Offline Noise Reduction Using Optimal Mass Transport Induced Covariance Interpolation

Filip Elvander*, Randall Ali†, Andreas Jakobsson*, and Toon van Waterschoot†
*Div. of Mathematical Statistics, Lund University, Sweden,
†Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Belgium

***Abstract*—In this work, we propose to utilize a recently developed covariance matrix interpolation technique in order to improve noise reduction in multi-microphone setups in the presence of a moving, localized noise source. Based on the concept of optimal mass transport, the proposed method induces matrix interpolants implying smooth spatial displacement of the noise source, allowing for physically reasonable reconstructions of the noise source trajectory. As this trajectory is constructed as to connect two observed, or estimated, covariance matrices, the proposed method is suggested for offline applications. The performance of the proposed method is demonstrated using simulations of a speech enhancement scenario.**

***Index Terms*—Noise reduction, speech enhancement, optimal mass transport, covariance interpolation**

I. Introduction

The problem of multi-microphone noise reduction finds many applications in both online and offline audio processing, such as in speech enhancement for hearing aids and for human-machine interaction [1], [2]. In multi-microphone setups, a commonly utilized technique is applying the minimum variance distortion less response (MVDR) beamformer [3] to the signal measured by the microphone array [4] in order to allow for suppressing unwanted interference while at the same time not distorting the desired signal component, such as a speech signal. Seen as a spatial filter, the MVDR beamformer requires a reliable estimate of the noise-only spatial covariance matrix in order to efficiently suppress any power associated with the noise source. Commonly, the noise covariance matrix is estimated using exponentially smoothed outer products of the measured microphone signals (see, e.g., [5]), thereby adapting to a possible spatial non-stationarity of the noise covariance, i.e., changes in the location of the noise. However, in order to be robust to the problem of self-nulling due to, e.g., array calibration errors, this technique requires the updating of the noise covariance estimate to be restricted to time instances where only noise is present in the measured signal [6]. In practice, such periods are detected using so-called voice activity detection algorithms [7] or by using the speech presence probability [8]. The MVDR approach to noise reduction may thus be summarized as determining spatial filters defined by covariance matrix estimates obtained during times where only the noise is present in the signal, and applying these filters at times where also the desired signal is present. The success of such a strategy thus relies on the assumption that the noise covariance matrix during periods where the speech component is present is equal to the covariance at times when actual estimates may be formed, i.e., it is assumed that the interferer does not move during voice activity periods. Violations of these assumptions may render the MVDR filter ineffective, as the locations of the nulls of the filter and the location of the interferer may then not coincide.

In this work, we propose to address this issue by means of a matrix interpolation strategy using recently developed tools from optimal mass transport (OMT) [9]. Originally introduced for modeling the problem of efficiently supplying construction sites with building material, the topic of OMT has lately gained interest in the fields of signal processing [10], automatic control [11], and machine learning [12], [13], with applications including convex clustering [14], sensor fusion [15], [16], smooth morphing of speech signals [17], spectral estimation [18], and graph signal processing [19].

Recently, OMT was used to define a measure of distance between covariance matrices [20] by relating these matrices to an underlying spectral domain. In this framework, the distance between two covariance matrices is defined in terms of the cost of morphing their spectral representations to each other. For spatial covariance matrices, this implies that the matrix distance directly corresponds to the distance between different source locations. In addition, the OMT framework presented in [20] provides a way of finding the most probable path of movement, as well as a way of defining interpolating covariance matrices.

In this work, we propose to utilize these ideas in order to improve the MVDR based noise reduction in time intervals in which the noise covariance matrix cannot be estimated directly from the data. Specifically, we propose to use two covariance matrix estimates, one estimated from data measured before the voice activation period, and one estimated from data measured after the voice activation period, in order to find an OMT induced covariance matrix interpolant, defining a

This work was supported in part by the Swedish Research Council, the Crafoord foundation, and the Royal Physiographic Society in Lund. This research work was carried out in part at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven Internal Funds C2-16-00449 and VES/19/004. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

sequence of matrices connecting the two estimates. Implicitly, this corresponds to reconstructing the spatial path taken by the interfering source during the non-observable period. As the proposed method constitutes a strategy for interpolating the noise covariance matrix, it is primarily intended for offline noise reduction applications. We demonstrate the performance of the proposed method using simulations of a speech enhancement scenario, demonstrating the viability of approach even if simplified models are used to construct the OMT interpolant.

II. Background

A. Speech enhancement

Consider a scenario where $p \in \mathbb{N}$ microphones measure a mixture of a desired signal, e.g., speech, and additive noise, with the objective being to reduce the interference of the noise. We will herein address the problem in the short-time Fourier transform (STFT) domain. Specifically, considering frequency f and time t, let the signal measured by the microphone array be modeled as

$$\mathbf{y}(f,t) = \mathbf{h}(f,t)s(f,t) + \mathbf{n}(f,t),$$

where the signal vector is detailed as

$$\mathbf{y}(f,t) = \begin{bmatrix} y_1(f,t) & y_2(f,t) & \dots & y_p(f,t) \end{bmatrix}^T. \quad (1)$$

Here, $s(t,f)$ denotes the desired speech signal, $\mathbf{n}(f,t)$ denotes the additive noise component, and $\mathbf{h}(f,t)$ denotes the acoustic transfer function corresponding to the source. Dropping the dependence on the frequency f for notational brevity[1], let $\mathbf{R}(t)$ denote the spatial noise covariance matrix at time t, i.e.,

$$\mathbf{R}(t) \triangleq \mathbb{E}\left(\mathbf{n}(t)\mathbf{n}(t)^H\right),$$

where $\mathbb{E}(\cdot)$ is the expectation operator. A common approach to performing noise reduction is then to estimate the speech signal $s(t)$ as $\hat{s}(t) = \mathbf{w}(t)^H\mathbf{y}(t)$, where the spatial filter $\mathbf{w}(t)$ is the MVDR beamformer, i.e.,

$$\mathbf{w}(t) = \arg\min_{\mathbf{w}} \ \mathbf{w}^H\mathbf{R}(t)\mathbf{w} \ \text{ subject to } \ \mathbf{w}^H\mathbf{h}(t) = 1. \quad (2)$$

In practice, the noise covariance $\mathbf{R}(t)$ is replaced by an estimate $\hat{\mathbf{R}}(t)$, commonly obtained using the recursively defined exponentially smoothed estimator

$$\hat{\mathbf{R}}(t) = (1-\eta)\hat{\mathbf{R}}(t-1) + \eta\mathbf{n}(t)\mathbf{n}(t)^H,$$

for some user-specified forgetting factor $\eta \in [0,1]$. However, in order to introduce robustness to the phenomenon of self-nulling [6], i.e., cancelling of the desired signal, estimates of $\mathbf{R}(t)$ have to be formed from signal samples in (1) wherein only the noise component is active. Thus, in practice, noise cancellation at time t, i.e., when the desired signal is present, relies on the assumption that $\mathbf{R}(t) = \mathbf{R}(t')$ for some other time t' at which an estimate of $\mathbf{R}(t')$ can be formed from the measured signal. This assumption may be violated in scenarios wherein the noise source is moving, potentially leading to severe performance degradation of the noise suppression due to the mismatch between the assumed and actual covariances.

In this work, we aim to mitigate such effects by exploiting the spectral representation of the noise covariance matrix and utilize recent results on matrix interpolation building on the concept of optimal mass transport [20]. Specifically, we will consider improving the performance of the MVDR beamformer in (2) on a time interval $[t_0, t_1]$ during which the speech component is active. Assuming that we have access to covariance matrix estimates $\hat{\mathbf{R}}(t_0)$ and $\hat{\mathbf{R}}(t_1)$, corresponding to the times just before and just after the speech activity period, respectively, we will construct interpolating matrices $\hat{\mathbf{R}}(t)$, for $t \in (t_0, t_1)$, using tools from OMT.

B. Spectral Representations

Assume that the space under consideration, i.e., the room in which the signal is observed, can be considered to be well described by the spatial domain $\mathcal{X} \subset \mathbb{R}^3$. Then, the spatial spectrum, $\Phi \in \mathcal{M}_+(\mathcal{X})$, where $\mathcal{M}_+(\mathcal{X})$ is the space of non-negative measures on $\mathcal{X}$, describes the distribution of noise power on $\mathcal{X}$. Also, there is a function $\mathbf{a} : \mathcal{X} \to \mathbb{C}^p$ such that the noise covariance matrix may be expressed as $\mathbf{R} = \Gamma(\Phi)$, where $\Gamma : \mathcal{M}_+(\mathcal{X}) \to \mathbb{C}^{p\times p}$ is the operator

$$\Gamma(\Phi) \triangleq \int_{\mathcal{X}} \mathbf{a}(x)\Phi(x)\mathbf{a}(x)^H dx, \quad (3)$$

where dx is the integration measure on $\mathcal{X}$. The function $\mathbf{a}$ is the array manifold vector, encoding the properties of the acoustic environment. The spectral representation of covariance matrices directly explains performance degradations caused by covariance mismatch: shifts in the distribution of spectral power, corresponding to movement of the noise source, implies that the nulls of the spatial filter $\mathbf{w}$ are not placed such that the resulting filter will effectively suppress the noise. It may be noted that for $\mathcal{X} = \mathbb{T} \triangleq [-\pi/2, \pi/2)$ and

$$\mathbf{a}(\theta) = \begin{bmatrix} 1 & e^{-i2\pi\frac{d\sin(\theta)}{\lambda}} & \dots & e^{-i2\pi(p-1)\frac{d\sin(\theta)}{\lambda}} \end{bmatrix}^T, \quad (4)$$

we get the far-field, uniform linear array (ULA) scenario in anechoic environment [21], with d and λ corresponding to the sensor spacing and signal wavelength, respectively, implying that the covariance matrix is in the range of the Toeplitz operator

$$\Gamma(\Phi) = \int_{\mathbb{T}} \mathbf{a}(\theta)\Phi(\theta)\mathbf{a}(\theta)^H \frac{d\theta}{2\pi}.$$

We will utilize this close connection between covariance matrices $\mathbf{R}$ and power spectra Φ in order to improve the estimate of the noise covariance.

C. Optimal Mass Transport and Matrix Interpolation

The Monge-Kantorovich problem of optimal mass transport is concerned with finding the most cost-efficient way of morphing one distribution of mass to another [9]. Considering

[1] We will throughout this work treat each frequency f independently. Thus, we omit the explicit frequency dependence in the considered quantities.

two power spectra $\Phi_0, \Phi_1 \in \mathcal{M}_+(\mathcal{X})$, one may define their distance as the minimum value of [18]

$$\begin{aligned} &\underset{M \in \mathcal{M}_+(\mathcal{X}^2)}{\text{minimize}} \quad \Psi_c(M) \\ &\text{subject to} \quad \int_{\mathcal{X}} M(\cdot, x_1) dx_1 = \Phi_0, \\ &\qquad\qquad \int_{\mathcal{X}} M(x_0, \cdot) dx_0 = \Phi_1, \end{aligned} \tag{5}$$

where $\Psi_c : \mathcal{M}_+(\mathcal{X}^2) \to \mathbb{R}_+$ is detailed as

$$\Psi_c(M) = \int_{\mathcal{X}^2} c(x_0, x_1) M(x_0, x_1) dx_0 dx_1, \tag{6}$$

with the cost function $c : \mathcal{X}^2 \to \mathbb{R}_+$ detailing the cost of transporting spectral mass between points of $\mathcal{X}$, and the transport plan M describes the amount of mass transported between different points. By relating covariance matrices to spectral representations, distances between covariance matrices have recently been defined in terms of transport problems [20]. Specifically, the distance between two covariance matrices $\mathbf{R}(0)$ and $\mathbf{R}(1)$ may be defined as the minimum value of

$$\begin{aligned} &\underset{M \in \mathcal{M}_+(\mathcal{X}^2)}{\text{minimize}} \quad \Psi_c(M) \\ &\text{subject to} \quad \Gamma\left(\int_{\mathcal{X}} M(\cdot, x_1) dx_1\right) = \mathbf{R}(0), \\ &\qquad\qquad \Gamma\left(\int_{\mathcal{X}} M(x_0, \cdot) dx_0\right) = \mathbf{R}(1). \end{aligned} \tag{7}$$

This formulation does not only allow for describing distances between matrices in terms of the geometry of the underlying space of interest, as reflected in the cost function c, it also directly provides the means for defining intermediate matrices $\mathbf{R}(\tau)$, for $\tau \in (0,1)$, in terms of the obtained transport plan M. As in [20], we herein consider the interpolant $\mathbf{R}(\tau) = \Gamma\left(\Phi_\tau^M\right)$ where

$$\Phi_\tau^M(x) = \int_{\mathcal{X}^2} \delta\left(x_0 + \tau x_1 - x\right) M(x_0, x_0 + x_1) dx_0 dx_1, \tag{8}$$

with δ denoting the Dirac delta, i.e., the mass transported from x_0 to x_1 is at interpolation time $\tau \in (0,1)$ located at $(1-\tau)x_0 + \tau x_1$. This interpolant is implied by imposing minimal assumptions on the structure of change in the spectral distribution power; in effect, any movement implied by the transport plan M is considered to be performed along straight lines. If no prior knowledge of, e.g., movement speed or acceleration is available, this is a reasonable assumption.

III. Proposed method

Recalling, for a time interval $[t_0, t_1]$ where the speech signal is present, one wishes to suppress the noise using the MVDR beamformer in (2), despite not being able to estimate the noise covariance matrix $\mathbf{R}(t)$, for $t \in [t_0, t_1]$, directly from the data. Without loss of generality, one may normalize the time axis such that the time interval of interest is indexed by $\tau \in [0,1]$. Assuming that one is able to form estimates $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$ corresponding to the times just before and after speech activation period, we propose to form estimates of the unobservable noise covariance matrix $\mathbf{R}(\tau)$, for $\tau \in (0,1)$, as the OMT induced interpolating matrix. That is, we propose to

1) obtain a transport plan M by solving (7) using $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$ as data,
2) compute interpolating covariance matrices as $\mathbf{R}(\tau) = \Gamma\left(\Phi_\tau^M\right)$, for $\tau \in (0,1)$, with Φ_τ^M defined in (8),
3) use the obtained $\mathbf{R}(\tau)$ to determine the MVDR filter in (2).

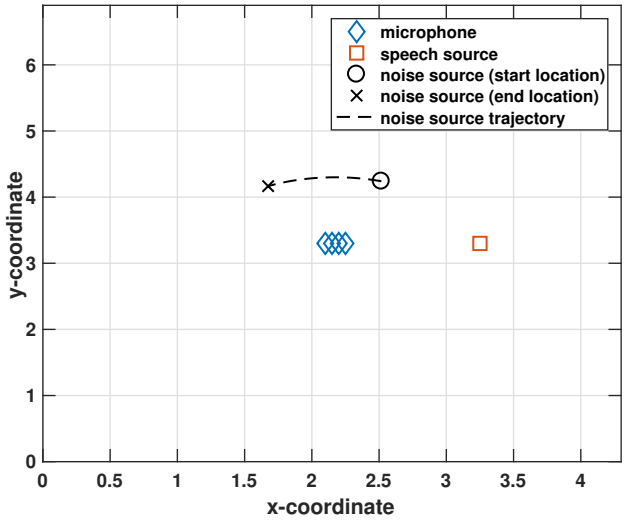

Fig. 1. Measurement setup consisting of 4 microphones arranged as a ULA, a target speech source, and a moving noise source.

Assuming that the interval $[t_0, t_1]$ is relatively short, e.g., corresponding to a few seconds, we argue that the linear interpolation on the underlying space $\mathcal{X}$ is a reasonable approximation, as it corresponds to the smoothest displacement of the noise source, i.e., minimal acceleration. This may be contrasted with the interpolant induced by the Euclidean metric, i.e., convex combinations of $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$, that, instead of modeling displacement, implies fading in and fading out of the interfering source (see also [20]).

From a practical point of view, it may be noted that due to imperfections in the assumed model, or even, simplifications of the model[2], reflected in the choice of $\mathcal{X}$ and $\mathbf{a}$, the estimated covariance matrices, as well as the actually true covariance matrices, may not be in the range of the chosen operator Γ. Also, in practical implementations of the method, we consider a discretization of the problem (7), i.e., the transport plan M and cost function c are represented by matrices $\mathbf{M}$ and $\mathbf{C}$, respectively. In order to efficiently compute the transport plan $\mathbf{M}$ used to define the interpolating covariance matrices, we propose to augment the objective function Ψ_c by an entropy regularization term $D(\mathbf{M})$, defined as

$$D(\mathbf{M}) = \epsilon \sum_{k,\ell=1}^{N} \left(m_{k,\ell} \log(m_{k,\ell}) - m_{k,\ell} + 1\right)$$

[2]One may, for example choose $\mathcal{X} = \mathbb{T}$ and use a far-field model.

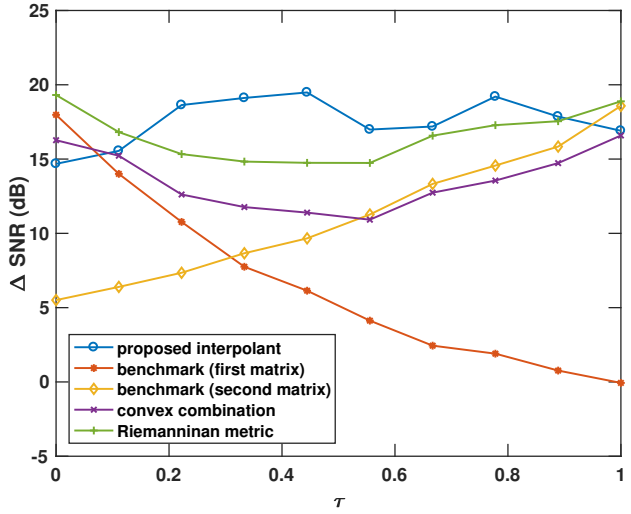

Fig. 2. Increase in SNR throughout the speech segment using different noise covariance matrix interpolants, with the time index τ being normalized to $\tau \in [0, 1]$. The reverberation time is 0 ms.

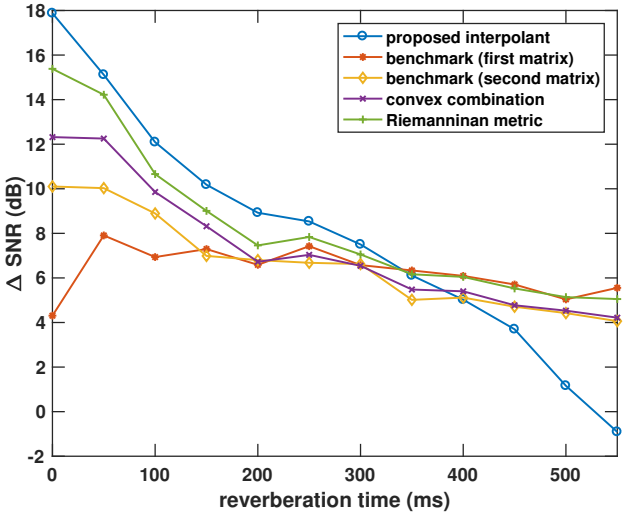

Fig. 3. Increase in SNR using different noise covariance matrix interpolants for different reverberation times.

where $\epsilon > 0$ is a regularization parameter, N is the number of points used in the discretization of $\mathcal{X}$, and where $m_{k,\ell}$ denotes the entries of the transport plan matrix $\mathbf{M}$, i.e., $m_{k,\ell}$ is the mass transported from discretization point k to discretization point ℓ. This type of regularization was in [22] proposed for regularizing problems of the form (5), as it allows for computationally efficient solvers based on Sinkhorn iterations.

As $D(\mathbf{M})$ introduces some smearing to the solution, the regularization parameter ϵ should be chosen small as to not greatly affect the obtained transport plan $\mathbf{M}$. However, the numerical precision used in the computations introduces a lower bound for the choice of ϵ (see [22] for a discussion on this). We have recently extended such iterations as to generalize to inverse problems on the form (7) (see [23] for details), and we will in the numerical examples herein use these in order to solve

$$\begin{aligned} \underset{\mathbf{M},\Delta_0,\Delta_1}{\text{minimize}} \quad & \langle \mathbf{M}, \mathbf{C} \rangle + D(\mathbf{M}) + \gamma \left(\|\Delta_0\|_F^2 + \|\Delta_1\|_F^2 \right) \\ \text{subject to} \quad & \Gamma\left(\mathbf{M}^T \mathbf{1}\right) = \Delta_0 + \hat{\mathbf{R}}(0), \\ & \Gamma\left(\mathbf{M}\mathbf{1}\right) = \Delta_1 + \hat{\mathbf{R}}(1), \end{aligned} \tag{9}$$

where $\gamma > 0$ is a regularization parameter, $\langle \cdot, \cdot \rangle$ is the matrix inner product, i.e., the discrete counterpart of (6), and $\mathbf{1}$ is a vector of ones of appropriate dimension. This problem is, with the addition of some additional terms, a direct discretization of the problem in (7). It may be noted that, in effect, augmenting the linear constraints by the deviation variables Δ_0, Δ_1, along with the penalization of their Frobenius norm, amounts to projecting the estimates $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$ onto the range of Γ, allowing for model mismatch. The operator Γ should in this context be interpreted as the discretized counterpart of (3).

IV. Numerical results

To illustrate the behavior of the proposed method, we consider the measurement setup illustrated in Figure 1. It consists of a ULA consisting of $p = 4$ microphones with spacing 0.05 meters, placed in a room of dimensions $4.3 \times 6.9 \times 2.6$ meters. The speech source is placed in the end-fire direction of the array. The speech source is active during 1.6 seconds, during which the localized noise source moves with constant speed between the two indicated locations along the detailed trajectory. The speech source consists of a male speaker uttering a phrase from the HINT database [24] and the noise source consists of a babble noise excerpt from [25]. The measured signals are constructed using acoustic impulse responses computed using the randomized image method [26]. We use the weighted overlap and add method [27] using 256 frequency bins and a sampling frequency of 16 kHz. Denoting the times just before and just after the voice activity period by $\tau = 0$ and $\tau = 1$, we estimate the spatial noise covariance matrix using the sample covariance matrix from the noise-only signal samples, thus obtaining two estimates $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$ per temporal frequency. We then construct matrix interpolants $\hat{\mathbf{R}}(\tau)$, for $\tau \in (0, 1)$, using the proposed method, where we for simplicity utilize a far-field model for the operator Γ, i.e., $\mathcal{X} = \mathbb{T}$, and use the cost function $c(\theta, \varphi) = \left| e^{i\theta} - e^{i\varphi} \right|^2$ for $\theta, \varphi \in \mathbb{T}$. When solving (9), the spatial domain was discretized into a uniform grid of $N = 250$ grid points, and the parameters $\epsilon = 10^{-3}$ and $\gamma = 100/p^2$ were used. The obtained interpolants $\hat{\mathbf{R}}(\tau)$ are then used in (2) to obtain an MVDR filter used to suppress the noise. The array steering vector used in the MVDR constraint corresponds to the simplified far-field model in (4). Figure 2 presents the resulting increase in signal-to-noise ratio (SNR) between the measured and filtered signals corresponding to the right-most microphone throughout the interpolation period for a non-reverberant scenario. Here, the increase in SNR in dB, at time τ, is defined as

$$\Delta\text{SNR}(\tau) = 10 \log_{10} \frac{\tilde{\sigma}_{\text{s}}^2(\tau)}{\tilde{\sigma}_{\text{n}}^2(\tau)} - 10 \log_{10} \frac{\sigma_{\text{s}}^2(\tau)}{\sigma_{\text{n}}^2(\tau)}$$

where $\tilde{\sigma}_s^2$ and $\tilde{\sigma}_n^2$ are the powers of the filtered speech and noise components, and σ_s^2 and σ_n^2 are the powers of the measured speech and noise components.

As comparison, we include results obtained using fixed interpolants, i.e., interpolants identical to either $\hat{\mathbf{R}}(0)$ or $\hat{\mathbf{R}}(1)$, as well as their (time-varying) convex combination $(1-\tau)\hat{\mathbf{R}}(0) + \tau\hat{\mathbf{R}}(1)$. Also included is the time-varying interpolant induced by the intrinsic Riemannian metric on the cone of positive definite matrices [28]. As can be seen, the performance of the comparison interpolants vary significantly throughout the considered period due to the movement of the noise source. It may be noted that the output SNR of the fixed interpolants $\hat{\mathbf{R}}(0)$ and $\hat{\mathbf{R}}(1)$ are strictly decreasing and increasing, respectively. In contrast, the proposed method is able to achieve a more consistent output SNR due to its ability to model spatial displacement of spectral power, even though a simplified model is considered here.

Considering the effect of increasing the deviation from the assumed model, Figure 3 presents the output SNR for different levels of reverberation. Here, the presented SNR is computed using the complete speech segment, excluding the beginning and the end of the speech segment, as to focus on intervals where interpolation is required. As may be noted, the performance of all considered interpolants decrease rapidly as the reverberation time increases. Interestingly, the performance of the fixed interpolant $\hat{\mathbf{R}}(0)$ actually increases, likely due to the smearing of the spatial spectrum induced by the reverberation. It may also be noted that the proposed method performs better than the comparison interpolants for moderate levels of reverberation, even though the presence of any form of reverberation is not taken into consideration in the utilized model. Audio samples from these simulations may be heard at [29].

V. Conclusion

In this work, we have proposed to utilize a matrix interpolation technique based on the concept of optimal mass transport in order to improve the performance of the MVDR beamformer in offline multi-microphone noise reduction, specifically addressing scenarios featuring moving interferers. In time intervals where the noise covariance matrix cannot be estimated directly from the microphone measurements, the proposed method uses two end point matrix estimates in order to reconstruct the spatial path taken by the interfering source during the non-observable period. Using realistic simulations, the proposed method has been shown to yield improved performance, even when using simplified signal models in the construction of the OMT interpolant.

References

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

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Room: Robustness Aagainst Reverberation for Automatic Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] J. Capon, "High Resolution Frequency Wave Number Spectrum Analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, 1969.

[4] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing, Techniques and Applications*. New York: Springer, 2001.

[5] S. Haykin, *Adaptive Filter Theory (4th edition)*. Englewood Cliffs, N.J.: Prentice Hall, Inc., 2002.

[6] H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*. John Wiley and Sons, Inc., 2002.

[7] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[8] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," *Proc. 2011 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '11)*, pp. 145–148, 2011.

[9] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008.

[10] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal Mass Transport: Signal processing and machine-learning applications," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 43–59, July 2017.

[11] Y. Chen, T. T. Georgiou, and M. Pavon, "Optimal Transport Over a Linear Dynamical System," *IEEE Trans. Autom. Control*, vol. 62, no. 5, pp. 2137–2152, May 2017.

[12] H. Ling and K. Okada, "An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, no. 5, pp. 840–853, 2007.

[13] J. Adler, A. Ringh, O. Öktem, and J. Karlsson, "Learning to solve inverse problems using Wasserstein loss," *arXiv preprint arXiv:1710.10898*, 2017.

[14] F. Elvander, S. I. Adalbjörnsson, J. Karlsson, and A. Jakobsson, "Using Optimal Transport for Estimating Inharmonic Pitch Signals," in *Proc. 42nd IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, New Orleans, LA, USA, March 5-9 2017, pp. 331–335.

[15] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Tracking and Sensor Fusion in Direction of Arrival Estimation Using Optimal Mass Transport," in *Proc. 26th European Signal Processing Conference*, Rome, Italy, Sep. 3-7 2018, pp. 1617–1621.

[16] ——, "Non-Coherent Sensor Fusion via Entropy Regularized Optimal Mass Transport," in *Proc. 44th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 13-17 2019, pp. 4415–4419.

[17] X. Jiang, Z. Q. Luo, and T. T. Georgiou, "Geometric Methods for Spectral Analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1064–1074, Mar. 2012.

[18] T. T. Georgiou, J. Karlsson, and M. S. Takyar, "Metrics for power spectra: an axiomatic approach," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 859–867, Mar. 2009.

[19] N. Saito, "How Can We Naturally Order and Organize Graph Laplacian Eigenvectors?" in *Proc. 2018 IEEE Stat. Signal Process. Workshop*, June 10-13 2018.

[20] F. Elvander, A. Jakobsson, and J. Karlsson, "Interpolation and Extrapolation of Toeplitz Matrices via Optimal Mass Transport," *IEEE Trans. Signal Process*, vol. 66, no. 20, pp. 5285 – 5298, Oct. 2018.

[21] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, N.J.: Prentice Hall, 2005.

[22] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2292–2300, 2013.

[23] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-Marginal Optimal Mass Transport with Partial Information," *arXiv:1905.03847*, 2019.

[24] M. Nilsson, S. D. Soli, and J. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[25] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec, St. Louis," St. Louis, 1997.

[26] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the Modeling of Rectangular Geometries in Room Acoustic Simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, Apr. 2015.

[27] R. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 28, no. 1, pp. 99–102, Feb. 1980.

[28] S. T. Smith, "Covariance, Subspace, and Intrinsic Cramér-Rao Bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, May 2005.

[29] (2018). [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/EUSIPCO2019/AudioOMT