INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On Nonlinear Spatial Filtering in Multichannel Speech Enhancement

Kristina Tesch, Robert Rehr, Timo Gerkmann

Signal Processing, Universität Hamburg, Germany
kristina.tesch@uni-hamburg.de, robert.rehr@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

Using multiple microphones for speech enhancement allows for exploiting spatial information for improved performance. In most cases, the spatial filter is selected to be a linear function of the input as, for example, the minimum variance distortionless response (MVDR) beamformer. For non-Gaussian distributed noise, however, the minimum mean square error (MMSE) optimal spatial filter may be nonlinear.

Potentially, such nonlinear functional relationships could be learned by deep neural networks. However, the performance would depend on many parameters and the architecture of the neural network. Therefore, in this paper, we more generally analyze the potential benefit of nonlinear spatial filters as a function of the multivariate kurtosis of the noise distribution.

The results imply that using a nonlinear spatial filter is only worth the effort if the noise data follows a distribution with a multivariate kurtosis that is considerably higher than for a Gaussian. In this case, we report a performance difference of up to 2.6 dB segmental signal-to-noise ratio (SNR) improvement for artificial stationary noise. We observe an advantage of 1.2 dB for the nonlinear spatial filter over the linear one even for real-world noise data from the CHiME-3 dataset given oracle data for parameter estimation.
Index Terms: Multichannel, speech enhancement, nonlinear filtering, acoustic beamforming, neural networks

1. Introduction

Many speech signals recorded in everyday environments, for example in a restaurant or next to a busy street, are corrupted by additional background noise. Therefore, speech enhancement algorithms that improve the perceived quality or intelligibility of a recorded speech signal by reducing noise or other disturbing effects such as reverberation are of great importance in a wide range of communication applications.

Noise reduction methods such as the Wiener filter [1, Sec. 11.3.1] and nonlinear optimal estimators of the clean speech Fourier coefficient [1, Sec. 11.4] or its magnitude [2] effectively reduce noise in single-channel microphone recordings. However, multichannel approaches often outperform single-channel methods as they incorporate not only tempo-spectral properties of the signals but can also include spatial information in the processing.

In most cases, the spatial filtering is based on a linear processing model, called beamforming, that weights the DFT coefficient of the different microphone channels with complex-valued coefficients before summation to suppress signal components from others than the target direction [3, Sec. 3.1]. The MVDR beamformer is a prominent example of a linear spatial filter that exploits the time delay of signal arrival determined by the spatial arrangement and further takes the correlation of the noise signals between the microphones into account.

It seems natural to include well-developed single-channel methods into multichannel speech enhancement by applying a single-channel algorithm, called a postfilter, to the output of a spatial filter. For Gaussian distributed noise, it has been shown that the sequential coupling of the spatially linear MVDR filter and a postfilter yields optimal results with respect to the MMSE, maximum a posteriori (MAP) and maximum likelihood (ML) criterion [4, 5]. In contrast, Hendriks et al. show that the optimal spatial filter is nonlinear and cannot be separated from spectral processing if the noise is not Gaussian distributed [6]. However, it remains open how large the potential benefit of using nonlinear spatial filters really is. This question gained importance in the context of the rise of neural networks in recent years: while it is demanding to derive optimal nonlinear spatial filters in a statistical framework, neural networks can learn to approximate nonlinear functions directly from data [7].

Neural networks have successfully been incorporated into single-channel speech enhancement [8, 9, 10, 11] often in the context of automatic speech recognition (ASR) [12] and they have also been very successful in estimating the parameters of linear spatial beamformers [13]. Sainath et al. propose a multichannel neural network approach to ASR that includes a spatial filtering layer [14, 15, 16]. Interestingly, the structure of their proposed time-convolutional layer imposes a linearity constraint on the spatial filter even though fixing a linear spatial filter might not lead to an optimal solution.

The goal of our research is to answer the question if investing in the development of neural networks that learn optimal *nonlinear* spatial filters is worth the effort. As a first step towards answering this question, in this paper, we analyze the potential benefit of nonlinear spatial filtering as compared to a standard linear spatial filter like the MVDR under ideal conditions.

In order to gain a better understanding of the role and potential of nonlinear spatial filters, we proceed as follows: Section 3 reviews the most relevant theoretical results on the optimality of linear versus nonlinear spatial filters. In section 4, we analyze the potential performance gain of an optimal nonlinear spatial filter in contrast to a linear spatial filter for noise with a known super-Gaussian distribution. Section 5 assesses the improvement potential of nonlinear spatial filters for real noise recordings from the CHiME-3 dataset [17].

2. Notation and Assumptions

We assume that a microphone array composed of D microphones records the target speech signal along with interfering noise. The time domain signals are windowed and transformed into the frequency domain using the discrete Fourier transform (DFT), which leads to the noisy DFT coefficients $Y_\ell(k,\, i)$ with microphone-channel index $\ell \in \{1,...,D\}$, frequency-bin index k and time-frame index i. We assume an additive noise signal model so that the noisy DFT coefficient $Y_\ell(k,i)$ can be represented as a sum of the clean speech DFT coefficient $S_\ell(k,i)$ and noise DFT coefficient $N_\ell(k,i)$ received at the ℓth microphone, i.e.,

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

http://dx.doi.org/10.21437/Interspeech.2019-2751

The DFT coefficients of the speech and noise signals are modeled as random variables. We denote random variables by uppercase letters, while lowercase letters are used for their respective realizations. The speech and noise coefficients are assumed to be uncorrelated and all DFT coefficients to be zero-mean and independent with respect to time and frequency. As a consequence, we can drop the indices (k,i) from the notation. Let $\mathbf{Y} = [Y_1,...,Y_D] \in \mathbb{C}^D$ be the vector containing the noisy DFT coefficients for all D channels and let $\mathbf{S} \in \mathbb{C}^D$ and $\mathbf{N} \in \mathbb{C}^D$ denote the vectors of speech and noise DFT coefficients, respectively. We work in a single source scenario, which means that there is only one target speaker, and model the signal propagation from the speaker to the microphones as a plane wave. Thus, the vector of clean speech DFT coefficients $\mathbf{S}$ can be obtained by multiplying the reference clean speech DFT coefficient S with a frequency-dependent vector $\mathbf{d} \in \mathbb{C}^D$, i.e., $\mathbf{S} = \mathbf{d}S$. We denote the noise correlation matrix by $\mathbf{\Phi}_n = \mathbb{E}[\mathbf{N}\mathbf{N}^H]$, while $\sigma_s^2 = \mathbb{E}[|S|^2]$ denotes the spectral power of S.

3. Linearity of the Optimal Spatial Filter

In this section, we review optimal multichannel estimators of the clean speech DFT coefficient to address the question under which conditions an optimal solution decomposes into a linear spatial filter and a spectral postfilter. First, we consider the case of multivariate complex Gaussian distributed noise DFT coefficients with zero mean and covariance matrix $\mathbf{\Phi}_n$. Since we assume that the noise is additive, the distribution of $\mathbf{Y}$ given the reference speech DFT coefficient s is Gaussian distributed with mean $\mathbf{d}s$ and covariance matrix $\mathbf{\Phi}_n$, i.e., $\mathbf{Y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{d}s, \mathbf{\Phi}_n)$. The corresponding conditional probability density function (PDF) is given by [18, Thm. 15.1]

$$p(\mathbf{y}|s) = \frac{1}{\pi^D |\mathbf{\Phi}_n|} \exp\Big\{ -(\mathbf{y} - \mathbf{d}s)^H \mathbf{\Phi}_n^{-1} (\mathbf{y} - \mathbf{d}s) \Big\}. \quad (2)$$

Balan and Rosca [4] use the concept of sufficient statistics to show that the MMSE estimator of the clean speech DFT coefficient S

$$T_{\text{MMSE}}(\mathbf{y}) = \arg\min_{\hat{s} \in \mathbb{C}} \mathbb{E}\big[|S - \hat{s}|^2 \, |\mathbf{y} \big] \quad (3)$$

separates into the well-known MVDR beamformer and a spectral postfilter under a Gaussian noise assumption. The MVDR beamformer T_{MVDR} is a sufficient statistic *in the classical sense* for the true clean speech DFT coefficient s if the conditional distribution of the noisy observation $\mathbf{Y}$ given $T_{\text{MVDR}}(\mathbf{y})$ does not depend on s [19, Def. IV.C.1], i.e. $T_{\text{MVDR}}(\mathbf{Y})$ contains all the information in $\mathbf{Y}$ that is useful for estimating s. Furthermore, T_{MVDR} is said to be a sufficient statistic of S *in the Bayesian sense* if

$$p(s|\mathbf{y}) = p(s|T_{\text{MVDR}}(\mathbf{y})) \quad (4)$$

holds for all observations $\mathbf{y}$ regardless of the prior distribution of S [20, Def. 2.4]. If the MVDR beamformer is a sufficient statistic, then no information about S is lost during spatial filtering even though the dimension of the output is reduced to one dimension. As a result, spatial processing and spectral processing can be performed separately in sequence. Since a statistic that is sufficient in the classical sense is also sufficient in the Bayesian sense [20, Thm. 2.14.2], we can infer that (4) holds by showing that the MVDR beamformer is a sufficient statistic in the classical sense. Resorting to the Fisher-Neyman factorization theorem [19, Prop. IV.C.1][21, Cor. 2.6.1], we deduce this property of the MVDR beamformer from the finding that the conditional PDF $p(\mathbf{y}|s)$ can be factorized as

$$\begin{aligned} p(\mathbf{y}|s) &= \underbrace{\frac{1}{\pi^D |\mathbf{\Phi}_n|} \exp\{-\mathbf{y}^H \mathbf{\Phi}_n^{-1} \mathbf{y}\}}_{h(\mathbf{y})} \\ &\quad \times \underbrace{\exp\Big\{ \mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{d} \big(2\text{Re}\{ s^* T_{\text{MVDR}}(\mathbf{y}) \} - |s|^2 \big) \Big\}}_{g(s, T_{\text{MVDR}}(\mathbf{y}))} \\ &= h(\mathbf{y}) g(s, T_{\text{MVDR}}(\mathbf{y})) \quad (5) \end{aligned}$$

with

$$T_{\text{MVDR}}(\mathbf{y}) = \frac{\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{y}}{\mathbf{d}^H \mathbf{\Phi}_n^{-1} \mathbf{d}}. \quad (6)$$

Using the fact that the MMSE estimator complies with the mean of the posterior [19, IV.B.1], we infer from (4) that

$$\begin{aligned} T_{\text{MMSE}}(\mathbf{y}) &= \mathbb{E}[S|\mathbf{y}] \quad (7) \\ &= \mathbb{E}[S|T_{\text{MVDR}}(\mathbf{y})] \quad (8) \end{aligned}$$

holds. The quantity $\mathbb{E}[S|T_{\text{MVDR}}(\mathbf{y})]$ can be seen as a single-channel filter working on the output of the MVDR beamformer. Because the relationship (4) holds for any prior distribution of S, a decomposition of the MMSE estimator into an MVDR beamformer and single-channel postfilter results independent of any further assumptions about the prior distribution of the reference speech DFT coefficient. The decomposition of the MMSE estimator is also described by Hendriks et al. [6] but derived without the concept of sufficient statistics.

From (4) we conclude that the MAP estimator also separates into a linear spatial filter and a single-channel postfilter. Furthermore, the MVDR beamformer can be identified as the ML estimator of the clean speech DFT coefficient S [5, Sec. 6.2.1.2].

However, the work of Hendriks et al. [6] reveals that the Gaussian noise assumption is fundamental to both the decomposability of the optimal estimator into a spatial and a spectral processing step and the linearity of the spatial filter. They derive an MMSE estimator for noise that follows a multivariate Gaussian mixture distribution. The M Gaussian mixture components are modeled as zero-mean with covariance matrices $\mathbf{\Phi}_m$, $m = 1,...,M$, and combined by positive weighting factors c_m that fulfill the constraint $\sum_{m=1}^M c_m = 1$. The resulting conditional PDF of $\mathbf{Y}$ is given by [22, Sec. 9.2]

$$p(\mathbf{y}|s) = \sum_{m=1}^{M} \frac{c_m}{\pi^D |\mathbf{\Phi}_m|} \exp\Big\{ -(\mathbf{y} - \mathbf{d}s)^H \mathbf{\Phi}_m^{-1} (\mathbf{y} - \mathbf{d}s) \Big\}. \quad (9)$$

Hendriks et al. assume the clean speech amplitude A to be distributed according to the PDF

$$p(a) = \frac{2 \left(\frac{\nu}{\sigma_s^2} \right)^{\nu}}{\Gamma(\nu)} a^{2\nu - 1} \exp\left\{ -\frac{\nu}{\sigma_s^2} a^2 \right\} \text{ with } \nu > 0, a \leq 0 \quad (10)$$

and the phase $\Psi \in [0, 2\pi)$ to be uniformly distributed and independent of the speech amplitude. Then the MMSE estimator is given by

$$\widetilde{T}_{\text{MMSE}}(\mathbf{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_m|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y} \right]} \frac{\sigma_s^2 T_{\text{MVDR}}^{(m)}(\mathbf{y}) \mathcal{M}(\nu+1, 2, P_m)}{\nu (\mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d})^{-1} + \sigma_s^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{|\mathbf{\Phi}_m|} e^{\left[-\mathbf{y}^H \mathbf{\Phi}_m^{-1} \mathbf{y} \right]} \mathcal{M}(\nu, 1, P_m)} \quad (11)$$

with

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

with $\mathcal{M}(\cdot,\cdot,\cdot)$ being the confluent hypergeometric function [23, Sec. 9.21]. Interestingly, the result shows that the MMSE estimator for the considered non-Gaussian model cannot be separated into an MVDR beamformer and a single-channel postfilter. Furthermore, the optimal spatial filter is not even linear [6].

4. Potential of Nonlinear Spatial Filters

In this section, we investigate the improvement potential of using the optimal spatially nonlinear MMSE estimator for Gaussian mixture distributed noise as opposed to a setup that combines a linear spatial filter and a spectral postfilter. To our knowledge, the MMSE estimator for non-Gaussian noise derived by Hendriks et al. has not been evaluated before.

We use a segment length of 32 ms and a square-root Hann window with 50% overlap for spectral analysis and synthesis. The clean speech signals have been taken from the WSJ0 dataset [24] and are balanced between male and female speakers (30 utterances each).

The noise DFT coefficients are generated by sampling a zero-mean Gaussian mixture distribution. The covariance matrix $\boldsymbol{\Phi}_n$ of the distribution is chosen to represent one of three scenarios [25]: spatially white noise, diffuse noise, and a directional noise source positioned at a 45 degree angle to the target source. In the latter cases, we add a small portion of spatially white noise ($\alpha_{\text{wn}} = 0.05$) to ensure that the noise correlation matrix is invertible. We obtain noise distributions that depart from normality by means of heavier tails by combining mixture components with scaled versions of the same covariance matrix. Thus, we set the mth mixture component's covariance matrix $\boldsymbol{\Phi}_m$ to be

$$\boldsymbol{\Phi}_m = \frac{b^{m-1}}{r} \boldsymbol{\Phi}_n \quad \text{with} \quad r = \sum_{m=1}^{M} c_m b^{m-1} \tag{12}$$

and scaling factor $b \in \mathbb{R}^+$. The constant r takes care of normalization so that the covariance matrix $\boldsymbol{\Phi}_n$ of the mixture distribution remains unchanged.

The kurtosis is a statistical measure that accounts for the shape of a distribution, specifically its heavy-tailedness [26, 27]. We extend Mardia's multivariate kurtosis definition [28] to complex-valued random vectors $\mathbf{X} \in \mathbb{C}^D$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_x$ to obtain

$$\kappa_{\mathbb{C}}(\mathbf{X}) = \mathbb{E}\left[\left(2 (\mathbf{X} - \boldsymbol{\mu})^H \mathbf{C}_x^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right)^2 \right]. \tag{13}$$

Using [29, Sec. 8.2.4], we find the multivariate complex kurtosis of the random vector $\mathbf{N}$ following a scaled Gaussian mixture distribution to be

$$\kappa_{\mathbb{C}}(\mathbf{N}) = 2D(2+2D) \underbrace{\sum_{m=1}^{M} c_m \frac{b^{2(m-1)}}{r^2}}_{q}. \tag{14}$$

The factor $2D(2 + 2D)$ corresponds to the kurtosis of the D-dimensional complex Gaussian distribution. Thus, the kurtosis of the scaled Gaussian mixture distribution equals the kurtosis of a Gaussian distribution multiplied by a factor that we name q. We see that the multivariate kurtosis depends on the dimensionality of the distribution and that the scaling factor b and the number components allow us to adjust the degree of heavy-tailedness of the noise distribution.

We use the MVDR beamformer as a linear spatial filter for the comparison setup because it is optimal with respect to the maximum likelihood criterion if the noise follows a scaled Gaussian mixture distribution as given in (12). This property can be deduced from the fact that the MVDR beamformer is the ML estimator for each Gaussian mixture component and that the MVDR beamformer is invariant against scaling of the noise correlation matrix. We then combine the MVDR beamformer with an MMSE optimal single-channel postfilter.

Since the input vector given the reference speech DFT coefficient s follows a multivariate complex Gaussian mixture distribution, i.e., $\mathbf{Y} \sim \sum_{m=1}^{M} c_m \mathcal{N}_{\mathbb{C}}(\mathbf{d}s, \boldsymbol{\Phi}_m)$, the output of the MVDR beamformer is distributed according to a one-dimensional complex Gaussian mixture distribution. More precisely, it is

$$p(T_{\text{MVDR}}(\mathbf{y}) | s) = \sum_{m=1} c_m \mathcal{N}_{\mathbb{C}} \Bigg(s, \underbrace{\frac{\mathbf{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{\Phi}_m \boldsymbol{\Phi}_n^{-1} \mathbf{d}}{(\mathbf{d}^H \boldsymbol{\Phi}_n^{-1} \mathbf{d})^2}}_{\sigma_m^2} \Bigg). \tag{15}$$

We adhere to the assumptions regarding speech phase and amplitude that Hendriks et al. introduced in [6] to compute the spatially nonlinear MMSE estimator and derive the postfilter using [23, Eq. 3.339, Eq. 6.643.2, Eq. 9.220.2] and [30, Eq. 10.32.3] in an analog way. We find the estimator $T_{\text{MVDR-MMSE}}$ that combines the MVDR beamformer with the MMSE postfilter to be

$$T_{\text{MVDR-MMSE}}(\mathbf{y}) = \nu \frac{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2} \right]} \frac{\sigma_s^2 T_{\text{MVDR}}(\mathbf{y}) \mathcal{M}(\nu+1,2,P_m)}{\nu \sigma_m^2 + \sigma_s^2}}{\sum_{m=1}^{M} \frac{c_m Q_m}{\sigma_m^2} e^{\left[-\frac{|T_{\text{MVDR}}(\mathbf{y})|^2}{\sigma_m^2} \right]} \mathcal{M}(\nu,1,P_m)} \tag{16}$$

with

$$\boldsymbol{\Phi}_n = \sum_{m=1}^{M} c_m \boldsymbol{\Phi}_m, \quad \sigma_m^2 = \frac{\mathbf{d}^H \boldsymbol{\Phi}_n^{-1} \boldsymbol{\Phi}_m \boldsymbol{\Phi}_n^{-1} \mathbf{d}}{(\mathbf{d}^H \boldsymbol{\Phi}_n^{-1} \mathbf{d})^2},$$
$$Q_m = \left(\frac{1}{\sigma_m^2} + \frac{\nu}{\sigma_s^2} \right)^{-\nu} \quad \text{and} \quad P_m = \frac{\sigma_s^2 \sigma_m^{-2} |T_{\text{MVDR}}(\mathbf{y})|^2}{\nu \sigma_m^2 + \sigma_s^2}.$$

Both the spatially nonlinear MMSE estimator and the MMSE postfilter require an estimate of the spectral power of the speech signal σ_s^2. We estimate the parameter for a given time frame by time-averaging over five successive segments of the clean speech data. The speech parameter ν in (10) is set to 0.25 for the nonlinear MMSE estimator and to 0.5 for the postfilter of the $T_{\text{MVDR-MMSE}}$ estimator because this gives the best results for scaled Gaussian mixture noise distributions with higher kurtosis values.

We model the microphone array as a linear array with five microphones at a distance of 5 cm and generate the vector of speech DFT coefficients $\mathbf{S}$ for a source that is located in endfire position. The noise and speech DFT coefficients are combined to give an SNR of 0 dB.

The left column of Figure 1 shows the segmental SNR improvement of the MVDR beamformer T_{MVDR}, the spatially nonlinear MMSE estimator $\widetilde{T}_{\text{MMSE}}$ derived by Hendriks et al. [6], and the MVDR beamformer combined with an MMSE postfilter $T_{\text{MVDR-MMSE}}$ with respect to the kurtosis factor q defined in (14). We compute the segmental SNR using a segment length of 10 ms as described in [31]. To measure the improvement of the segmental SNR, we compare the mean segmental SNR of

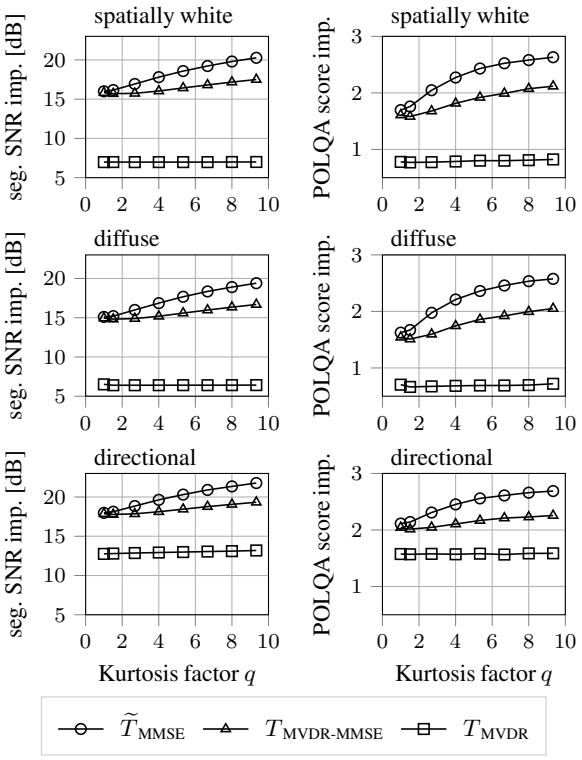

Figure 1: *Segmental SNR and POLQA improvement for noise distributions with increasing kurtosis in three noise scenarios (spatially white, diffuse and directional).*

the noisy microphone recordings to the segmental SNR of the enhanced speech signal. The gap between the top curves (circle and triangle) quantifies the advantage of the nonlinear spatial filter over the linear spatial filter. The difference amounts to values in the order of 2.6 dB for noise that obeys a significantly more heavy-tailed distribution than a Gaussian.

The right column of Figure 1 depicts the perceptual objective listening quality analysis (POLQA) score [32] improvement achieved by the three processing methods. POLQA is the successor of the perceptual evaluation of speech quality (PESQ) measure [33] and returns the expected mean opinion score (MOS) [34] that ranges from one (bad) to five (excellent). As for the segmental SNR improvement, there is a measurable performance difference ($\sim$ 0.5 POLQA score improvement) between the spatially linear and nonlinear estimator. We conclude that the use of a nonlinear spatial filter could be worthwhile if real noise follows a distribution that is considerably more heavy-tailed than a Gaussian distribution.

5. Evaluation on Real-World Noise Data

Using the same estimators as in the previous section, we aim to assess if performance improvements obtained by a nonlinear spatial filter also hold for real-world noise recordings, as provided by the CHiME-3 dataset [17]. We use the five recordings that correspond to the front-facing microphones placed in a frame around a tablet computer that has been used to record noise in four different locations: a bus, a cafeteria, a pedestrian area, and a busy street. We place the target source in the same plane as the tablet, perpendicular to the upper edge, and combine the speech noise signals to obtain an SNR of 0 dB.

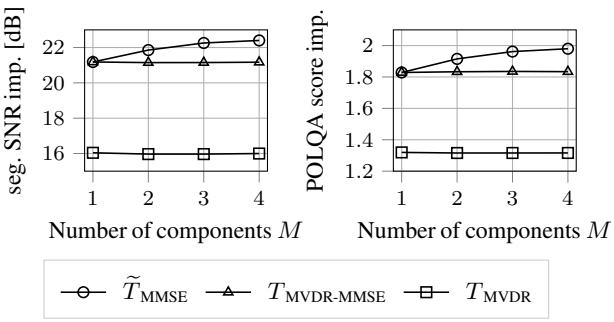

Figure 2: *Segmental SNR and POLQA improvement for CHiME-3 noise recordings from four locations (bus, cafeteria, pedestrian area, street) with respect to the number of mixture components used to fit the noise distribution.*

The estimators $\widetilde{T}_{\text{MMSE}}$ and $T_{\text{MVDR-MMSE}}$ require the parameters of a zero-mean Gaussian mixture distribution to be estimated from data. We obtain time-variant estimates of the component covariance matrices with the expectation maximization (EM) algorithm [22] applied to signal segments of length 750 ms that overlap by 50% and set the speech parameter $\nu = 0.25$ for both estimators as this gave the best results for the CHiME-3 data.

The left side of Figure 2 depicts the segmental SNR improvement results with respect to the number of components M in the mixture distributions that have been fitted to the data. We find that the use of a postfilter significantly increases the segmental SNR improvement (the difference of 5 dB between the results of T_{MVDR} and $T_{\text{MVDR-MMSE}}$), but the postfilter following the linear spatial filter in $T_{\text{MVDR-MMSE}}$ delivers a very similar performance regardless of the number of components of the distribution model. In contrast, we observe that the $\widetilde{T}_{\text{MMSE}}$ estimator with a nonlinear spatial filter achieves better results when we model the distribution through a Gaussian mixture with more components. The performance difference between $\widetilde{T}_{\text{MMSE}}$ and $T_{\text{MVDR-MMSE}}$ that we attribute to the usage of a nonlinear spatial filter amounts to 1.2 dB averaged over all locations. We make similar observations for the individual locations.

The right plot of Figure 2 shows the improvement with respect to the POLQA measure. The results obtained with the perceptively motivated POLQA measure exhibit the same structure as the results obtained with the segmental SNR and, thus, we find that using a nonlinear spatial filter instead of a linear spatial filter increases the speech quality predicted by POLQA for real-world noise data.

6. Conclusions

In this paper, we showed that using the MMSE optimal nonlinear spatial filter instead of a classical concatenation of a linear spatial filter and a postfilter may yield a performance gain of up to 2.6 dB segmental SNR improvement if the noise follows a distribution with considerably higher multivariate kurtosis than a Gaussian distribution. Also for the real-world noise recordings from the CHiME-3 dataset still moderate improvements of 1.2 dB are achieved when the parameters are estimated on oracle speech and noise data. Future work will analyze the achievable benefit when the filter parameters are estimated blindly from noisy data.

7. Acknowledgment

We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

8. References

[1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Chichester, England Hoboken, NJ: John Wiley, 2006.

[2] 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.

[3] J. Benesty, *Microphone array signal processing*. Berlin: Springer, 2008.

[4] 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.

[5] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[6] 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.

[7] Y. B. Ian Goodfellow and A. Courville, *Deep Learning*. MIT Press, 2016.

[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] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, Lyon, France, Aug. 2013, pp. 436–440.

[10] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

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

[12] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic, Aug. 2015.

[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] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and Andrew, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, USA, Dec. 2015, pp. 30–36.

[15] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5075–5079.

[16] 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.

[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)*, Scottsdale, USA, Dec. 2015, pp. 504–511.

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

[19] H. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY: Springer New York, 1994.

[20] M. Schervish, *Theory of Statistics*. New York, NY: Springer New York, 1995.

[21] E. L. Lehmann, *Testing statistical hypotheses*. New York: Springer, 2005.

[22] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[23] I. S. Gradshteyn, *Table of integrals, series, and products*. San Diego: Academic Press, 2000.

[24] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," Linguistic Data Consortium, Philadelphia, May 2007.

[25] 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.

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

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

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

[29] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," https://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf, November 2012.

[30] "NIST Digital Library of Mathematical Functions," Release 1.0.20 of 2018-09-15, f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds. [Online]. Available: http://dlmf.nist.gov/

[31] 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.

[32] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[33] "P.862.3 : Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," International Telecommunication Union, Nov. 2007, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3-200711-I/en

[34] "P.800 : Methods for subjective determination of transmission quality," International Telecommunication Union, Aug. 1996, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en