# On the Integration of Sampling Rate Synchronization and Acoustic Beamforming

Tobias Gburrek, Joerg Schmalenstroeer and Reinhold Haeb-Umbach
Paderborn University, Germany
Department of Communications Engineering,
{gburrek, schmalen, haeb}@nt.uni-paderborn.de

***Abstract*—Time synchronization of the nodes of an acoustic sensor network is important to be able to apply acoustic beamforming for signal extraction. While former solutions perform the tasks of synchronization and beamformer coefficient estimation in cascade, we here show how they can be carried out jointly. The key observation is that the spatial covariance matrices of the speakers contain all information necessary to carry out both tasks. We simulate a meeting transcription system with asynchronous sensor nodes and show that the joint treatment not only saves some computations, but also leads to improved sampling rate offset estimation and beamforming. However, the final transcription performance turns out to be insensitive to those improvements.**

***Index Terms*—sampling rate offset, spatial covariance matrix, synchronization, beamforming, ad-hoc acoustic sensor networks, speech enhancement**

## I. Introduction

It is well known that the spatial distribution of the sensor nodes in a wireless acoustic sensor network (WASN) offers the opportunity for improved signal capture. This can be achieved by acoustic beamforming, whose full potential, however, can only be leveraged if the sampling clocks of the sensor nodes are synchronized [1]. Several techniques for sampling rate offset (SRO) estimation solely from the observed speech signals have been proposed [2]–[4]. They typically rely on observing the drift over time of the correlation or coherence of the microphone signals in different channels.

On the other hand, acoustic beamforming also relies on the estimation of cross-channel signal statistics in order to compute the beamformer coefficients. It therefore seems natural to investigate if some of the computations for SRO estimation and acoustic beamforming can be shared. A first indication that this is indeed possible is the observation in [5] that the dominant eigenvector of the spatial covariance matrix (SCM) of the microphone signals is directly related to the SRO. However, while in that publication a single sinusoidal source signal in white noise was considered, we here look at the practically more relevant setup of a meeting transcription system.

In a meeting scenario, the signal captured by the WASN is the speech signal of the participants, which may be corrupted by environmental noise. Participants are located at different spatial positions, e.g., are sitting around a table, they speak alternatingly, and sometimes more than a single speaker is active at the same time. This all poses additional challenges to SRO estimation and acoustic beamforming. Some of those challenges have been addressed in [3] and [6].

Usually, a cascade of synchronization and beamforming is proposed. The first task can be achieved, in its simplest form, by coarsely time-aligning stream segments in regular intervals to compensate for a sampling time offset (STO) between the channels, disregarding any SRO [7]. In more sophisticated approaches the channels are resampled to additionally compensate for an SRO [8], [9]. For the second task, acoustic beamforming, state of the art techniques employ a time-frequency mask estimator to obtain the activity patterns of each of the speakers in the meeting [10]–[12]. With those masks speaker-specific SCMs are estimated, from which in turn the beamformer coefficients are computed. Indeed, it is important that the SRO is compensated before beamformer estimation, because it has been shown that an SRO can have an irreversible detrimental effect on the SCM estimation process [13].

In this contribution we consider SCM estimation and SRO estimation and compensation jointly. First, a spatial mixture model is employed to estimate speaker-specific time-frequency masks [12]. Those are used to estimate the SCMs of the speakers, from which a matrix of all pair-wise microphone channel coherences is computed. From this matrix, an SRO estimate is derived by computing it's dominant eigenvector. SRO compensation is achieved by an upstream short-time Fourier transform (STFT) resampling [14] prior to SCM estimation. The coefficients of the acoustic beamformers operating on the time-aligned channels can be readily computed from the above mentioned SCMs. Compared to the cascaded approach, this joint treatment has two major advantages: the covariance terms only need to be computed once and outliers caused by computing the drift between segments originating from different speakers are avoided. Finally and arguably, the integrated approach is more elegant.

The paper is organized as follows: In Sec. II, the considered scenario is presented, followed by the mask-based SCM estimator in Sec. III. Section IV discusses the SCM-based SRO estimation including an integrated SRO compensation, and Sec. V briefly summarizes the source separation using minimum variance distortionless response (MVDR) beamforming. Experiments on simulated meeting data are discussed in Sec. VI before a brief conclusion is given in Sec. VII.

## II. Scenario description

The scenario at hand is an ad-hoc WASN consisting of $D \geq 2$ devices at fixed, however unknown, positions. Some devices, e.g., smartphones, may have a single microphone and some devices, e.g., smart devices like Alexa, may have multiple microphones. Overall, this results in a set of $M$ microphones being available. All devices are placed on a table and record a meeting-like conversation of $I$ speakers, which sit at fixed positions around the table. Mostly, a single speaker is active during the conversation. However, the conversation also includes quiet periods as well as a significant amount of time with two speakers being active at the same time.

The discrete-time signal recorded by the $m$-th microphone, $m \in \{1, \dots M\}$, is given by

$$y_m[n] = \sum_{i=1}^{I} h_{i,m}[n] * x_i[n] + v_m[n], \tag{1}$$

with $x_i[n]$ being the source signal emitted by the $i$-th speaker and $v_m[n]$ representing Gaussian sensor noise. $h_{i,m}[n]$ denotes the room impulse response (RIR) modeling the sound propagation from the position of the $i$-th speaker to the position of the $m$-th microphone. $n$ corresponds to the discrete-time sample index. The STFT of the $m$-th microphone signal is denoted by $Y_m(\ell, k)$ with frame index $\ell$, frequency bin index $k$, frame size $N$ and frame shift $B$.

As proposed in [3] the deviation of the sampling frequency of the $m$-th microphone $f_m[\ell] = (1+\varepsilon_m[\ell]) \cdot f_s$ from the nominal sampling frequency $f_s$ is modeled by the time-varying SRO $\varepsilon_m[\ell]$. Additionally, sampling of the microphone signals will start at different points in time resulting in an STO $T_m$. We assume that all STOs are handled by an initial synchronization procedure, e.g., the one explained in [6]. Consequently, we will omit the effect of STOs in the following, i.e., $T_m = 0$. If a device has multiple microphones it is assumed that the corresponding microphone signals are sampled with the same sampling frequency (see [15] for a hardware example), i.e., these signals show the same SRO.

## III. Spatial covariance matrix estimation

As it will be explained later we use speaker-dependent SCMs for source extraction via beamforming and SRO estimation. To this end a mask-based approach to SCM estimation is employed here. Let $\gamma_i(\ell, k), i \in \{1, \dots, I\}$, be a time-frequency mask indicating the activity of speaker $i$ in time frame $\ell$ and frequency bin $k$ and $\gamma_v(\ell, k)$ be the time-frequency mask of the noise. Furthermore, assume a frame-wise estimate of the speaker's activity $a_i[\ell] \in \{0, 1\}$ to be given. The sensor noise is assumed to be always active.

The SRO estimator and the beamformer are based on a block-wise processing. Therefore, both rely on SCMs $\boldsymbol{R}_i(\ell, k)$ which are calculated based on blocks consisting of $N_w$ consecutive frames:

$$\boldsymbol{R}_i(\ell, k) = \frac{\sum_{\tilde{\ell}=\ell-N_w+1}^{\ell} \gamma_i^2(\tilde{\ell}, k) \cdot a_i[\tilde{\ell}] \cdot \boldsymbol{Y}(\tilde{\ell}, k) \cdot \boldsymbol{Y}^H(\tilde{\ell}, k)}{\sum_{\tilde{\ell}=\ell-N_w+1}^{\ell} a_i[\tilde{\ell}]}, \tag{2}$$

with $\boldsymbol{Y}(\ell, k) = [Y_1(\ell, k), \dots, Y_m(\ell, k), \dots, Y_M(\ell, k)]^T$ corresponding to the stacked STFTs of the microphone signals. The quadratic weighting of the dyade product in (2), as proposed for example in [16], has shown improved experimental results for block-online beamforming compared to a linear weighting. Accordingly, the noise SCM $\boldsymbol{R}_v(\ell, k)$ is estimated using $\gamma_v(\ell, k)$ as mask.

## IV. SCM-based SRO estimation and compensation

In [3] we proposed the dynamic weighted average coherence drift (DWACD) method for SRO estimation, which has shown reliable results for estimating time-varying SROs in a scenario with speaker changes similar to the meeting scenario at hand. Here, we extend DWACD by deriving the coherence drift from speaker-dependent SCMs which are also used for beamforming as explained later.

### A. Dynamic weighted average coherence drift

In the following the basic concept of the DWACD method is shortly recapitulated on the basis of estimating the SRO between microphone $q$ and $r$, i.e., $\varepsilon_{rq}[\ell] = \varepsilon_q[\ell] - \varepsilon_r[\ell]$. The coherence between the microphones $q$ and $r$ is calculated as

$$\Gamma_{rq}(\ell, k) = \Phi_{rq}(\ell, k) / \sqrt{\Phi_{rr}(\ell, k) \cdot \Phi_{qq}(\ell, k)}, \tag{3}$$

with $\Phi_{rq}(\ell, k)$ denoting the power spectral density (PSD) of the channels $r$ and $q$ which is calculated via the Welch method:

$$\Phi_{rq}(\ell, k) = \frac{1}{N_w} \sum_{\tilde{\ell}=\ell-N_w+1}^{\ell} Y_r(\tilde{\ell}, k) \cdot Y_q^*(\tilde{\ell}, k). \tag{4}$$

The SRO is estimated on the basis of the complex-conjugated product of two consecutive coherence functions which is computed as

$$P_{rq}(\ell, k) = \Gamma_{rq}(\ell, k) \cdot \Gamma_{rq}^*(\ell - \ell_d, k), \tag{5}$$

where $\ell_d$ corresponds to the temporal distance between the two coherence functions. Before estimating the SRO, the complex-conjugated product of consecutive coherence functions is smoothed via an first-order autoregressive process, resulting in $\bar{P}_{rq}(\ell, k)$. Eventually, the SRO is estimated via

$$\widehat{\varepsilon}_{rq}[\ell] = -\frac{1}{\ell_d \cdot B} \cdot \left( \underset{\lambda}{\operatorname{argmax}} \; |\bar{p}_{rq}(\ell, \lambda)| \right), \tag{6}$$

where $\bar{p}_{rq}(\ell, \lambda) = \mathrm{IFFT}_N\left(\bar{P}_{rq}(\ell, k)\right)$ denotes the $N$-point inverse fast Fourier transform (IFFT) of $\bar{P}_{rq}(\ell, k)$. Note that the SRO is estimated only every $B_{\mathrm{S}}$ frames and therefore $P_{rq}(\ell, k)$ is also computed only every $B_{\mathrm{S}}$ frames.

### B. SCM-based SRO estimation

Let $\Phi_{i,rq}(\ell,k)$ be the PSD of the channels $q$ and $r$ for the signal of speaker $i$. By comparing (2) and (4) it can be seen that $\Phi_{i,rq}(\ell,k)=(\boldsymbol{R}_i(\ell,k))_{r,q}$ holds if only speaker $i$ is active. Here, $(\boldsymbol{R}_i(\ell,k))_{r,q}$ corresponds to the $r$-th row and $q$-th column element of the SCM $\boldsymbol{R}_i(\ell,k)$. Thus, the coherences $\Gamma_{i,rq}(\ell,k)$ as well as the products of the coherence functions $P_{i,rq}(\ell,k)$ can be computed speaker-dependently based on the elements of the SCMs $\boldsymbol{R}_i(\ell,k)$ via (3) and (5). This reduces the computational overhead since the speaker-dependent SCMs $\boldsymbol{R}_i(\ell,k)$ are also used for beamforming and no additional speaker-independent coherences $\Gamma_{rq}(\ell,k)$ need to be computed.

The speaker-dependent complex-conjugated product of coherence functions comes with the advantage that the two coherence functions $\Gamma_{i,rq}(\ell,k)$ and $\Gamma_{i,rq}(\ell-\ell_d,k)$ always belong to the same source position. Thus, the phase of $P_{i,rq}(\ell,k)$ only depends on the SRO as discussed in [3]. This might not hold for $P_{rq}(\ell,k)$ for all blocks due to the speaker changes.

The phase of the coherence $\Gamma_{rq}(\ell,k)$ is typically too volatile for accurate SRO estimation when a coherent source is active in only a few frames that are used to estimate $\Gamma_{rq}(\ell,k)$. In the DWACD method, these frequency bins are dominated by sensor noise, resulting in a small absolute value of the coherence estimate. Hence, these frequency bins get a small weight when estimating the SRO. However, the speaker-dependent coherence $\Gamma_{i,rq}(\ell,k)$ might still have a large absolute value in this case due to the normalization in (3). To mitigate this effect we weigh $P_{i,rq}(\ell,k)$ by $w_i(\ell,k)=\sqrt{\widetilde{w}_i(\ell,k)\widetilde{w}_i(\ell-\ell_d,k)}$. Hereby, $\widetilde{w}_i(\ell,k)$ reflects the average amount of frames within the $\ell$-th block in which speaker $i$ is dominant for frequency bin $k$. The decision if speaker $i$ is dominant is made by quantizing the mask $\gamma_i(\ell,k)$ to zero or one using a threshold of $0.9$.

The speaker-dependent complex-conjugated products of coherence functions $P_{i,rq}(\ell,k)$ of all active speakers are added up in blocks where speech from multiple speakers overlaps. Therefore, the complex-conjugated product of consecutive coherence $P_{rq}(\ell,k)$ is given by

$$P_{rq}(\ell,k)=\sum_{i\in\widetilde{\mathcal{A}}[\ell]} w_i(\ell,k)\cdot P_{i,rq}(\ell,k), \tag{7}$$

with $\widetilde{\mathcal{A}}[\ell]$ denoting the set of speaker indices belonging to the speakers which show a suitable activity in both signal segments used to calculate $P_{i,rq}(\ell,k)$.

Moreover, the SCMs $\boldsymbol{R}_i(\ell,k)$ provide all values to calculate the complex-conjugated products of coherence functions $P_{rq}(\ell,k)$ for all channel combinations. Thus, an extension of the pair-wise SRO estimator to a multi-channel version, which takes into account the relationships between all SROs $\varepsilon_{rq}[\ell]$, is possible without much overhead. First of all, all pair-wise estimates of $\bar{P}_{rq}(\ell,k)$ are combined into a matrix $\bar{\boldsymbol{P}}(\ell,k)$ with $(\bar{\boldsymbol{P}}(\ell,k))_{r,q}=\bar{P}_{rq}(\ell,k)$. As shown in [3] $P_{rq}(\ell,k)$ can be decomposed into a signal-to-noise ratio (SNR)-related weight $\psi_{rq}(\ell,k)$ with zero-phase and a phase term $\varphi_{rq}(\ell,k)=\exp\left(j2\pi k\ell_d B\varepsilon_{rq}[\ell]/N\right)$ that depends on the SRO. Utilizing channel $0$ as reference for SRO estimation, $\bar{\boldsymbol{P}}(\ell,k)$ can be written as

$$\bar{\boldsymbol{P}}(\ell,k)=\operatorname{diag}(\boldsymbol{\varphi}(\ell,k))\cdot\boldsymbol{\Psi}(\ell,k)\cdot(\operatorname{diag}(\boldsymbol{\varphi}(\ell,k)))^H, \tag{8}$$

with $(\boldsymbol{\Psi}(\ell,k))_{r,q}=\psi_{rq}(\ell,k)$, $(\boldsymbol{\varphi}(\ell,k))_m=\varphi_{0m}(\ell,k)$ and $\operatorname{diag}(\boldsymbol{\varphi}(\ell,k))$ corresponding to a diagonal matrix formed by the elements of $\boldsymbol{\varphi}(\ell,k)$. Thus, $\boldsymbol{\varphi}(\ell,k)$ and hence the SRO can be estimated from the dominant eigenvector of $\bar{\boldsymbol{P}}(\ell,k)$ as proposed in [5]. To this end, after each update of $\bar{\boldsymbol{P}}(\ell,k)$, we perform a single power iteration round to track the dominant eigenvector $\widehat{\boldsymbol{d}}(\ell,k)$ using its previous estimate as initialization. Finally, $\widehat{\boldsymbol{d}}(\ell,k)$ is weighted by it's corresponding eigenvalue to retain the benefit of the SNR-related weights $\psi_{rq}(\ell,k)$, and takes the role of $\bar{P}_{rq}(\ell,k)$ for the SRO estimation procedure via (6).

As described in [1] and [17], a non-zero SRO can bias SRO estimates. In addition, it has a detrimental effect on the SCM estimates as shown in [13]. Therefore, we use the SRO estimates prior to SCM estimation in an upstream online STFT resampling [14] procedure.

## V. Source extraction via MVDR beamforming

To extract the signals of the individual speakers from the recordings we build upon the approach based on a speaker-dependent, time-varying MVDR beamformer we presented in [9]. Thereby, the STFT of the signal of the $i$-th speaker is estimated as

$$\widehat{X}_i(\ell,k)=\boldsymbol{W}_i^H(\ell,k)\cdot\boldsymbol{Y}(\ell,k). \tag{9}$$

The beamformer filter coefficients are calculated according to [18] with

$$\boldsymbol{W}_i(\ell,k)=\frac{\left(\widetilde{\boldsymbol{\Phi}}_i(\ell,k)\right)^{-1}\cdot\boldsymbol{\Phi}_i(\ell,k)}{\operatorname{tr}\left\{\left(\widetilde{\boldsymbol{\Phi}}_i(\ell,k)\right)^{-1}\cdot\boldsymbol{\Phi}_i(\ell,k)\right\}}\cdot\boldsymbol{u}, \tag{10}$$

where $\operatorname{tr}\{\cdot\}$ is the trace operator, and $\boldsymbol{u}$ a unit vector pointing to a reference microphone. $\boldsymbol{\Phi}_i(\ell,k)$ denotes the SCM of the target speaker and $\widetilde{\boldsymbol{\Phi}}_i(\ell,k)$ is the SCM of the interference, i.e., the SCM of sensor noise and interfering speakers.

Similar as in [9] the SCMs $\boldsymbol{\Phi}_i(\ell,k)$ and $\widetilde{\boldsymbol{\Phi}}_i(\ell,k)$ are determined based on the speaker-dependent SCM estimates $\boldsymbol{R}_i(\ell,k)$. $\boldsymbol{\Phi}_i(\ell,k)=\boldsymbol{R}_i(\ell,k)$ is chosen for the SCM of the target speaker. The SCM of the interference $\widetilde{\boldsymbol{\Phi}}_i(\ell,k)$ is given by the sum of the SCM estimates of all interfering sources:

$$\widetilde{\boldsymbol{\Phi}}_i(\ell,k)=\boldsymbol{R}_v(\ell,k)+\sum_{g\in\mathcal{A}[\ell]\setminus\{i\}}\boldsymbol{R}_g(\ell,k). \tag{11}$$

Here, $\mathcal{A}[\ell]$ corresponds to the set of speaker indices for which an activity is detected in the $\ell$-th block.

To avoid too frequent and too large updates of the beamforming coefficients, which can, e.g., deteriorate the performance of an automatic speech recognition (ASR) system, the speaker-dependent SCMs are updated in a block-online manner. Consequently, $\boldsymbol{\Phi}_i(\ell,k)$ and $\widetilde{\boldsymbol{\Phi}}_i(\ell,k)$ and, therefore, the beamforming coefficients $\boldsymbol{W}_i(\ell,k)$, are only updated once

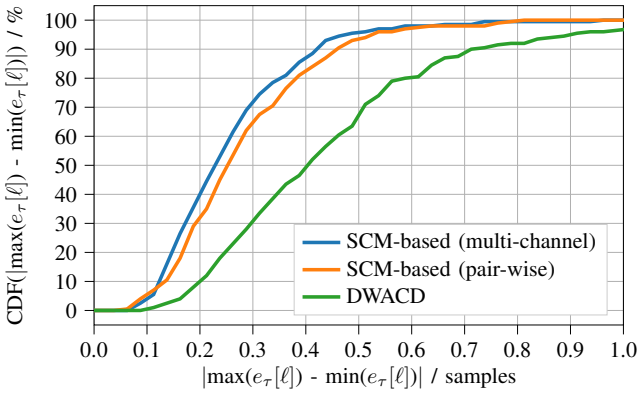


Fig. 1. Cumulative distribution function (CDF) of the peak-to-peak value of the remaining synchronization error after compensating for SROs

per block and are used to filter all frames within each block. Moreover, the speaker-dependent SCM estimates $\boldsymbol{R}_i(\ell, k)$ are smoothed via a first-order autoregressive process.

## VI. Experiments

The proposed integrated sampling rate synchronization and beamforming system[1] is evaluated on a data set consisting of 100 meetings which are simulated with the meeting generator from [19]. All meetings are 2 min long and use the spatial setups of the meeting data set from [6]. In 60% of the total meeting duration a single speaker is active, in 23% of the total meeting duration two speakers are concurrently active and in the remaining time no speaker is active. The audio signals are recorded by $D$=3 independent devices. Thereby, one device is equipped with a microphone array having a quadratic layout with an edge length of 5 cm and the other two devices are equipped with a single microphone. Furthermore, time-varying SROs are simulated as described in [3]. The mean value of the simulated SRO trajectories is randomly drawn from a uniform distribution in the interval $[-100\,\text{ppm}, 100\,\text{ppm}]$.

For mask estimation a complex Angular Central Gaussian Mixture Model (cACGMM) [20] with time-varying instead of frequency-dependent mixture weights [21] is employed on the signals of the microphone array. The cACGMM is initialized as described in [12] and delivers posterior probabilities, which are used as the masks $\gamma_i(\ell, k)$ and $\gamma_v(\ell, k)$. As proposed in [12] frame-wise estimates of the speakers' activity $a_i[\ell]$ are gathered by comparing a smoothed version of the prior probabilities $\pi_i[\ell]$ of the cACGMM to a threshold. Note that the device with the microphone array is only needed due to the usage of an offline version of the cACGMM. By replacing the cACGMM by a single-channel mask estimator or an online version of the cACGMM, which works on the resampled signals from multiple devices, the need for a device with multiple microphones can be removed. Moreover, the ASR results are obtained using an acoustic model which is trained on 16 kHz SMS-WSJ data [22] and configured as described in [22].

[1]Code & parametrization are available at https://github.com/fgnt/paderwasn

TABLE I
Transcription performance for varying microphone constellations and synchronization conditions

| Setup | SRO Compensation | cpWER / % |
|---|---|---|
| Clean audio | — | 5.97 |
| Single-array | — | 18.57 |
| Array + 2 sync. mics | — | 11.81 |
| Array + 2 async. mics | — | 16.08 |
| Array + 2 async. mics | DWACD-based | 11.88 |
| Array + 2 async. mics | SCM-based | 11.97 |

Fig. 1 shows the cumulative distribution function (CDF) of the absolute value of the peak-to-peak value of the synchronization error $e_\tau[\ell]$ which remains after compensating for the SROs. Here, the remaining synchronization error is defined as

$$e_\tau[\ell] = \frac{N}{2} \cdot (\varepsilon_{0m}[0] - \widehat{\varepsilon}_{0m}[0]) + \sum_{\tilde{\ell}=1}^{\ell} (\varepsilon_{0m}[\tilde{\ell}] - \widehat{\varepsilon}_{0m}[\tilde{\ell}]) \cdot B. \quad (12)$$

Note that this metric was chosen since it reflects the long-term synchronization stability better than error metrics directly applied to the SRO estimates. It can be seen that the proposed approach to SRO estimation based on speaker-dependent SCMs is able to outperform the DWACD method. Due to the large similarity between both approaches, this might be mostly explained by the fact that the SCMs-based SRO estimator always combines only coherence functions which belong to the same source position. Moreover, it can be seen that the proposed multi-channel extension to the SCM-based SRO estimator leads to a small additional gain in performance. Thus, the multi-channel version of the SCM-based SRO estimator will be used in the following experiments.

The transcription performance for different microphone constellations and different synchronization conditions is compared in Table I on the basis of the concatenated minimum-permutation word error rate (cpWER) [23]. It can be seen that the cpWER can be improved a lot by using two additional spatially distributed microphones. However, it becomes obvious that the sampling rate of the devices have to be synchronized to make use of the whole potential of the additional microphones. Compensating for the SROs either by the proposed integrated sampling rate synchronization or SRO estimation via the DWACD method with following offline STFT resampling leads to a transcription performance which is very close to the one for synchronous devices.

In the meeting scenario at hand the periods in time with overlapping speech are most crucial for the transcription performance. Therefore, Fig. 2 visualizes the distribution of the invasive SDR for these periods in time. It can be seen that the additional devices are beneficial for the suppression of concurrent speakers and a compensation for SROs is necessary. Note that the SDR of the proposed integrated system coincides with the SDR which can be achieved when synchronous devices are used. Thus, the small difference in the cpWER cannot be explained by a reduced suppression of concurrent speakers.

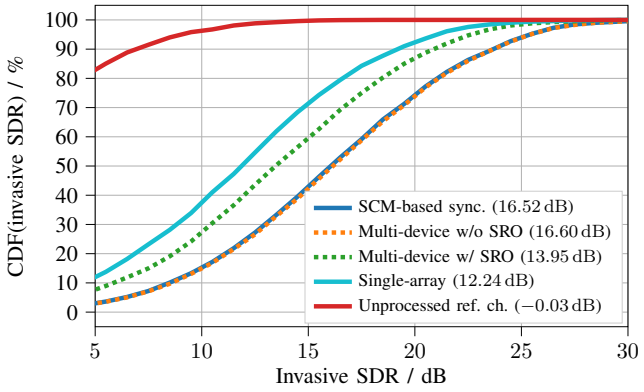


Fig. 2. Cumulative distribution function (CDF) of the invasive SDR for overlapping speech regions. The numbers in brackets correspond to the the average SDR values. The curve for DWACD + downstream resampling coincides with the curve for the integrated SCM-based sampling rate synchronization and is therefore omitted.

## VII. Conclusions

In this paper an integrated sampling rate synchronization and acoustic beamforming system is introduced. For the considered meeting scenario the proposed system allows to make use of the benefits which can be gained from distributed recording devices enabling a better suppression of interfering sources via beamforming. However, the sampling rates of independent recording devices typically differ from each other prohibiting to utilize the whole potential of beamforming. Therefore, the proposed system estimates the SRO based on speaker-dependent SCMs which are also used for beamforming. Beyond reducing the number of computational operations this also improves the SRO estimation performance by ensuring that the SRO is always estimated based on periods in time with constant source positions. Eventually, these SROs are fed back into an upstream online resampling procedure before SCM estimation. For simulated meeting data it was shown that the proposed system is able to achieve the same suppression of concurrent speakers as a system utilizing synchronous recording devices.

## Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project 282835863.

## References

[1] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddeker, 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)*, October 2017.

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

[3] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[4] A. Chinaev, P. Thüne, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021.

[5] A. Ganti and J. Krolik, "Sampling rate offset estimation in acoustic sensor networks," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, 2021, pp. 436–440.

[6] T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, and R. Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," *arXiv preprint arxiv.2205.00944*, 2022.

[7] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, and X. Huang, "Meeting Transcription Using Asynchronous Distant Microphones," in *Proc. Interspeech 2019*, 2019, pp. 2968–2972.

[8] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 32–39.

[9] T. Gburrek, J. Schmalenstroeer, J. Heitkaemper, and R. Haeb-Umbach, "Informed vs. blind beamforming in ad-hoc acoustic sensor networks for meeting transcription," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[10] N. Furnon, R. Serizel, S. Essid, and I. Illina, "DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2310–2323, 2021.

[11] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 835–848, 2023.

[12] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, and R. Haeb-Umbach, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models," in *Proc. Interspeech 2022*, 2022, pp. 271–275.

[13] J. Schmalenstroeer and R. Haeb-Umbach, "Insights into the interplay of sampling rate offsets and MVDR beamforming," in *ITG 2018, Oldenburg, Germany*, 2018.

[14] ——, "Efficient sampling rate offset compensation - an overlap-save based approach," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018.

[15] H. Afifi, J. Schmalenstroeer, J. Ullmann, R. Haeb-Umbach, and H. Karl, "MARVELO - a framework for signal processing in wireless acoustic sensor networks," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech 2016*, 2016, pp. 1981–1985.

[17] A. Chinaev, S. Wienand, and G. Enzner, "Control architecture of the double-cross-correlation processor for sampling-rate-offset estimation in acoustic sensor networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 801–805.

[18] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[19] T. Cord-Landwehr, T. von Neumann, C. Boeddeker, and R. Haeb-Umbach, "MMS-MSG: A multi-purpose multi-speaker mixture signal generator," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022.

[20] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[21] ——, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[22] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[23] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.