NONLINEAR SPATIAL FILTERING FOR MULTICHANNEL SPEECH ENHANCEMENT IN INHOMOGENEOUS NOISE FIELDS

Kristina Tesch and Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
kristina.tesch@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

A common processing pipeline for multichannel speech enhancement is to combine a linear spatial filter with a single-channel postfilter. In fact, it can be shown that such a combination is optimal in the minimum mean square error (MMSE) sense if the noise follows a multivariate Gaussian distribution. However, for non-Gaussian noise, this serial concatenation is generally suboptimal and may thus also lead to suboptimal results. For instance, in our previous work, we showed that a joint spatial-spectral nonlinear estimator achieves a performance gain of 2.6 dB segmental signal-to-noise ratio (SNR) improvement for heavy-tailed large-kurtosis multivariate noise compared to the traditional combination of a linear spatial beamformer and a postfilter.

In this paper, we show that a joint spatial-spectral nonlinear filter is not only advantageous for noise distributions that are significantly more heavy-tailed than a Gaussian but also for distributions that model inhomogeneous noise fields while having rather low kurtosis. In experiments with artificially created noise we measure a gain of 1 dB for inhomogenous noise with low kurtosis and up to 2 dB for inhomogeneous noise fields with moderate kurtosis.

Index Terms— Multichannel, speech enhancement, nonlinear filtering, acoustic beamforming

1. INTRODUCTION

Speech enhancement algorithms are used to recover a target speech signal from microphone recordings that are corrupted by background noise. These techniques are fundamental to many communication applications such as telephony, hearing aids, and the emerging field of human-machine interaction with an automatic speech recognition (ASR) system.

If several recordings of the target signal from multiple microphones are available, multichannel speech enhancement methods can be used. The advantage of these methods over single-channel approaches, e.g., [1, 2, 3], is that not only tempo-spectral but also spatial information can be included in the processing [4]. In many cases, the spatial filtering is carried out using a so-called *beamformer* that emphasizes a signal from a certain direction and suppresses the signal components originating from other directions. Beamforming is a linear operation: the discrete Fourier transform (DFT) coefficients of all channels are multiplied by complex weights and summed [5].

Commonly, a single-channel method is applied to the output of such a linear spatial filter to further exploit spectral characteristics for suppressing the remaining noise. It is often referred to as a *postfilter*. However, this common processing pipeline, despite its prevalence, is in general suboptimal if the noise does not follow a multivariate complex Gaussian distribution.

Balan and Rosca [6] have shown that the clean speech MMSE estimator for multivariate complex Gaussian noise can be separated into an minimum variance distortionless response (MVDR) beamformer and a single-channel postfilter. In contrast, the work of Hendriks et al. [7] revealed that the MMSE solution for noise that follows a multivariate complex Gaussian mixture distribution inseparably joins the spatial and spectral processing and is even nonlinear in the spatial filter. From these results, it becomes clear that the noise distribution plays an important role in determining whether joint spatial-spectral nonlinear processing could lead to an improved performance. In the sequel, we may refer to an estimator that joins the spatial and spectral processing into a single nonlinear operation a *nonlinear spatial filter*.

It is important to note that characterizing the noise scenarios in which a nonlinear filtering is advantageous gains particular relevance in the context of the neural network revolution. Evermore often, neural networks are trained to solve single-channel and multichannel speech enhancement tasks, e.g., [8, 9, 10, 11]. While neural networks could potentially be used to elegantly approximate nonlinear joint spatial-spectral filters, most neural network approaches for multichannel speech enhancement restrict the spatial filter to be linear [10, 12] or use neural networks just for estimating the beamformer parameters [13]. In contrast, using neural networks for modeling a nonlinear spatial filter is far less common, e.g., [11]. This is also because the potential benefit of using nonlinear spatial filters is not fully understood. Tackling this problem experimentally by trying out different network architectures does not seem to be a satisfying approach to fundamentally understand the potential gain of nonlinear spatial filtering. For instance, network architectures that are more complex than necessary are generally undesirable as they require more data and training time. Therefore, it is important to understand for which noise scenarios learning a nonlinear spatial filter is worthwhile and for which not. For this, we compare the performance obtained by statistical MMSE-optimal estimators to be able to gain more general insights without depending e.g. on specific neural network architectures.

Already in our recent previous work [14], we evaluated the benefit of the optimal MMSE solution of Hendriks et al. with joint spatial-spectral nonlinear filtering by comparing it to the best matching estimator composed of an MVDR beamformer and an MMSE single-channel postfilter. However, in this analysis we obtained Gaussian mixtures by combining Gaussian components with the *same spatial structure* but different scaling. We observed for noise distributions with a high kurtosis, which measures the heavy-tailedness of a distribution [15, 16], a gain of 2.6 dB segmental SNR and 0.5 POLQA score improvement. Furthermore, in [14] we observed a gain of 1.2 dB segmental SNR improvement for noisy mixtures with real-world noise recordings taken from the CHiME-3 data set [17] when fitting a zero-mean multivariate complex Gaussian mixture with four components to the data. Since the nonlinear spatial filter delivers better results than separated processing with a linear spatial filter and a

postfilter, one may conclude that the fitted distribution is not Gaussian and may speculate that the distribution has a notably larger kurtosis than a Gaussian distribution. However, examining the kurtosis of the distributions fitted to the CHiME-3 data revealed that the kurtosis is surprisingly low (Section 3). Thus, it seems that the advantage of a joint nonlinear spatial-spectral filter does not only depend on the kurtosis of the noise distribution but also on other properties. The goal of this paper is to analyze how much the *spatial structure* of the noise model impacts performance when using a joint nonlinear spatial-spectral filter instead of the traditional serial concatenation of a linear beamformer and a postfilter.

Section 2 introduces the modeling assumptions and statistical estimators that our analysis is based on. In Section 3, we show that solely the kurtosis of the noise distribution is not sufficient to characterize when the use of a nonlinear spatial filter could be worthwhile and in Section 4 we evaluate to what extent spatial properties of the noise distribution influence the gain achieved with a nonlinear spatial filter.

2. THEORETICAL BACKGROUND

2.1. Signal model

We assume that the target speech signal is disturbed by additive noise and recorded by a microphone array with D microphones. The recorded time-domain signal for every microphone-channel $\ell \in \{1, ..., D\}$ is transformed to the frequency-domain using a windowed DFT yielding DFT coefficients $Y_\ell(k,i) \in \mathbb{C}$ with frequency-bin index k and time-frame index i. As we assume the noise to be additive, we obtain the noisy DFT coefficient $Y_\ell(k,i) \in \mathbb{C}$ as the sum of the clean speech and the noise DFT coefficients $S_\ell(k,i) \in \mathbb{C}$ and $S_\ell(k,i)$, i.e.,

$$Y_\ell(k,i) = S_\ell(k,i) + N_\ell(k,i). \tag{1}$$

We model the DFT coefficents as random variables and assume that they are independent with respect to the frequency-bin index and time-frame index. As a result, we can consider every time-frequency bin separately and omit the frequency-bin and time-frame indices from the notation. Uppercase letters will denote random variables and lowercase letters will be used for their realizations. Further, we assume speech and noise to be uncorrelated and zero-mean.

The noise DFT coefficients of all channels are combined into a vector $\mathbf{N} = [N_1, ..., N_D]^T \in \mathbb{C}^D$ with correlation matrix $\mathbf{\Phi}_\text{n} = \mathbb{E}[\mathbf{N}\mathbf{N}^H]$. The vector of clean speech DFT coefficients is given by $\mathbf{S} = \mathbf{d}S \in \mathbb{C}^D$ with the so-called steering vector $\mathbf{d} \in \mathbb{C}^D$ modeling the propagation path from the single target speaker to the microphones. Then, the vector $\mathbf{Y} = \mathbf{S} + \mathbf{N} \in \mathbb{C}^D$ contains the noisy DFT coefficients for every channel. The spectral power of the clean speech signal S is denoted by $\sigma_\text{s}^2 = \mathbb{E}[|S|^2]$.

2.2. Estimators

In our previous work [14], we gathered theoretical results to point out that the MMSE solution can be separated into an MVDR beamformer defined as

$$T_\text{MVDR}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_\text{n}^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_\text{n}^{-1} \mathbf{d}} \tag{2}$$

and a single-channel postfilter *if the noise follows a multivariate complex Gaussian distribution*. However, this also implies that for the separability into a linear spatial filter concatenated with a spectral postfilter the distribution of the noise plays a decisive role. This becomes clear from the result of Hendriks et al. [7] who show that the MMSE-optimal estimator of the clean speech DFT coefficient S for noise that is distributed according to a multivariate complex Gaussian mixture distribution is in general a nonlinear and non-separable joint spatial-spectral filter.

This Gaussian mixture distribution combines M zero-mean Gaussian components with covariance matrices $\mathbf{\Phi}_\text{m} \in \mathbb{C}^{D \times D}$, $m = 1, ..., M$, and the correspondig noise probability density function (PDF) is given by

$$p(\mathbf{n}) = \sum_{m=1}^{M} c_m \frac{1}{\pi^D |\mathbf{\Phi}_\text{n}|} \exp\left\{-\mathbf{n}^H \mathbf{\Phi}_\text{n}^{-1} \mathbf{n}\right\} \tag{3}$$

with component weights c_m that sum to one. The estimator T_MMSE for multivariate complex Gaussian mixture distributed noise has been derived by Hendriks et al. [7] under the additional assumption that the clean speech signal amplitude follows a generalized-Gamma distribution with a shape parameter $\nu \in \mathbb{R}^+$ and that the phase $\Psi \in [0, 2\pi)$ is uniformly distributed and independent of the speech amplitude. Then, the MMSE solution is given by

$$T_\text{MMSE}(\mathbf{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_\text{m}|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \frac{\sigma_\text{s}^2 T_\text{MVDR}^{(m)}(\mathbf{y}) \mathcal{M}(\nu+1,2,P_m)}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_\text{s}^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_\text{m}|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{4}$$

with

$$T_\text{MVDR}^{(m)}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d}}, \quad Q_m = (\nu + \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \sigma_\text{s}^2)^{-\nu},$$

$$\text{and} \quad P_m = \frac{\sigma_\text{s}^2 \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d} \left|T_\text{MVDR}^{(m)}(\mathbf{y})\right|^2}{\nu(\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_\text{s}^2}$$

and $\mathcal{M}(\cdot,\cdot,\cdot)$ being the confluent hypergeometric function [18, Sec. 9.21].

This MMSE estimator cannot be separated into a spatial filter and a spectral postfilter since the observation $\mathbf{y}$ is the input of the linear function $T_\text{MVDR}^{(m)}$, which in turn depends on the summation index, and also occurs in the quadratic term $\exp\left\{-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y}\right\}$. The latter highlights the spatial nonlinearity of the solution.

In [14], we have experimentally quantified the benefit of the nonlinear joint spatial-spectral MMSE-optimal solution T_MMSE over a separated solution $T_\text{MVDR-MMSE}$ that combines an MVDR beamformer with an MMSE-optimal postfilter. We derived the MMSE postfilter under the same assumptions used to compute T_MMSE to allow for a meaningful comparison. This results in the composite estimator

$$T_\text{MVDR-MMSE}(\mathbf{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_\text{MVDR}(\mathbf{y})|^2}{\sigma_m^2}\right]} \frac{\sigma_\text{s}^2 T_\text{MVDR}(\mathbf{y}) \mathcal{M}(\nu+1,2,P_m)}{\nu\sigma_m^2 + \sigma_\text{s}^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_\text{MVDR}(\mathbf{y})|^2}{\sigma_m^2}\right]} \mathcal{M}(\nu, 1, P_m)} \tag{5}$$

with

$$\mathbf{\Phi}_\text{n} = \sum_{m=1}^{M} c_m \mathbf{\Phi}_\text{m}, \quad \sigma_m^2 = \frac{\mathbf{d}^H \mathbf{\Phi}_\text{n}^{-1} \mathbf{\Phi}_\text{m} \mathbf{\Phi}_\text{n}^{-1} \mathbf{d}}{(\mathbf{d}^H \mathbf{\Phi}_\text{n}^{-1} \mathbf{d})^2},$$

$$Q_m = \left(\frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_\text{s}^2}\right)^{-\nu} \quad \text{and} \quad P_m = \frac{\sigma_\text{s}^2 \sigma_m^{-2} |T_\text{MVDR}(\mathbf{y})|^2}{\nu\sigma_m^2 + \sigma_\text{s}^2}.$$

The separability into the MVDR beamformer and a single-channel postfilter can be seen from the fact that the observation is contained in this equation only as input to the MVDR beamformer.

Our previous experiments with known noise distributions indicate a dependence of the performance gain achieved by the spatially nonlinear T_{MMSE} on the kurtosis and, thus, on the heavy-tailedness of the noise distribution. For the real-world noise recordings from the CHiME-3 dataset [17] we observed a moderate improvement by using a non-linear spatial filter but did not yet investigate the kurtosis value of the fitted distribtions.

3. MULTIVARIATE KURTOSIS OF CHIME-3 NOISE DATA

The CHiME-3 dataset provides multichannel recordings obtained in different environments: on a moving bus, in a cafeteria, next to a busy street and in a pedestrian area [17]. For our analysis, we use recordings from five front-facing microphones that have been embedded in a frame around a tablet computer. To approximate the unknown and potentially time-variant distribution of the recorded noise data with a zero-mean multivariate complex Gaussian mixture distribution, we apply the expectation maximization (EM) algorithm to windows of length 750 ms that overlap by 50%.

We use the definition of the multivariate kurtosis by Mardia [19], which we extend for the complex-valued case based on the equivalence of a D-dimensional complex Gaussian distribution with a $2D$-dimensional real Gaussian distribution [20, Thm. 15.1]. Then, the kurtosis of a complex-valued random vector $\mathbf{X} \in \mathbb{C}^D$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_{\text{x}}$ is given by

$$\kappa_{\mathbb{C}}(\mathbf{X}) = \mathbb{E}\left[\left(2(\mathbf{X}-\boldsymbol{\mu})^H \mathbf{C}_{\text{x}}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)^2\right]. \tag{6}$$

The kurtosis of a D-dimensional complex Gaussian distributed random vector $\mathbf{X} \in \mathbb{C}$ depends solely on the dimension D through

$$\kappa_{\mathbb{C}}(\mathbf{X}) = 2D(2D+2). \tag{7}$$

We now normalize all kurtosis values by the kurtosis of the Gaussian distribution with the corresponding dimensionality and name the result the *kurtosis factor* q. Thus, a kurtosis factor of one indicates a Gaussian distribution, while a larger kurtosis indicates a heavy-tailed distribution.

Figure 1 shows the histograms for the estimated kurtosis factors of the distributions that have been fitted to the CHiME-3 data using the EM algorithm. For this, a different number of mixture components M is used. The kurtosis as given in (6) is estimated by averaging over 1000 samples drawn from the distribution that we obtained with the EM algorithm. Using a single mixture component means to fit a Gaussian distribution and, as a result, we observe a peak at a kurtosis factor of 1 for the blue histogram. Estimating higher order statistics is generally difficult and this is reflected in the width of the peak, which shows that the estimate obeys some variance even when estimated from 1000 samples. If we add more components, i.e., $M \in \{2,3,4\}$, the peak of the histogram shifts to the right and we tend to observe larger kurtosis factors. The graphic was clipped at a kurtosis factor of 2 to improve the readability but all results are summarized in Table 1 which shows the mean and median values that confirm the observation.

In [14] we have observed that the gain obtained from T_{MMSE} in comparison $T_{\text{MVDR-MMSE}}$ reaches a value of 1.2 dB segmental SNR improvement as the number of components used to fit the noise distribution is increased to four. Here, we find that the kurtosis factor increases with the number of components and, thus, the noise distributions tend to shift towards more heavy-tailed distributions.

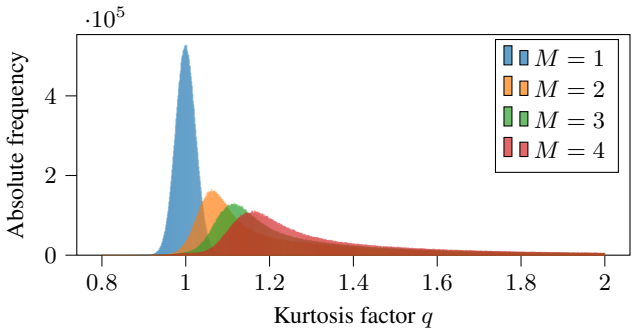

Fig. 1. Histogram of the estimated kurtosis factor for mixture distributions with M components fitted to the CHiME-3 noise data.

M	Mean	Median
1	1.00	1.00
2	1.26	1.13
3	1.36	1.20
4	1.42	1.26

Table 1. Mean and median kurtosis factor per number of components M averaged for all CHiME-3 locations (BUS, CAF, STR, PED).

However, the increase of the mean kurtosis factor up to value of 1.42 for four components is surprisingly small in comparison with the kurtosis factors that we experimented with in [14]. As a result, we conclude that the kurtosis is not the only property of the noise distribution that determines the advantage that we can expect from using the joint spatially and spectrally nonlinear estimator T_{MMSE} compared to a linear spatial filter followed by a postfilter such as $T_{\text{MVDR-MMSE}}$.

4. NONLINEAR FILTERING FOR INHOMOGENEOUS NOISE SCENARIOS

Next, we investigate the influence of spatial properties of the noise distribution on the performance of the nonlinear joint spatial-spectral T_{MMSE} compared to the concatenation of linear spatial filtering and postfiltering in $T_{\text{MVDR-MMSE}}$. For this, we set up a Gaussian mixture distribution whose Gaussian components are constructed to reassemble the spatial properties of noise point sources placed in different directions and we obtain the noise signal from sampling this multivariate complex Gaussian mixture distribution. Note that this implies that noise sources associated with this overall mixture distribution are non-Gaussian or not active for the same time-frequency bins, which is a common assumption in source separation [21].

The creation of the noise distribution is illustrated in Figure 2a. The center of the image shows a microphone array with two microphones m_1 and m_2 positioned at a distance of 5 cm. The first directional noise source n_1 stays in a fixed position 30 degrees from the target source as depicted in Figure 2a. The second noise source n_2 is placed in 20 different directions, which are indicated by the colored boxes on the circle.

For the noise sources n_1 and n_2, we can compute the steering vectors $\mathbf{d}_{n_1}$ and $\mathbf{d}_{n_2}$, which model the relative time delays of signal arrival at the microphones, based on the noise source incidence angle and the microphone array geometry. From this we construct the correlation matrices modeling the directional noise sources and some additional spatially white noise as [22]

$$\boldsymbol{\Phi}_{\text{n}_\text{i}} = (1-\alpha_{\text{wn}})\mathbf{d}_{n_i}\mathbf{d}_{n_i}^H + \alpha_{\text{wn}}\mathbf{I} \quad \text{with } i = \{1,2\}. \tag{8}$$

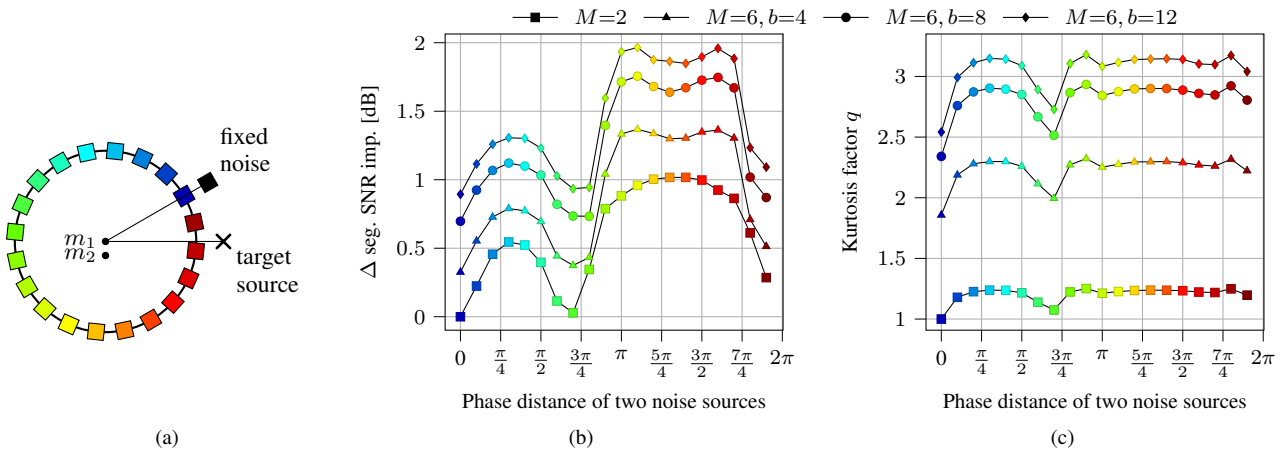

Fig. 2. (a) Illustration of the creation of a multivariate Gaussian mixture distribution modelling an inhomogeneous noise field. (b) Performance gain of T_{MMSE} over $T_{\text{MVDR-MMSE}}$ for Gaussian mixture noise modeling two directional sound sources whose placement is illustrated in Figure 2a. (c) Frequency-averaged kurtosis factor of the Gaussian mixture noise modeling two directional sound sources.

The parameter α_{wn} describes the amount of spatially white noise, which we set to $\alpha_{\text{wn}} = 0.05$, and $\mathbf{I}$ denotes the identity matrix. In our first test case we use two equally weighted zero-mean Gaussian components with correlation matrices as given in (8) to construct the overall Gaussian mixture noise distribution. In a second setting, we use three Gaussian components to model one noise source n_i and obtain their correlation matrices $\mathbf{\Phi}_{\text{n}_{\text{ij}}}, j = \{1, 2, 3\}$, by scaling the matrix $\mathbf{\Phi}_{\text{n}_{\text{i}}}$ under the constraint $\sum_{j=1}^{J} \mathbf{\Phi}_{\text{n}_{\text{ij}}} = \mathbf{\Phi}_{\text{n}_{\text{i}}}$ with $J = 3$. Based on a varying scale factor $b \in \mathbb{R}^{+}$ we compute the component correlation matrices as

$$\mathbf{\Phi}_{\text{n}_{\text{ij}}} = \frac{b^{j-1}}{r} \mathbf{\Phi}_{\text{n}_{\text{i}}} \quad \text{with} \quad r = \sum_{j=1}^{J} \frac{1}{J} b^{j-1}. \tag{9}$$

The overall Gaussian mixture distribution is then scaled such that the noisy observation has an SNR of 0 dB.

For spectral analysis and synthesis we use square-root Hann windows of length 32 ms and a 50% overlap. The speech power σ_s^2 is estimated from the clean speech signal by time-averaging over five successive time-frequency bins. We set the speech shape parameter to $\nu = 0.25$ for both estimators and evaluate each configuration on 48 speech signals that have been taken from the WJS0 dataset [23] and balanced between male and female speakers.

Figure 2b shows the performance gain of the spatially and spectrally nonlinear estimator T_{MMSE} over the classic setup $T_{\text{MVDR-MMSE}}$ based on the segmental SNR improvement. We evaluate the segmental SNR of the signals using segments of length 10 ms in which speech is present as proposed, e.g., in [24]. The mean segmental SNR of the two noisy signals is compared to the segmental SNR of the enhanced signal to obtain a measure of the improvement. The performance results are displayed with respect to phase distance of the two noise sources' incidence angles, whereby one of the noise sources moves around the microphone array counterclockwise. The marker colors have been chosen such that they indicate the moving noise source's direction in accordance with the representation in Figure 2a.

The lowest line in Figure 2b with square markers represents the results for a Gaussian mixture distributed noise with two Gaussian components. If the two noise sources are placed in the same direction (zero phase distance, dark blue marker), the Gaussian mixture distribution reduces to a Gaussian distribution and, in accordance with the theory, we cannot observe a benefit from using the joint spatial-spectral nonlinear T_{MMSE} estimator. However, we observe a clear influence of the spatial properties of the noise field and performance gains up to 1 dB.

Our previous conjecture that the kurtosis is not the only property of the noise distribution that affects the performance gain achieved with nonlinear spatial filter is confirmed by Figure 2c. It depicts the normalized kurtosis estimate from 1500 samples which has been averaged over the frequencies on the y-axis and, again, uses the phase distance between the noise sources on the x-axis. We observe rather flat courses and for instance a small kurtosis factor of about 1.2 for the lowest line representing two mixture components ($M = 2$). In particular, the performance difference of 0.5 dB segmental SNR improvement between the first maximum, located at a phase distance of roughly $\frac{\pi}{4}$, and second maximum at a phase distance between $\frac{5\pi}{4}$ and $\frac{3\pi}{2}$ of the lowest curve in Figure 2b do not go along with an increased kurtosis.

The same observation can also be made if three scaled components are used to model each noise source ($M = 6$). A larger scaling factor leads to a higher kurtosis as can be seen in Figure 2c and as we would expect. For example, we observe a kurtosis factor of 3.2 for the scaling factor $b = 12$, but still the performance difference of 0.7 dB segmental SNR improvement for the two spatial scenarios leading to the first and second maximum cannot be predicted from the kurtosis alone.

5. CONCLUSIONS

For multivariate non-Gaussian noise, the traditional concatenation of linear beamforming and spectral postfiltering is not generally optimal. Instead, the MMSE-optimal estimator generally results in a non-separable nonlinear joint spatial-spectral filter. In this paper, we provide further insights into which properties of the multichannel noise impact the potential performance gain when replacing the traditional concatenation of linear beamforming and spectral postfiltering by a joint nonlinear spatial-spectral filter. We show that besides its heavy-tailedness also the spatial structure of the noise distribution plays an important role. In our exemplary setup, we obtain performance gains of up to 2 dB segmental SNR improvement for spatially inhomogeneous noise fields with moderate kurtosis.

6. REFERENCES

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

[2] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[3] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, Ltd, 2018, ch. 5, pp. 65–85.

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

[5] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley, 2006.

[6] R. Balan and J. P. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, Rosslyn, Virginia, Aug. 2002, pp. 209–213.

[7] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 885–888, Oct. 2009.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[9] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Interspeech*, San Francisco, USA, 2016, pp. 3768–3772.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[11] H. Lee, H. Y. Kim, W. H. Kang, J. Kim, and N. S. Kim, "End-to-end multi-channel speech enhancement using inter-channel time-restricted attention on raw waveform," in *Interspeech*, Sep. 2019, pp. 4285–4289.

[12] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, G. Chen, Y. Zhang, M. Mandel, D. Yu, and M. L. Seltzer, "Deep beamforming networks for multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5745–5749.

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

[14] K. Tesch, R. Rehr, and T. Gerkmann, "On nonlinear spatial filtering in multichannel speech enhancement," in *Proc. Interspeech 2019*, Sep. 2019, pp. 91–95.

[15] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychological Methods*, vol. 2, pp. 292–307, Sep. 1997.

[16] P. H. Westfall, "Kurtosis as peakedness, 1905–2014. R.I.P." *The American Statistician*, vol. 68, no. 3, pp. 191–195, 2014.

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

[18] I. S. Gradshteyn and J. M. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 2000.

[19] K. V. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970.

[20] S. M. Kay, *Fundamentals Of Statistical Signal Processing*. Pearson, 2009.

[21] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[22] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, Jan. 2014.

[23] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," May 2007.

[24] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.