Insights into the Interplay of Sampling Rate Offsets and MVDR Beamforming

Joerg Schmalenstroeer, Reinhold Haeb-Umbach

Department of Communications Engineering, Paderborn University, Germany
Email: {schmalen, haeb}@nt.uni-paderborn.de

Abstract

It has been experimentally verified that sampling rate offsets (SROs) between the input channels of an acoustic beamformer have a detrimental effect on the achievable SNR gains. In this paper we derive an analytic model to study the impact of SRO on the estimation of the spatial noise covariance matrix used in MVDR beamforming. It is shown that a perfect compensation of the SRO is impossible if the noise covariance matrix is estimated by time averaging, even if the SRO is perfectly known. The SRO should therefore be compensated for prior to beamformer coefficient estimation. We present a novel scheme where SRO compensation and beamforming closely interact, saving some computational effort compared to separate SRO adjustment followed by acoustic beamforming.

1 Introduction

In multi-channel acoustic signal processing it is usually assumed that the different channels are sampled synchronously, using the same sampling clock for all channels. While this holds true for microphone arrays which have a compact form factor, the assumption is often violated in case of distributed acoustic sensor nodes. A prominent example are Wireless Acoustic Sensor Networks (WASNs), which consist of spatially distributed devices equipped with sensors (microphones or microphone arrays), processing units and communication interfaces [1]. The spatial distribution of the sensors offers great opportunities for improved signal acquisition and enhancement. However, again due to their spatial distribution, the microphones will no longer share the same sampling clock signal. Each sensor node usually has its own crystal oscillator driving the sampling process. Even if these oscillators run on the same nominal sampling frequency, there will inevitably be frequency and phase differences between the sampled signals at the different devices. These are due to imperfect hardware and environmental influences, such as temperature, see, e.g., [2] for a discussion of the hardware issues involved.

If those signals recorded at the sensor nodes are combined for multi-channel signal processing, such as acoustic beamforming, the sampling rate offsets (SROs) lead to significant degradation of the achievable performance. Several studies in the literature report a significant drop in SNR gain from beamforming in the presence of uncompensated SRO, see, e.g., [3–5]. The negative effect of a SRO on adaptive echo cancellation is discussed in [6] and on blind source separation in [7].

While these effects have been observed in experiments, existing literature is usually not concerned with the development of an analytic model explaining the effect of SRO on, say, adaptive beamforming. The works concentrate instead on how to blindly estimate a SRO and how to realign the signals once an estimate of the SRO has been found: In [8] a recursive band-limited interpolation is proposed for SRO estimation, while a correlation maximization approach is proposed in [9]. Others examine the coherence function between pairs of input channels [3, 10].

Experiments on beamforming and SRO estimation were published by Markovich et al. in [4] and by Cherkassky et al. in [5], where for example the performance of an Minimum Variance Distortionless Response (MVDR) beamformer in WASNs was examined. To this end the authors modeled the non-zero SRO in the frequency domain by a multiplication with a phase term, according to the ideas in [11].

However, estimating the SRO is out of the scope of this paper. We rather assume that this task has already been solved and an estimate of the SRO is available. Instead of relying on experimental evidence, the purpose of this paper is to offer an analytic model to explain the impact of a SRO on adaptive beamforming. To be specific, we consider a scenario with a desired and and undesired signal, both modeled as a point source in a reverberant environment, and compute the impact of the SRO on the estimation of the spatial noise covariance matrix. This matrix is needed in many multi-channel signal processing tasks, such as, e.g., for the computation of the beamformer coefficients of a MVDR beamformer. This analysis shows that the effect of the SRO on a covariance matrix estimated by time averaging cannot be undone perfectly, even if the SRO were known. We conclude that the SRO compensation has to be done prior to the beamformer coefficient computation and the beamforming operation. For this we offer an implementation where SRO compensation and beamforming are closely coupled, saving some computation compared to a separate SRO compensation by resampling followed by acoustic beamforming.

The paper is organized as follows. The basic signal model is described in Sec. 2, followed by a short introduction to MVDR beamforming in Sec. 3. Sec. 4 presents a detailed analysis of the spatial noise covariance matrix estimation in the presence of a SRO. The new SRO aware MVDR beamformer is introduced in Sec. 5. We continue with some experiments in Sec. 6 and summarize our findings in Sec. 7.

2 Signal model

Let $x_i(t)$ denote the microphone signal at node i. It is sampled with the sampling frequency f_i. If there is a SRO between the oscillators at nodes i and j, the sampling frequency at node j can be expressed by

$$f_j = (1+\epsilon_{ij}) \cdot f_i, \tag{1}$$

where $\epsilon_{ij} \ll 1$ denotes the SRO between nodes i and j.

The N-point Short Time Fourier Transform (STFT) $X_i(k,l)$ of the l-th block (with block shift B) using a periodic Hann window $w(n)$ is given by

$$X_i(k,l) = \sum_{n=0}^{N-1} w(n) \cdot x_i(n+l\cdot B) \cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}kn}, \tag{2}$$

where k denotes the frequency bin and the $x_i(n)$ are the time domain samples of $x_i(t)$, sampled with frequency f_i.

We assume a single target speech source signal $s(t)$ and a single coherent noise source $v(t)$ at fixed positions in the room. The corresponding signals at node i can be expressed in the STFT domain as

$$X_i(k,l) = H_i(k,l)\cdot S_i(k,l) + \underbrace{G_i(k,l)\cdot V_i(k,l)}_{=N_i(k,l)}. \tag{3}$$

where $H(k,l)$ and $G(k,l)$ denote the corresponding acoustic transfer functions. Note that all terms have an index i to indicate that they are sampled with node i's sampling frequency f_i. In contrast to the model in [10] we do not include spatially uncorrelated noise in the model to simplify the derivation, although it will be included lateron in the simulations.

Now consider another sensor node j observing the same signals, however under the sampling frequency f_j. Following [4] and [11] the STFTs of the speech components are linked via a phase term, which is due to the difference in sampling phase and frequency:

$$S_i(k,l) \approx S_j(k,l)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}[\tau_{ij}+(\frac{N}{2}+lB)\epsilon_{ij}]k}, \tag{4}$$

where the initial sampling phase offset is caused by τ_{ij}, the difference in the recording start times at nodes i and j.

The SRO introduces a delay between the two streams which increases/decreases over time and which is given by $(N/2 + lB)\epsilon_{ij}$, according to (4). To shorten the notation we denote the complex exponential caused by sampling phase and frequency offset by

$$\xi_{ij}(k,l) := \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}[\tau_{ij}+(\frac{N}{2}+lB)\epsilon_{ij}]k}. \tag{5}$$

Correspondingly, the noise source signal part can be described by

$$V_i(k,l) \approx V_j(k,l) \cdot \xi_{ij}(k,l). \tag{6}$$

To summarize, a constant SRO leads to a linearly increasing or decreasing time shift, which corresponds to a multiplication with a complex exponential term in the STFT domain. This model has been successfully used in a variety of publications, e.g., [3,9,11]. However, the model validity strongly depends on the SRO and the STFT size, as has been discussed in [12].

3 MVDR beamforming

To study the influence of the SRO on beamforming we select the widely used MVDR beamformer in its narrow-band formulation. In the frequency domain the MVDR beamformer filter coefficients are given by (see [13])

$$\boldsymbol{W}(k,l) = \frac{\boldsymbol{R}^{-1}(k,l)\cdot\boldsymbol{d}(k,\boldsymbol{p})}{\boldsymbol{d}^H(k,\boldsymbol{p})\cdot\boldsymbol{R}^{-1}(k,l)\cdot\boldsymbol{d}(k,\boldsymbol{p})}. \tag{7}$$

Here, $\boldsymbol{d}(k,\boldsymbol{p})$ denotes the steering vector, which we model in this work to consist of phase-only terms, assuming anechoic signal transmission, to point the beamformer towards the desired speaker at position $\boldsymbol{p}$. Further, $\boldsymbol{R}(k,l)$ denotes the spatial correlation matrix of the noise.

Applying the beamformer to the input signals results in the output signal $Y(k,l)$ with

$$Y(k,l) = \boldsymbol{W}^H(k,l)\cdot\boldsymbol{X}(k,l), \tag{8}$$

which is subsequently transformed back to the time domain to synthesize the output signal.

Here we assume that speaker and node positions are known, and thus that the steering vector $\boldsymbol{d}(k,\boldsymbol{p}_s)$ is known. However, the noise correlation matrix $\boldsymbol{R}(k,l)$ will be estimated from the sensor nodes' signals. The impact of SROs between the nodes on this estimation will be discussed next.

4 Noise correlation matrix estimation

At times where the desired source signal is absent, the noise correlation matrix can be estimated from the recorded signals $X_i(k,l) = N_i(k,l)$, $i = 1,\ldots,K$, where K is the number of sensor nodes. These signals are gathered in the vector $\boldsymbol{N}(k,l) = [N_1(k,l),\ldots,N_K(k,l)]^T$.

For illustration purposes consider a sensor network with $K = 3$ nodes, which are denoted by h,i and j. From the instantaneous observation $\boldsymbol{N}(k,l) = [N_h(k,l), N_i(k,l), N_j(k,l)]^T$ we can compute the dyade

$$\begin{aligned}
\widetilde{\boldsymbol{R}}(k,l) &= \boldsymbol{N}(k,l)\cdot\boldsymbol{N}^H(k,l)\\
&= |V_j(k,l)|^2 \begin{bmatrix} G_h(k,l)\xi_{hj}(k,l)\\ G_i(k,l)\xi_{ij}(k,l)\\ G_j(k,l)\end{bmatrix}\begin{bmatrix} G_h(k,l)\xi_{hj}(k,l)\\ G_i(k,l)\xi_{ij}(k,l)\\ G_j(k,l)\end{bmatrix}^H\\
&= |V_j(k,l)|^2 \begin{bmatrix} |G_h|^2 & G_hG_i^*\xi_{hj}\xi_{ij}^* & G_hG_j^*\xi_{hj}\\ G_iG_h^*\xi_{hj}^*\xi_{ij} & |G_i|^2 & G_iG_j^*\xi_{ij}\\ G_jG_h^*\xi_{hj}^* & G_jG_i^*\xi_{ij}^* & |G_j|^2\end{bmatrix}_{k,l},
\end{aligned} \tag{9}$$

where the notation $[...]_{k,l}$ means that all terms within the matrix depend on k and l. The phase terms can be summarized which we demonstrate for the product of ξ_{hj} and ξ_{ij}^*:

$$\begin{aligned}
\xi_{hj}\xi_{ij}^* &= \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}[\tau_{hj}+(\frac{N}{2}+lB)\epsilon_{hi}]k}\mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}[\tau_{ij}+(\frac{N}{2}+lB)\epsilon_{ij}]k} &(10)\\
&= \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}[\tau_{hj}-\tau_{ij}+(\frac{N}{2}+lB)(\epsilon_{hj}-\epsilon_{ij})]k}\\
&\approx \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}[\tau_{hi}+(\frac{N}{2}+lB)(\epsilon_{hi})]k} = \xi_{hi}. &(11)
\end{aligned}$$

Here we used that $\tau_{hj} - \tau_{ij} = \tau_{hi}$ and $\epsilon_{hj} - \epsilon_{ij} \approx \epsilon_{hi}$. The latter can be seen as follows

$$\begin{aligned}
\epsilon_{hj} - \epsilon_{ij} &= (\frac{f_j}{f_h} - 1) - (\frac{f_j}{f_i} - 1) = \frac{f_j}{f_h} - \frac{f_j}{f_i}\\
&= \left(\frac{f_i}{f_h} - 1\right)\frac{f_j}{f_i} = \epsilon_{hi}\frac{f_j}{f_i} \approx \epsilon_{hi}
\end{aligned} \tag{12}$$

because $f_j/f_i \approx 1$ since $\epsilon_{ij} \ll 1$.

Thus (9) can be rewritten as

$$\begin{aligned}
&\widetilde{\boldsymbol{R}}(k,l) = \boldsymbol{N}(k,l)\cdot\boldsymbol{N}^H(k,l) =\\
&= |V_j(k,l)|^2 \cdot \begin{bmatrix} |G_h|^2 & G_hG_i^*\xi_{hi} & G_hG_j^*\xi_{hj}\\ G_iG_h^*\xi_{hi}^* & |G_i|^2 & G_iG_j^*\xi_{ij}\\ G_jG_h^*\xi_{hj}^* & G_jG_i^*\xi_{ij}^* & |G_j|^2\end{bmatrix}_{k,l} &(13)\\
&= |V_j(k,l)|^2 \cdot \boldsymbol{D}(k,l)\begin{bmatrix} |G_h|^2 & G_hG_i^* & G_hG_j^*\\ G_iG_h^* & |G_i|^2 & G_iG_j^*\\ G_jG_h^* & G_jG_i^* & |G_j|^2\end{bmatrix}_{k,l}\boldsymbol{D}^H(k,l).
\end{aligned}$$

The matrix $\boldsymbol{D}(k,l)$ reflects the SRO influence on the result with node j acting as a reference node with

$$\boldsymbol{D}(k,l) = \begin{bmatrix}\xi_{hj}(k,l) & 0 & 0\\ 0 & \xi_{ij}(k,l) & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{14}$$

Since $\boldsymbol{D}(k,l)$ is an unitary matrix the inverse is given by its Hermitian transpose and we can compute a SRO compensated dyade by multiplying from the left and right side with $\boldsymbol{D}^H(k,l)$ and $\boldsymbol{D}(k,l)$, respectively. Thus the SRO compensated instantaneous estimate of $\boldsymbol{R}$ at the l-th block and k-th bin is given by:

$$\widehat{\boldsymbol{R}}(k,l) = \boldsymbol{D}^H(k,l)\cdot\boldsymbol{N}(k,l)\cdot\boldsymbol{N}^H(k,l)\cdot\boldsymbol{D}(k,l). \tag{15}$$

The estimates of the noise correlation matrix, $\widetilde{\boldsymbol{R}}(k,l)$ or $\widehat{\boldsymbol{R}}(k,l)$, based on a single time frame are, however, much too noisy. Furthermore, they are of rank one and thus cannot be inverted, as is required in (7). Both issues can be solved by averaging across L frames (if L is chosen large enough):

$$\overline{\boldsymbol{R}}(k) = \frac{1}{L}\sum_{l=0}^{L-1}\boldsymbol{N}(k,l)\cdot\boldsymbol{N}^H(k,l). \tag{16}$$

Here, we considered the uncompensated instantaneous estimate $\widetilde{\boldsymbol{R}}(k,l)$ to showcase the influence of the SRO on the correlation matrix entries. To study the statistical properties of $\overline{\boldsymbol{R}}(k)$ we compute the expectation of the (h,i)-th entry

$$\begin{aligned}
\mathbb{E}\left[\overline{R}_{hi}(k)\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{l=0}^{L-1}G_h(k,l)V_h(k,l)G_i^*(k,l)V_i^*(k,l)\right]\\
&= \frac{1}{L}\sum_{l=0}^{L-1}\underbrace{\mathbb{E}\left[|V_h(k,l)|^2G_h(k,l)G_i^*(k,l)\right]}_{\approx R_{hi}(k)}\xi_{hi}\\
&= R_{hi}(k)\cdot\frac{1}{L}\sum_{l=0}^{L-1}\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}(\frac{N}{2}+lB)\epsilon_{hi}k} &(17)\\
&= R_{hi}(k)\cdot\frac{1}{L}\cdot\mathrm{e}^{-\mathrm{j}\pi\epsilon_{hi}k}\cdot\frac{1-\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}B\epsilon_{hi}kL}}{1-\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}B\epsilon_{hi}k}} &(18)
\end{aligned}$$

where $R_{hi}(k)$ is the correlation in absence of a SRO, and in (17) we assumed $\tau_{hi} := 0$ (zero initial phase offset).

Obviously, the SRO introduces a bias in the estimate of $R_{hi}(k)$. If ϵ_{hi} is known, the bias can be computed. Still the bias cannot be removed perfectly because the sum of the harmonic exponentials can become zero. This will be studied in more detail in the experimental section.

5 Leaping MVDR

In our model of the MVDR beamformer given in Section 3 the steering vector consists of complex exponentials representing signal delays. However, delays larger than one sample will create cyclic wrap around effects, if the delay is compensated by a multiplication with a phase term in the STFT domain, and therefore cause signal artifacts. Additionally, the SRO compensation of the microphone signals requires also signal delay compensations following (15). In this section we show how these two delay terms can be jointly and efficiently handled.

Our idea is here to combine the SRO compensation and the MVDR beamformer delays into an integrated processing block, where the block shift of the analysis window is selected at each processing time step individually, fitting the requirements of the SRO compensation and the MVDR beamformer. Since this individual block shift is not constant over all blocks we call this beamformer "Leaping Minimum Variance Distortionless Response (L-MVDR)".

Figure 1 depicts the block diagram of the L-MVDR. At first each input channel is pre-processed by a STFT based SRO compensation as explained in [12]. Here the h-th channel is shown which employs the following input signals/data: Audio signal $x_h(n)$, block index counter l, SRO ϵ_{hj} between h-th and reference channel j, and optimal Time Differences of Arrival (TDOA)

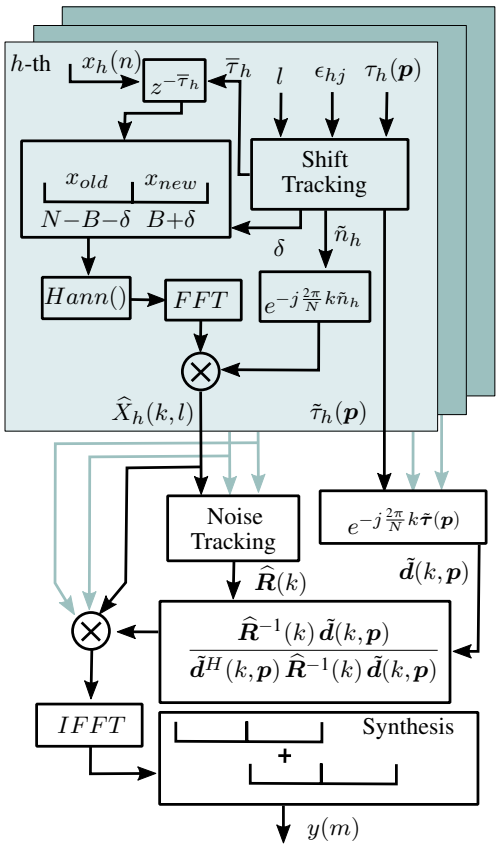

Figure 1: Block diagram of L-MVDR, shown is the h-th channel pre-processing for synchronization (gray box) and below the multi-channel MVDR filter.

$\tau_h(\boldsymbol{p})$, which is a function of the positions of the sensor node h and the speaker. The purpose of the shift tracking block is described in Sec. 5.1 further below. After this preprocessing the noise correlation matrix is estimated (block "noise tracking"), the beamformer coefficients are computed, and the input signal is filtered. Then the time domain signal $y(m)$ is reconstructed by an overlap-add operation. Note that $y(m)$ is a digital signal sampled at the same sampling rate as the reference channel j.

5.1 Delay compensation

The MVDR steering vector $\boldsymbol{d}(k,\boldsymbol{p})$, which consists of delays, accounts for TDOA of the target source signal at the microphones. Additionally, delay compensation is required for removing the SROs, which is, however, better known under the name "resampling". In the following we describe how these delays are treated.

The main task of the shift tracker in Fig. 1 is to keep the delays which are realized by phase shifts in the STFT domain smaller than one sample in order to avoid cyclic wrap around effects.

First consider the delay which is required for the SRO compensation, according to the model in (4). This requires the tracking of the continuously increasing/decreasing delay $n_h(l)$, where

$$n_h(l) = (1+\epsilon_{hj}) \cdot (lB + B/2). \tag{19}$$

This delay, representing the average delay between the h-th and the j-th channel, due to the SRO ϵ_{hj}, is split into an integer part

$$\overline{n}_h(l) = \lfloor (1+\epsilon) \cdot (lB + B/2) \rceil, \tag{20}$$

and a fractional part

$$\tilde{n}_h(l) = \overline{n}_h(l) - (1+\epsilon) \cdot (lB + B/2), \tag{21}$$

where $\lfloor\rceil$ denotes rounding towards the next integer. The buffer leap signal δ is triggered by the integer parts of $n_h(l)$

$$\delta(l) = \left(\overline{n}_h(l) - \overline{n}_h(l-1)\right), \tag{22}$$

while the fractional part $\tilde{n}_h(l)$ is compensated for by a multiplication with a complex exponential.

The second kind of delays, those required for the steering vector, are also treated by the shift tracker. Consider the TDOA $\tau_h(\boldsymbol{p})$ of node h. It is also split into an integer part $\overline{\tau}_h(\boldsymbol{p}) = \lfloor \tau_h(\boldsymbol{p}) \rceil$ and a fractional part $\tilde{\tau}_h(\boldsymbol{p}) = \tau_h(\boldsymbol{p}) - \overline{\tau}_h(\boldsymbol{p})$. A variable delay (see Fig. 1, block $z^{-\overline{\tau}_h}$) takes care of $\overline{\tau}_h(\boldsymbol{p})$ while the fractional parts of all channels are summarized in the vector $\tilde{\boldsymbol{\tau}}(\boldsymbol{p}) = [\tilde{\tau}_h(\boldsymbol{p}), \tilde{\tau}_i(\boldsymbol{p}), \tilde{\tau}_j(\boldsymbol{p})]^T$ and together form the latency reduced steering vector $\tilde{\boldsymbol{d}}(\boldsymbol{p})$.

From experiments on the STFT resampling procedure (documented in [12]) we know that for a Fast Fourier Transform (FFT) size of 1024 samples the precision of this STFT resampling method used here is limited to a maximum of approximately 55 dB Signal-to-Interpolation-Noise Ratio (SINR). Other resampling methods, e.g., the Overlap-Save method (OSM) method from [12], achieve better performance in terms of the SINR. Hence, the L-MVDR precision observed here may be limited by the precision of the resampling method which will be investigated in the following experiments.

6 Experiments

We consider a room with distributed microphones, where each microphone signal is sampled independently using oscillators whose frequencies slightly differ. We assume presence of a single target source (a speaker in the middle of the room) and a single coherent noise source (in one of the corners) at fixed positions in the room. The target speech source signal is taken from the TIMIT corpus by concatenating recordings to utterances of 30 s duration. All data was generated with the image method software from [14] where we tried to select the parameters according to

the results from [15]. The simulated room had the dimensions $5\,\text{m} \times 3\,\text{m} \times 3\,\text{m}$. We also added some uncorrelated sensor noise at 20 dB SNR on each channel to enhance the realism of the data.

Since our scenario assumes asynchronously sampled data we artificially introduced SROs by resampling the signals (see [12] for details) generated by the image method from [14].

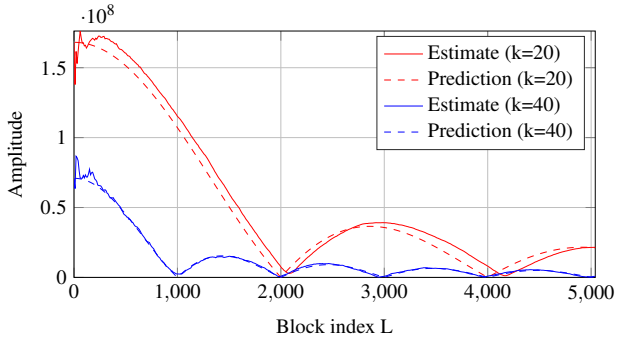

Figure 3: Estimation of matrix element $\boldsymbol{R}_{12}$, averaged across 150 experiments. Shown are the estimate (solid lines) and the prediction following (18) (dashed lines) at an SRO of 50 ppm for bins $k = [20, 40]$ (reverberation time $T_{60} = 350\,\text{ms}$).

6.1 Noise correlation matrix

To verify the result of (18), we generated 150 recordings of 60 s duration using the image method as described above setting the reverberation time to $T_{60} = 350\,\text{ms}$. Fig. 3 compares the estimate $\overline{R}_{12}(k)$ computed from the simulated signals with the theoretical result of (18) for the bins $k = 20$ and $k = 40$. The SRO between the nodes (1 and 2) was set to 50 ppm.

The estimates (solid lines) follow the values predicted by (18) (dashed lines) quite well. Note that the ground truth values were fixed at $|R_{12}(k=20)| = 1.68e{+}8$ and $|R_{12}(k=40)| = 7.09e{+}7$. The block shift B was set to half of the FFT size ($N = 128$), which means the first zero of the function $\boldsymbol{R}_{12}(k=40)$ is reached at block index $L = 1000$. With a sampling rate of 16 kHz this corresponds to 4 s after starting the system.

For values $L_\nu = (\nu \cdot N)/(B\epsilon_{hi}k)$, $\nu \in \mathbb{N}$ the estimates $\tilde{R}_{hi}(k)$ tend to be zero, as predicted by (18). The location of the zeros thus depends on the frequency bin k. If the value is zero, a multiplicative compensation of the bias is impossible. Only if no zeros occur for any of the frequencies an exact bias compensation is possible, i.e., only if $l < 2/(B\epsilon_{hi})$. These intermittent zeros may also explain why the approach of [5] suffers so much from local maxima during the optimization as described by the authors. So a correct estimate of $\boldsymbol{R}(k)$ requires a SRO compensation on the signals prior to time averaging.

6.2 Beamforming experiments

Table 1 summarizes the performance of the L-MVDR in terms of SNR gain and compares it to combinations of MVDR beamformer with the resampling methods STFT and OSM [12]. In case of absence of clock synchronization the MVDR performance degrades rapidly with increasing values of the SRO.

It can be seen that all explicit resampling methods and the new L-MVDR achieve roughly the same SNR gains, which is independent of the SRO. Comparing the execution times of the plain MVDR with the other methods, it is obvious that the L-MVDR is computationally more efficient than the variants with an explicit resampling stage, with the STFT resampling coming close, which is not surprising due to its similarity with the proposed method. Furthermore, the results reveal that the difference in precision of the tested resampling methods is so small that it does not affect the SNR gain obtained by the beamformer.

Fig. 2 depicts the improvements of the L-MVDR over an MVDR with uncompensated SRO in terms of the SNR gain ΔSNR (ΔSNR = $\text{SNR}_{\text{L-MVDR}}$ - SNR_{MVDR}). On the abscissa, the SNR of the coherent noise source is varied between −5 dB and 30 dB, while the sensor noise is fixed at 20 dB. MVDR and L-MVDR achieve the same results, i.e., $\Delta\,\text{SNR} = 0$ for synchronized data ($\epsilon_{ij} = 0, \forall i, j$), while for SROs of 50 and 100 ppm the L-MVDR achieves several dB gain over the MVDR, which, however not surprisingly, diminishes as the reverberation time and the SNR to the coherent noise source increase.

Table 1: Processing time per 1 s of a 4-channel audio segment and average SNR gain ($T_{60} = 200\,\text{ms}$, 10 dB SNR coherent noise, 20 dB SNR sensor noise)

Beamformer	Sync.	SRO 0 ppm	SRO ±50 ppm	SRO ±100 ppm	Avg. Time [ms]
MVDR	-	8.92	3.69	3.23	293.52
MVDR	STFT	8.92	8.92	8.92	310.67
MVDR	OSM	8.91	8.91	8.91	376.64
L-MVDR	-	8.92	8.92	8.92	308.13

7 Summary

We have presented an analytical model to explain the effect of SRO on the estimation of the spatial noise covariance matrix. The theoretical results were confirmed by experimentation. Further, we have presented the new L-MVDR beamformer, which is a SRO aware narrowband MVDR beamformer operating in the STFT domain. It has a reduced computational complexity compared to separate resampling and beamforming components, and the developed shift tracker reduces cyclic wrap around effects.

Acknowledgment

This work was supported by *Deutsche Forschungsgemeinschaft* (DFG) under contract no. SCHM 3301/1-1 within the framework of the Research Unit FOR2457 “Acoustic Sensor Networks”.

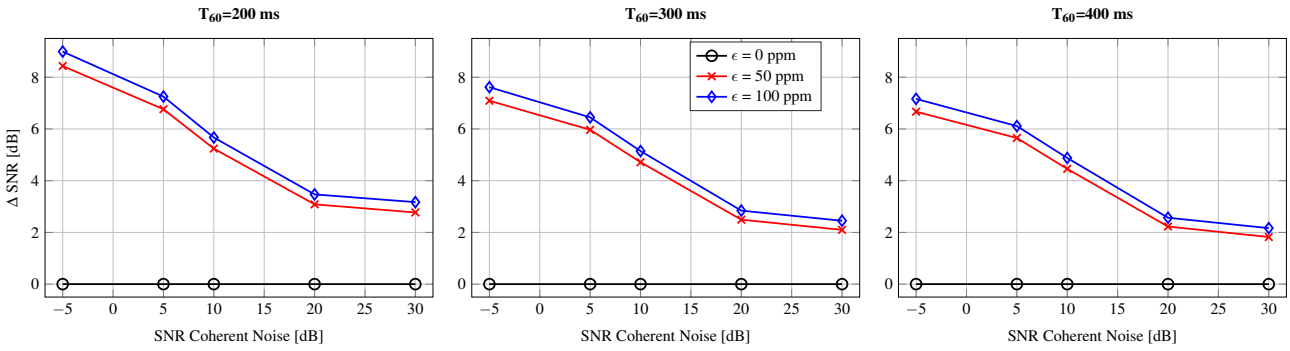

Figure 2: Improvements ΔSNR by L-MVDR in comparison to an MVDR beamformer processing the same data for different T_{60}-times, SROs and coherent noise source signal strengths (sensor noise of 20 dB).

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Nov 2011, pp. 1–6.

[2] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware-software approach for acoustic sensor network synchronization," *Signal Processing*, vol. 107, no. 0, pp. 171–184, 2015.

[3] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 2017.

[4] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 4–6, 2012.

[5] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization," in *23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 245–249.

[6] M. Pawig, G. Enzner, and P. Vary, "Adaptive sampling rate correction for acoustic echo control in voice-over-IP," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 189–199, Jan. 2010.

[7] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation," *Proc. IEEE Sixth International Symposium on Multimedia Software Engineering*, pp. 18–25, 2004.

[8] D. Cherkassky and S. Gannot, "Blind Synchronization in Wireless Acoustic Sensor Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017.

[9] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Speech and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[10] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind Sampling Rate Offset Estimation for Wireless Acoustic Sensor Networks Through Weighted Least-Squares Coherence Drift Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 674–686, 2017.

[11] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 674–678, 2013.

[12] J. Schmalenstroeer and R. Haeb-Umbach, "Efficient sampling rate offset compensation - an Overlap-Save based approach," in *26th European Signal Processing Conference (EUSIPCO) (EUSIPCO 2018)*, Rome, Italy, Sep. 2018.

[13] S. Haykin, *Adaptive Filter Theory - Fourth Edition*. Prentice-Hall, Information and system science series, 2002.

[14] E. Habets, "Emanuël habets github," *https://github.com/ehabets*.

[15] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan 2010.