatial domain, algorithms known as beamformers, for example delay and sum, and super-directive (such as maximum directivity and differential microphone arrays), have been studied extensively [11]–[13].

Other more optimal approaches require statistical knowledge of both the desired source and the undesired noise [12], [14]. Methods such as minimum-variance distortionless-response (MVDR) and linearly-constrained minimum-variance (LCMV) provide optimal filter coefficients for separating the two types of signals whilst not distorting the desired signal. The LCMV formulation allows for a linear constraint on multiple transfer functions if they are available. Alternative formulations relying only on the inference of the desired or undesired, as well as information of the mixture of both, are known as the minimum-power distortionless-response (MPDR) and linearly-constrained minimum-power (LCMP) methods.

While the aforementioned approaches are capable of providing an optimal solution given knowledge of the signal parameters, it is still challenging to estimate those parameters from the signal alone. Several methods for spatial filter parameter estimation exist, however, they often assume speech source signals. The methods typically consider the temporal activity and probability of whether or not speech has occurred [15], [16]. These are usually classified as voice activity detectors or the popular single-channel speech presence probability (SPP) [16]. The SPP has also been investigated for use with multi-channel arrays to leverage the spatial domain to determine the probability of speech in different spatial locations [17], [18]. These methods rely on the assumption that speech is either the target signal of interest, primarily the noise signal or both. This assumption limits the ability of speech based methods, particularly when interfering sources may be music, vehicles, animals or otherwise that are not temporally or spectrally similar to speech. Other techniques use long temporal history and statistics to separate sources, which can result in long latency in real-time applications [19].

Given the limitations of relying on speech signals only, we propose a spatial signal enhancement method that is independent of the spectral and temporal statistics of the desired and undesired source signals. The proposed method performs the adaptive parameter estimation solely using spatial domain processing techniques. The proposed algorithm's independence on the type of source signal is especially valuable when spatial noise sources other than speech are present. We assume that each source contributes a portion of signal to the received samples and that the relative contribution correlates to the usefulness of the samples in estimating spatial source parameters. Under these assumptions, the method starts with initial direction of arrival (DOA) estimates [20] and tracking to determine the location of the sources. The locations are used to seed an LCMV process that is then used to determine a relative source contribution estimate (SCE) of each source to the currently received signal. The SCE value is used as a weight for adaptively learning statistical parameters of the sound sources in noisy reverberant environments, which are also used to compute the parameters of a final set of MVDR filters. The final filters are used to perform spatial signal enhancement thus enhancing a selected desired sound source. The parameters used for the final MVDR are fed back into the first LCMV and, hence, adaptively refined over time.

In section II, the signal model used throughout the rest of the paper is introduced along with a brief description of spatial filtering and parameter estimation. A description of the proposed parameter estimation method is given in section III, which goes into detail on the proposed SCE method and the associated statistical adaptation. The setup used for the simulated analysis along with results from the analysis of the proposed methods are outlined in section IV and the paper is concluded in section V.

II. Problem Description

Throughout this work, we consider the scenario where we have several, potentially concurrent, sound sources in a room with ambient noise. The challenge is that each source is mixed together in the received microphone signals. Our goal is to extract a chosen source while removing the remaining sources and ambient noise. The extraction process can be accomplished via spatial filtering. Typically, spatial filtering requires knowledge of the desired and undesired components of the mixture, these components are commonly termed the *parameters* of the model. An open challenge with spatial filtering is estimating the correct parameters for signal separation.

A. Signal Model

Assuming the received signals at the microphones are a mixture of individual spatially stationary sources with locations $\psi \equiv (\theta, \phi)$, in the time-frequency domain we have,

$$\mathbf{x}_{M\times 1}(t,f) = \sum_{n=1}^{N} s_n(t,f)\cdot \mathbf{h}_{M\times 1}(f,\psi_n) + \mathbf{u}_{M\times 1}(t,f), \quad (1)$$

$$\mathbf{x} = \mathbf{H}\cdot\mathbf{s} + \mathbf{u}, \quad (2)$$

where $s_n(t,f) \in \mathbb{C}$ is a source signal, $\mathbf{h}_{M\times 1}(f,\psi_n) \in \mathbb{C}^{M\times 1}$ is an array transfer function (ATF) for the M element microphone array in the direction of the n-th source, $\psi_n \equiv (\theta_n, \phi_n), \forall n \in \{1,\ldots,N\}$ and $\mathbf{u}_{M\times 1}(t,f) \in \mathbb{C}^{M\times 1}$ is a noise signal containing all unwanted sounds. We assume that $\mathbf{h}_{M\times 1}(f,\psi_n)$ sufficiently describes the response over the time segment of $\mathbf{x}_{M\times 1}(t,f)$. In complete vector notation, $\mathbf{H}$ is the ATF matrix of size $M \times N$, $\mathbf{s}$ is the source vector of size $N \times 1$ and $\mathbf{u}$ is the noise vector of size $M \times 1$. Throughout this work we denote discrete time with t and discrete frequency with f. For brevity, from here on we omit the function arguments of time, (temporal) frequency, angular position and subscripted size, unless they are necessary to facilitate a description. In addition, from here on, symbols noted with accents are an estimated value or have been derived from one.

B. Spatial Filtering

A popular method for separating a particular desired source signal from the noise and other interfering sources is spatial filtering, such as MVDR or LCMV. We describe the separation process in this work with an MVDR formulation, however, an LCMV process may be a reasonable alternative for some applications. The MVDR filter for extracting the signal for a desired index, n, is given by,

$$\mathbf{w}_n(t,f) = \frac{\mathbf{\Phi}_n^{-1}(t,f)\cdot\mathbf{h}_n}{\mathbf{h}_n^H\cdot\mathbf{\Phi}_n^{-1}(t,f)\cdot\mathbf{h}_n}, \quad (3)$$

where $(\cdot)^H$ denotes a Hermitian transposition of the corresponding matrix, the transfer function of the desired source is $\mathbf{h}_n$ and all the remaining undesired signals for that desired source are defined by the noise covariance matrix, $\mathbf{\Phi}_n(t,f)$. The filter from (3) is used to extract an approximation of the desired source signal frame from the received microphone signals with,

$$s_n \approx \mathbf{w}_n^H\cdot\mathbf{x}. \quad (4)$$

C. Parameter Estimation

In (3), we assume that $\mathbf{h}_n$ and $\mathbf{\Phi}_n$ are known or can be reliably estimated. These two parameters are often obtained off-line by a calibration process (transfer function measurement and/or singular value decomposition) that is provided clean samples of the desired and/or undesired sources.

The off-line approach fails in a real-time scenario as desired and undesired samples are unknown. A voice activity detector (VAD) is a popular approach for understanding when to use the correct samples, however, VAD's are not always accurate in noisy conditions and are tuned to speech signals. Another method is the multi-channel speech presence probability (M-SPP), however, this method is also reliant on the statistics of speech. The challenge of correctly and reliably estimating which components of the signal are desired and which are undesired, so that they can be filtered out, is the crux of the problem.

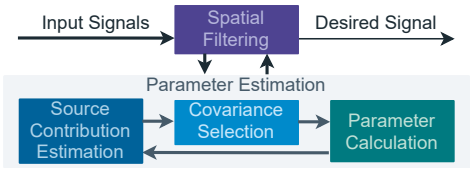

Fig. 1. High level system diagram of the proposed parameter estimation approach to spatial filtering.

III. Proposed Parameter Estimation Method

In this section, we outline the proposed approach to estimating the spatial filtering parameters in real-time. The method assumes that each source contributes a portion of signal to any given frame of samples. We also assume that the relative contribution a source makes to a frame correlates to the usefulness of that frame in estimating the parameters for that source. A high level depiction of the proposed spatial filtering method is shown in Fig. 1.

A. Initially Required Values

We first start by describing the values required for initialization of the process depicted in Fig. 1 and indicate estimated values using a caret. We begin with the assumption that we can obtain a reasonable estimate (e.g. by measurements or other simulations) of the free-field relative transfer function (RTF), $\widehat{\mathbf{h}}(f,\psi)$, of size $M \times 1$, for a set of ψ that is representative of the response in most directions [1], [21].

The initial noise covariance matrices are isotropic noise covariances obtained by using the estimated RTFs,

$$\widehat{\mathbf{\Phi}}_n(0,f) = \sum_{p\in\mathcal{P}} \widehat{\mathbf{h}}(f,p)\cdot\widehat{\mathbf{h}}^H(f,p), \quad (5)$$

where $t = 0$ and $\mathcal{P}$ is a set of points, $p \equiv (\theta,\phi)$, sampled approximately equidistant on a sphere.

We also assume that we know, or can reliably estimate, the N directions, $\{\psi_n\}_{n=1}^N$, of the sound sources that we wish to control using the adaptive procedure. For example, by using a DoA and tracking algorithm [20].

B. Multi-source Contribution Estimation

In this section we describe the proposed multi-source contribution estimation procedure. We begin with using a linearly-constrained minimum-variance (LCMV) beamformer, which is described by,

$$\widehat{\mathbf{w}}_n = \widehat{\mathbf{\Phi}}_n^{-1}\cdot\widehat{\mathbf{H}}\cdot\left(\widehat{\mathbf{H}}^H\cdot\widehat{\mathbf{\Phi}}_n^{-1}\cdot\widehat{\mathbf{H}}\right)^{-1}\cdot\mathbf{g}_n, \quad (6)$$

where $\widehat{\mathbf{w}}_n$ is of size $M \times 1$, $\widehat{\mathbf{\Phi}}_n$ is $M \times M$, $\widehat{\mathbf{H}}$ is $M \times N$ and $\mathbf{g}$ is $N \times 1$. In (6), we use a set of RTFs for each localized source,

$$\widehat{\mathbf{H}} = \left[\widehat{\mathbf{h}}(f,\psi_1) \quad \widehat{\mathbf{h}}(f,\psi_2) \quad \cdots \quad \widehat{\mathbf{h}}(f,\psi_N)\right]. \quad (7)$$

We then attempt to extract each source signal in the scene arriving from each tracked direction,

$$\widehat{s}_n = \widehat{\mathbf{w}}_n^H\cdot\mathbf{x}, \quad (8)$$

where the noise covariance, $\widehat{\mathbf{\Phi}}_n(t,f)$, is initially the isotropic covariance in (5) at $t = 0$ and a source is selected with the constraints,

$$\mathbf{g}_n = \left[0_{n-1\times 1} \quad 1_{1\times 1} \quad 0_{N-n\times 1}\right]^T. \quad (9)$$

Using the estimated DoA's, we arrive at an $N \times 1$ size vector of estimated source signal strengths,

$$\widehat{\mathbf{s}} = \left[\widehat{s}_1 \quad \widehat{s}_2 \quad \cdots \quad \widehat{s}_N\right]^T. \quad (10)$$

These source signal strengths are the basis of the multi-source contribution estimation method.

C. Contribution Equalization and Normalization

In order to ensure that the contribution estimates for all sources both, sum to unity and are valid during noise only frames, for which the LCMV may be invalid, an equalization is performed using the noise-only excited LCMV auto-correlation values. The total energy of the equalized signal is then used to normalize the contribution estimates.

The noise-only LCMV output gain, used for equalization, is obtained by finding the auto-correlation values of the LCMV from (6) initialized with the isotropic noise covariance from (5). We start by finding the noise-only LCMV energy for each source,

$$\mathbf{W}_{\mathrm{u}} = \widehat{\mathbf{\Phi}}_n^{-1}(0,f) \cdot \widehat{\mathbf{H}} \cdot \left(\widehat{\mathbf{H}}^H \cdot \widehat{\mathbf{\Phi}}_n^{-1}(0,f) \cdot \widehat{\mathbf{H}} \right)^{-1} \cdot \mathbf{I}_{N \times N} \quad (11)$$

and then the equalization values for each source contribution estimate are given by the $N \times 1$ vector,

$$\widehat{\mathbf{u}} = \sqrt{\mathrm{diag}(\mathbf{W}_{\mathrm{u}}^H \cdot \mathbf{W}_{\mathrm{u}})}. \quad (12)$$

Finally, we equalize for noise-only lower bound values and normalize to the total energy of equalized signal strengths to arrive at an estimate of the contribution, $\widehat{\mathbf{c}}$, of each source to the particular time-frequency sample,

$$\widehat{\mathbf{c}}(t,f) = \frac{\widehat{\mathbf{s}} \oslash \widehat{\mathbf{u}}}{||\widehat{\mathbf{s}} \oslash \widehat{\mathbf{u}}||_1}, \quad (13)$$

where $\oslash$ is the Hadamard division operator. In an effort to make the estimate more robust to spatial aliasing of the array and microphone self-noise, we propose an optional step of using the mean contribution over a useful range of frequencies for the particular application, as,

$$\bar{\mathbf{c}}(t) = \frac{1}{F} \sum_{f \in \mathbb{F}} \widehat{\mathbf{c}}(t,f), \quad (14)$$

where $\mathbb{F}$ is a set of size F containing all pre-defined frequencies for which the mean is taken and an overbar denotes the direct result of a mean. Otherwise, the values from (13) can be used to weight the contributions per frequency directly. The estimated contribution can then be used to weight statistics from which the parameters of the MVDR beamformer in (3) can be computed.

D. Parameter Adaptation

We propose to replicate, for the estimated number of sources, the sample covariance matrix obtained from the current frame of microphone signals and weight it by the corresponding element of $\widehat{\mathbf{c}}$ from (13),

$$\widehat{\mathbf{\Omega}}_n = \widehat{c}_n \cdot \mathbf{x} \cdot \mathbf{x}^H. \quad (15)$$

We propose to use a buffering set, $\widetilde{\mathbf{\Omega}}_{n,l}, l \in \{1,...,L\}$, of L weighted sample covariance matrices for each source and two companion sets containing the associated contribution values, $\tilde{\mathbf{c}}_{n,l}, l \in \{1,...,L\}$, and the time at which they were included in the set, $\tilde{\mathbf{t}}_{n,l}, l \in \{1,...,L\}$. We denote the buffering and companion sets using a tilde. The buffering set is initialized with covariances from (5) and the companion sets with zeros,

$$\widetilde{\mathbf{\Omega}}_{n,l} = \left\{ \widehat{\mathbf{\Phi}}_n(0,f) \right\}_{l=1}^{L}, \quad (16)$$

$$\tilde{\mathbf{c}}_{n,l} = \{0\}_{l=1}^{L}, \quad (17)$$

$$\tilde{\mathbf{t}}_{n,l} = \{0\}_{l=1}^{L}. \quad (18)$$

At each discrete time, $t \in \mathbb{Z}$, a new weighted sample covariance matrix, $\widehat{\mathbf{\Omega}}_n$, and associated contribution, $\widehat{c}_n$, are added to the sets such that they replace the element of the buffering set with the lowest corresponding $\widehat{c}_n$ if their contribution value is greater,

$$\widetilde{\mathbf{\Omega}}_{n,\arg\min_l(\tilde{\mathbf{c}}_{n,l})} = \widehat{\mathbf{\Omega}}_n, \quad \text{s.t. } \widehat{c}_n > \min_l(\tilde{\mathbf{c}}_{n,l}), \quad (19)$$

$$\tilde{\mathbf{c}}_{n,\arg\min_l(\tilde{\mathbf{c}}_{n,l})} = \widehat{c}_n, \quad \text{s.t. } \widehat{c}_n > \min_l(\tilde{\mathbf{c}}_{n,l}), \quad (20)$$

$$\tilde{\mathbf{t}}_{n,\arg\min_l(\tilde{\mathbf{c}}_{n,l})} = t, \quad \text{s.t. } \widehat{c}_n > \min_l(\tilde{\mathbf{c}}_{n,l}). \quad (21)$$

In order to ensure that the learned statistics remain relevant over extended periods of time, a forgetting process is implemented that naturally forces the statistics back to those that are under more generalized assumptions, e.g. a specular noise field based on the provided DoA of the noise sources or a diffuse noise field. The forgetting procedure is defined as,

$$\widetilde{\mathbf{\Omega}}_{n,\arg\min_l(\tilde{\mathbf{t}}_{n,l})} = \widehat{\mathbf{h}}(f,\psi_n) \cdot \widehat{\mathbf{h}}(f,\psi_n)^H, \quad \text{s.t. } t - \mathcal{T} > \min_l(\tilde{\mathbf{t}}_{n,l}), \quad (22)$$

$$\tilde{\mathbf{c}}_{n,\arg\min_l(\tilde{\mathbf{t}}_{n,l})} = 1/N, \quad \text{s.t. } t - \mathcal{T} > \min_l(\tilde{\mathbf{t}}_{n,l}), \quad (23)$$

$$\tilde{\mathbf{t}}_{n,\arg\min_l(\tilde{\mathbf{t}}_{n,l})} = t, \quad \text{s.t. } t - \mathcal{T} > \min_l(\tilde{\mathbf{t}}_{n,l}), \quad (24)$$

where $\mathcal{T}$ is a time-based forgetting threshold after which statistics are forgotten. The procedure that is (19), (20), (21), (22), (23) and (24), is performed as an update at every new time frame.

We then estimate the covariance matrix by taking the average as,

$$\overline{\mathbf{\Omega}}_n = \frac{1}{L} \sum_{l=1}^{L} \widetilde{\mathbf{\Omega}}_{n,l}. \quad (25)$$

The statistical adaptation process is applied for all frequency components unless (14) is used. The adaptation procedure involves updating the filter coefficients described in (6) by using updated parameters.

E. Estimating Tracked Steering Vectors

We estimate a new reverberant RTF for the desired source, $\widehat{\mathbf{h}}_n$, at index n, based on the estimated signal covariance from (25). We do this using the method of Eigen-value decomposition (EVD),

$$\overline{\mathbf{\Omega}}_n \widehat{\mathbf{h}}_n = \lambda \widehat{\mathbf{h}}_n, \quad (26)$$

where the principal Eigen-vector, $\widehat{\mathbf{h}}_n$, satisfies the above equation with the largest Eigen-value, λ. This estimated reverberant RTF, $\widehat{\mathbf{h}}_n$, can be used in the MVDR formulation in (3) replacing $\mathbf{h}_n$.

F. Estimating Untracked-Source Noise Statistics

Under some circumstances, the DoA estimation and source tracking stage might not track sources. In this work, we learn the parameters of the untracked sources as noise, which could be diffuse, volumetric, quiet or from other sources not identified by the DoA or tracking.

We propose to perform exponential smoothing on sample covariance matrices when the maximum of all SCE values is below a given threshold. The untracked-source noise covariance is given by,

$$\widehat{\mathbf{\Phi}}_{\mathrm{u}}(t,f) = \alpha \cdot \mathbf{x} \cdot \mathbf{x}^H + (1-\alpha) \cdot \widehat{\mathbf{\Phi}}_{\mathrm{u}}(t-1,f), \qquad t > 0 \quad (27)$$

where the forgetting (smoothing) factor is,

$$\alpha = \begin{cases} \alpha_{\mathrm{forget}}, & \text{if } \max(\widehat{\mathbf{c}}) < \beta_{\mathrm{thr}} \\ 0, & \text{otherwise} \end{cases}, \quad (28)$$

α_{forget} is the conditional forgetting factor and β_{thr} is the threshold for classifying untracked-sources. $\widehat{\mathbf{\Phi}}_{\mathrm{u}}$ is initialized at $t = 0$ equivalent to (5). We then use the updated covariances, $\widehat{\mathbf{\Phi}}_{\mathrm{u}}$, in (6) to find new filters that correspond to updated SCEs.

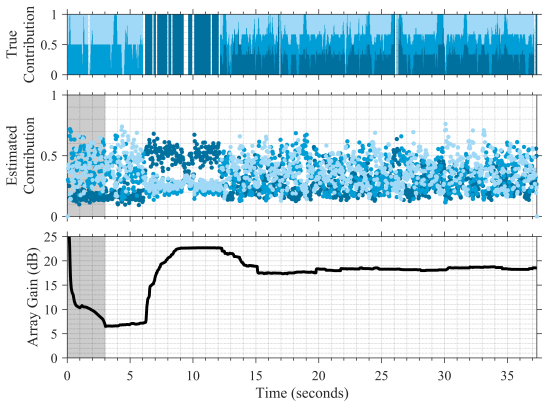

Fig. 2. The true contribution, estimated contribution and beamformer array gain are shown from top to bottom, respectively. The three colors correspond to three speech sources. The top plot shows a stacked bar graph. The gray area indicates algorithm ramp-up from initialization ($L = 94, 3072$ ms).

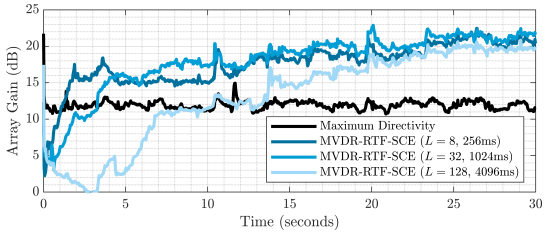

Fig. 3. The true signal array gain in decibels is shown for four different beamformers to highlight their adaptation speed. All sources are continually active for the full duration.

G. Estimating Total Undesired Source Statistics

When using MVDR or LCMV, suppressing tracked and untracked sources requires statistical inference of their spatial relationship. We propose combining the statistics evenly. After N frames, we iteratively update the total undesired source covariance to be,

$$\widehat{\mathbf{\Phi}}_n = \mathbf{\Phi}_n = \widehat{\mathbf{\Phi}}_{\mathrm{u}} + \frac{1}{N-1} \sum_{n' \in \{1,\ldots,N\} \setminus \{n\}} \overline{\mathbf{\Omega}}_{n'}. \qquad (29)$$

The estimated undesired covariance can be used in (3) to produce the filtered output and also in (6) when re-estimating the SCEs. The process described in section III is then repeated indefinitely.

IV. Results and Discussion

In this section, we outline the simulated experimental setup and provide details for reproducing results. The performance results in terms of array gain (signal-to-noise ratio improvement), as defined in [12], are discussed with respect to oracle MVDR performance trained with known source activity.

A. Experimental Setup

Recordings of speech in rooms with various reverberation times were simulated. The size of the rooms were $6\,\mathrm{m} \times 7\,\mathrm{m} \times 3\,\mathrm{m}$ and the wall absorption coefficients were adjusted using the Eyring formula [22] so that the reverberation time to 60 dB (RT60) varied between $0.15\,\mathrm{s}$ to $0.9\,\mathrm{s}$ with a total of 6 different RT60 values. The signals were simulated using a speed of sound in air of $343\,\mathrm{m\,s^{-1}}$ at a sampling frequency of $16\,\mathrm{kHz}$. A circular microphone array with 6 equally spaced microphones was centered at $(2, 3.5, 1.5)$. White Gaussian sensor noise was added to all received signals at a level of $30\,\mathrm{dB\,SPL}$. Three speech sources, $N = 3$, were positioned randomly around the microphone array with a distance of $1\,\mathrm{m}$ to $2\,\mathrm{m}$, an angular separation greater than 20° in azimuth and elevation within $\pm 10^\circ$. The performance was analyzed from 20 random sets of positions.

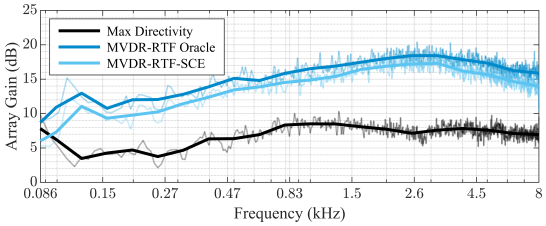

Fig. 4. The octave-band -mean frequency-domain array gain is shown for an RT60 of 750 ms. The result is an average over 20 sets of simulations where each includes three speech sources.

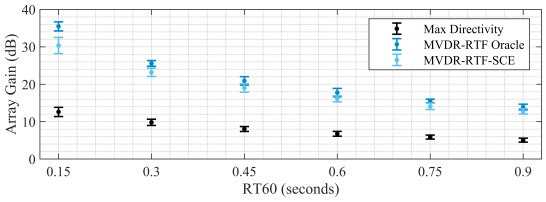

Fig. 5. The mean array gain is shown for RT60 ranging from 150 ms to 900 ms. 95% confidence intervals over the full bandwidth and 20 sets of source positions are shown.

The received signals were then processed using all steps of the proposed algorithm, named here-on as 'MVDR-RTF-SCE', along with a maximum directivity index beamformer, 'Max Directivity', and an oracle MVDR trained on known source activity, 'MVDR-RTF Oracle'. Simulated RTFs were used for all methods. A weighted over-lap add (WOLA) block-based process was used with a block size of 1024, an overlap of 50 % and a square-root hanning window for analysis and synthesis. The lengths of the buffers were $L = 94$, corresponding to approximately 3 s, and the frequency range in (14) was 100 Hz to 5 kHz. The time-based forgetting threshold was $\mathcal{T} = \infty$ in order to investigate adaptation behavior and the forgetting factor was $\alpha_{\mathrm{forget}} = 10^{-2}$. The threshold for classifying untracked-sources was $\beta_{\mathrm{thr}} = 1.2/N$. The mixture signals contained 6 s of desired-only and 6 s of undesired-only segments of sound to begin with.

B. Adaptation Behavior

We show the behavior of the proposed algorithm in Fig. 2 for a specific simulation (RT60 of 150 ms). The top plot shows the oracle activity as determined using a voice activity detector on the ideal clean source signals. The middle plot shows the SCEs per time and the bottom most plot shows the estimated array gain of the adapted filter. When there are clean segments of speech the spatial adaptation returns SCEs that correlate well with the true activity. The algorithm maintains performance in the mixed source signal segments, as measured by the array gain. The specific scenario provides a clear opportunity for the algorithm to determine the correct parameters, thus allowing it to achieve a close match to an ideal MVDR.

In more difficult scenarios, like the one from the results in Fig. 3, mixed sources results in few segments where only one source is active, thus limiting the ability for algorithms to learn correct parameters. The proposed algorithm adaptively improves performance in the more difficult conditions as shown by the increasing array gain. The array gain increases as rapidly as in Fig. 2 and matches performance in less than 2 s when $L = 8$. It is important to note that in Fig. 3 the values converge on similar performances after 30 s regardless of L. The array gain also surpasses the 'Maximum Directivity' beamformer in approximately 1 s.

C. Array Gain Performance

Array gain performance is analyzed in Fig. 4. A similar method to the one in Fig. 2 is used and repeated for the 20 sets of random positions. Adaptation is paused after $12\,\mathrm{s}$ to ensure the analysis is reflective of the maximum performance. The array gain of the proposed approach is slightly lower than oracle performance but significantly above the 'Maximum Directivity' case. The proposed method's peak performance reaches $17\,\mathrm{dB}$ on average, $1\,\mathrm{dB}$ less than the $18\,\mathrm{dB}$ of the 'Oracle' method. The proposed method's mean performance remains $1\,\mathrm{dB}$ to $2\,\mathrm{dB}$ less than the 'Oracle' case for the wideband speech range and consistently results in $4\,\mathrm{dB}$ to $9\,\mathrm{dB}$ greater array gain than the 'Maximum Directivity' method.

Further, performance as a function of RT60, shown in Fig. 5, shows an inverse correlation of array gain and reverberation, which is expected as spatial diffuseness increases with reverberation. The proposed method's maximum performance is similar to the 'Oracle' MVDR for all reverberation levels as it is exposed to source signals partly separated in time. The proposed method achieves more than $7\,\mathrm{dB}$ higher array gain than the equivalent 'Max Directivity' beamformer (RT60 of $900\,\mathrm{ms}$).

V. Conclusions

In this work, a method is proposed to automatically estimate statistical parameters for a spatial filtering process. The method estimates the spatial energy contribution of tracked sources within an environment and uses the contributions to bias a statistical representation of the environment learned over time. The method is analyzed in terms of adaptation behavior, adaptation speed and converged performance in terms of array gain. It is shown that the method correctly adapts given general initialization conditions and in noisy multi-talk scenarios. The adaptation can match the performance of an ideal spatial filter and when using a small buffer size can produce array gain that outperforms a maximum directivity beamformer in under two seconds. The resulting spatial filters have distortionless constraints and the method is not limited to speech sources.

References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[2] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *2016 Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Shanghai: IEEE, Mar. 2016, pp. 5745–5749.

[3] T. Ochiai *et al.*, "Speaker adaptation for multichannel end-to-end speech recognition," in *2018 Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 6707–6711.

[4] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 1–5.

[5] N. Gossling and S. Doclo, "RTF-steered binaural MVDR beamforming incorporating an external microphone for dynamic acoustic scenarios," in *2019 Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brighton, United Kingdom: IEEE, May 2019, pp. 416–420.

[6] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[7] D. Marquardt and S. Doclo, "Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1261–1274, Jul. 2018, ISSN: 2329-9290, 2329-9304.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv:1804.04121*, Jun. 19, 2018.

[9] K. Tan *et al.*, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *arXiv:1909.07352*, Apr. 10, 2020. arXiv: 1909.07352. (visited on 04/14/2020).

[10] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug. 10, 2018, ISSN: 0730-0301, 1557-7368.

[11] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[12] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[13] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and applications of spherical microphone array processing*. Springer, 2017, vol. 9.

[14] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[15] A. Ivry, I. Cohen, and B. Berdugo, "Evaluation of deep-learning-based voice activity detectors and room impulse response models in reverberant environments," in *2020 Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 406–410.

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[17] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Letters*, vol. 11, no. 12, pp. 956–959, 2004.

[18] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, 2009.

[19] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Trans. on Signal Process.*, vol. 66, no. 24, pp. 6458–6473, 2018.

[20] V. Tourbabin, J. Donley, B. Rafaely, and R. Mehra, "Direction of arrival estimation in highly reverberant environments using soft time-frequency mask," in *Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, IEEE, Oct. 2019, pp. 383–387.

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

[22] C. F. Eyring, "Reverberation time in "dead" rooms," *J. Acoust. Soc. Am.*, vol. 1, no. 2A, pp. 217–241, 1930.