Multi-Stage Coherence Drift Based Sampling Rate Synchronization for Acoustic Beamforming

Joerg Schmalenstroeer, Jahn Heymann, Lukas Drude, Christoph Boeddecker and Reinhold Haeb-Umbach
Department of Communications Engineering, Paderborn University
Email: {schmalen, heymann, drude, haeb}@nt.uni-paderborn.de

***Abstract*—Multi-channel speech enhancement algorithms rely on a synchronous sampling of the microphone signals. This, however, cannot always be guaranteed, especially if the sensors are distributed in an environment. To avoid performance degradation the sampling rate offset needs to be estimated and compensated for. In this contribution we extend the recently proposed coherence drift based method in two important directions. First, the increasing phase shift in the short-time Fourier transform domain is estimated from the coherence drift in a Matched Filter-like fashion, where intermediate estimates are weighted by their instantaneous SNR. Second, an observed bias is removed by iterating between offset estimation and compensation by resampling a couple of times. The effectiveness of the proposed method is demonstrated by speech recognition results on the output of a beamformer with and without sampling rate offset compensation between the input channels. We compare MVDR and maximum-SNR beamformers in reverberant environments and further show that both benefit from a novel phase normalization, which we also propose in this contribution.**

I. Introduction

A common scenario for Wireless Acoustic Sensor Networks (WASN) is given by a setup where each sensor node of the network has an independent oscillator driving the local Analog-to-Digital Converter (ADC). Hence, the data streams originating from the individual sensor nodes are sampled at slightly different rates. However, many practically relevant signal processing techniques require synchronized data streams and are known to deteriorate with an increasing Sampling Rate Offset (SRO), e.g., echo cancellation [1], blind source separation [2] and beamforming [3]. Consequently, SRO estimation and compensation becomes an essential task for signal processing on WASN data streams.

The SRO can be estimated either by a time stamp exchange algorithm (e.g., [4]), by using timing information from the sampling process as proposed in [5], or by examining the audio streams to obtain SRO estimates in the time [6] or frequency [3], [7], [8] domain. Subsequently, to compensate for the SRO, either the hardware is reconfigured [5], or the signals are resampled in software (e.g., using Lagrange polynomials [3], band-limited interpolation [6] or frequency domain methods [9]), or the estimated SRO is directly taken care of in the original signal processing task.

In [3] the authors propose to use the coherence function to estimate the SRO between two data streams. By observing the phase drift between the coherence functions computed on two temporally adjacent signal segments an estimate of the SRO can be obtained. Bahari et al. extended this idea in [10] by replacing the temporal averaging of the observations with a least squares approach, and in [7] an additional outlier detection was introduced.

Some authors use an exhaustive search for determining SRO values, where either in the time domain a scaling [6] or in the frequency domain a resampling [8] of all possible SROs and delays are evaluated against a cost function. If the cost function itself is smooth enough an iterative optimization procedure or a smart grid search can be applied to reduce the overall computational complexity (e.g., [6]).

In this paper we extend the coherence drift based SRO estimator in two directions. First, we introduce an SNR-related weighting, and second, we propose a multi-stage procedure, where SRO estimation and compensation by resampling are alternatingly carried out. The latter is motivated by an observed bias which increases with the true SRO. We compare the performance of this modified estimator against the approaches from [3] and [8].

SRO compensation is required for the subsequent signal processing tasks. Here, we consider acoustic beamforming. The impact of a fixed but unknown delay between channels or even SRO on the beamforming result depends on the implemented beamforming technique. Especially error prone are techniques which rely on the array geometry and require a geometrically motivated steering vector (e.g., delay and sum beamformer or Minimum Variance Distortionless Response (MVDR) beamformer with conventional steering vector consisting of pure delays) [11], [12]. A short discussion on the detrimental effects of even small SROs on Direction-of-Arrival (DoA) estimates can be found in [5].

In case the beamforming technique at hand blindly estimates the beamforming vector from the observed data, it may compensate for moderate delays between channels. If the beamforming vector is extracted by using cross power spectral density matrices only, small delays will change the beamforming vector such that it incorporates a compensation of those delays. Nevertheless, this renders the beamforming vector to not be geometrically interpretable anymore.

Of particular relevance is an MVDR beamformer where the Acoustic Transfer Function (ATF) vector is obtained as the principal component of the target covariance matrix. This vector is then used in the MVDR formalism to obtain the beamforming vector [13].

An alternative, statistically motivated, beamforming approach is the maximum-SNR, also called Generalized Eigenvalue (GEV) beamformer [14], which will be put into use here. Much in the sense of the MVDR, a target and a noise cross power spectral density matrix is obtained from the multichannel observation. The method can therefore also compensate small delays inherently.

It is important to note that if the ATF vector is obtained by eigenvector decomposition, there is still a phase ambiguity, even in the case of the MVDR beamformer. In this contribution we propose to fix the phase by minimizing the group delay.

To allow an objective comparison, we evaluate the performance of the used algorithms in terms of word error rates (WERs) with a backend based on a Wide Residual Network (WRN) [15], [16] on a dataset based on the 4th CHiME challenge [17] to ensure that possible gains are still visible with a rather robust acoustic model.

The paper is organized as follows. In Section II the SRO model is introduced and in Section III the coherence drift is explained. Section IV discusses approaches for estimating the SRO from the coherence function either in a single or multi-stage fashion. Beamforming and phase normalization techniques are presented in Section V and the utilized ASR backend in Section VI. After discussing some experiments in Section VII we end with some conclusions.

II. Sampling Rate Offset Model

Assume we select two arbitrary nodes R and S from a sensor network. Although the sampling frequency is nominally the same, the nodes will operate at slightly different sampling frequencies f_S and f_R, since both nodes have different hardware oscillators.

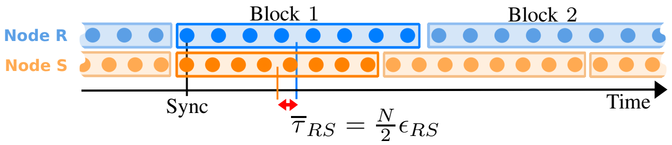

Fig. 1. Visualization of average delay $\overline{\tau}_{RS}$, introduced by SRO ϵ_{RS} during block-oriented processing of data streams ($N = 8, B = 8$).

Without loss of generality we select f_R of node R to be the reference sampling rate and define the SRO ϵ_{RS} between the nodes R and S to be:

$$f_S = (1 + \epsilon_{RS}) \cdot f_R. \tag{1}$$

Each node samples the impinging microphone signals and generates a sequence of time-discrete values $x_i(n)$ with $i \in \{R, S\}$ which are further processed in the Short Time Fourier Transform (STFT) domain.

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

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

where the $x_i(n)$ are the time domain samples. Here, the index $i \in \{R, S\}$ indicates that the signal has been sampled by the i-th node's oscillator. As proposed in [7] we assume the input signals to be an additive composition of a coherent source signal $S_i(l, k)$ (a speaker in our scenario), filtered by an unknown transfer function $H_i(l, k)$, and a spatially uncorrelated noise term $V_i(l, k)$:

$$X_i(l,k) = H_i(l,k) \cdot S_i(l,k) + V_i(l,k), \tag{3}$$

A non-zero SRO will introduce an average delay $\overline{\tau}_{RS}$ between the data streams of the nodes (see Fig.1), which can be approximated by $\overline{\tau}_{RS} \approx \frac{N}{2}\epsilon_{RS}$ (see [9], [8]). Furthermore, it is reasonable to assume that nodes in a WASN start the sampling process asynchronously and that the nodes are at different distances from the source. Hence, a fixed delay τ_{RS} between the nodes' data streams has to be regarded in the following. The signal modifications by the fixed delay (starting point) and the increasing delay (SRO) can be modeled by a multiplication with a time-variant phase term in the STFT domain. If the overall delay between the channels remains small compared to the STFT size the following correspondence between the coherent signal parts can be assumed:

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

III. Coherence drift estimation

SRO estimation is basically the task of robustly determining the phase term in Eq. (4) [3]. To this end the coherence function $\Gamma_{R,S}(l, k)$ of the l-th block

$$\Gamma_{R,S}(l,k) = \frac{\Psi_{R,S}(l,k)}{\sqrt{\Psi_{R,R}(l,k) \cdot \Psi_{S,S}(l,k)}} \tag{5}$$

is employed, where $\Psi_{i,j}(l, k)$ $(i, j \in \{R, S\})$ denotes the Power Spectral Density (PSD), which is estimated via the Welch method:

$$\Psi_{i,j}(l,k) = \frac{1}{N_W} \sum_{\kappa=0}^{N_W - 1} X_i(l+\kappa, k) \cdot X_j(l+\kappa, k)^*. \tag{6}$$

$\Psi_{R,S}(l, k)$ is the cross PSD and $\Psi_{R,R}(l, k)$ and $\Psi_{S,S}(l, k)$ are the auto PSDs.

In the following only a single source scenario is regarded to keep the expressions compact, however, as explained in [7] it may be extended towards multiple sources. Inserting the model Eq. (3) into the Welch method Eq. (6) and using the expressions within Eq. (5) results in Eq. (23) (see last page), where we assumed that the unknown transfer functions $H_i(l, k)$ are constant within the window size of the Welch method, i.e., it is assumed that the movement of the speaker is negligible during the duration of a window. Eq. (23) consists of three terms where

$$H_{R,S}(k) = \frac{H_R(k) H_S^*(k)}{|H_R(k)||H_S(k)|} \tag{7}$$

summarizes the transfer functions,

$$W_{R,S}(l,k) = \frac{\sum\limits_{\kappa=0}^{N_W-1} |S_R(l{+}\kappa,k)|^2 \, \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}\kappa Bk\epsilon_{RS}}}{\sqrt{\overline{|X_R(l{+}\kappa,k)|^2} \cdot \overline{|X_S(l{+}\kappa,k)|^2}}} \tag{8}$$

comprises a signal-to-noise ratio (SNR) related weighting term and

$$\phi_{R,S}(l,k) = \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}[\tau_{RS}+\epsilon_{RS}(\frac{N}{2}+lB)]k} \tag{9}$$

is the desired term for calculating the phase information. To ease the notation we summarize the denominator terms describing the input signal energy by

$$\overline{|X_R(l,k)|^2} = \sum_{\kappa=0}^{N_W-1} |S_R(l{+}\kappa,k)|^2 + \frac{|V_R(l{+}\kappa,k))|^2}{|H_R(k)|^2} \tag{10}$$

and

$$\overline{|X_{RS}(l,k)|^2} = \sqrt{\overline{|X_R(l,k)|^2} \cdot \overline{|X_S(l,k)|^2}}. \tag{11}$$

IV. SRO Estimation

Following [3] or [7] the phase information can be approximately retrieved by dividing two consecutive coherence functions:

$$\frac{\Gamma_{R,S}(l{+}p,k)}{\Gamma_{R,S}(l,k)} \approx \mathrm{e}^{+j\frac{2\pi}{N}(pBk)\epsilon_{RS}}. \tag{12}$$

However, inspecting the detailed result in Eq. (24) reveals that this approximation relies on the assumption that the ratio

$$\frac{\sum\limits_{\kappa=0}^{N_W-1} |S_R(l{+}\kappa,k)|^2 \, \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}(\kappa Bk)\epsilon_{RS}}}{\sum\limits_{\kappa=0}^{N_W-1} |S_R(l{+}\kappa{+}p,k)|^2 \, \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}(\kappa Bk)\epsilon_{RS}}} \tag{13}$$

is real-valued. But this only holds if either all $|S_R(l{+}\kappa,k)|^2$ equal $|S_R(l{+}\kappa+p,k)|^2$ or if ϵ_{RS} is close to zero. Usually it can be assumed that speech signals are sparse and thus violate the assumption to have equal signal power in consecutive frames (or in frames p block sizes apart). Hence, the estimate Eq. (12) will deteriorate with increasing values of SRO ϵ_{RS} and frequency bins k.

Furthermore, the approach drops the important information about the actual presence of a coherent source. Signal segments with active sources and segments without any coherent source are treated equally, disregarding the fact that segments with coherent sources provide more reliable information for phase estimation. However, reliability information is available through the magnitude of the coherence functions.

A. Weighted SRO estimation

We propose to use the complex conjugate product of consecutive coherence functions

$$\Gamma_{R,S}(l{+}p,k) \cdot \Gamma^*_{R,S}(l,k) = W_{\text{SNR}} \cdot \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}(pBk)\epsilon_{RS}} \tag{14}$$

to estimate a reliability weighted phase term, where the magnitude is proportional to the product of the coherence function values with $W_{\text{SNR}} = W_{R,S}(l{+}p,k) \cdot W^*_{R,S}(l,k)$. Now averaging the complex conjugate products across an utterance will automatically weigh each individual estimate by its SNR.

We define the temporally averaged phase information $P(k)$ for the Average Coherence Drift (ACD) approach by

$$P^{\text{ACD}}(k) = \frac{1}{L}\sum_{l=1}^{L} \frac{\Gamma_{R,S}(l{+}p,k)}{\Gamma_{R,S}(l,k)}, \tag{15}$$

and for our new proposed Weighted Average Coherence Drift (WACD) method by

$$P^{\text{WACD}}(k) = \frac{1}{L}\sum_{l=1}^{L} \Gamma_{R,S}(l{+}p,k) \cdot \Gamma^*_{R,S}(l,k), \tag{16}$$

where L is the number of coherence functions averaged. The SRO can be either estimated via the ACD approach from [3] with

$$\hat{\epsilon}^{\text{ACD}}_{RS} = \frac{1}{K_{\max}} \sum_{k=1}^{K_{\max}} \frac{N}{2\pi pBk} \angle\left\{P^{\text{ACD}}(k)\right\} \tag{17}$$

or by the proposed WACD via

$$\hat{\epsilon}^{\text{WACD}}_{RS} {=} \frac{\epsilon_{\max}}{\pi} \angle \left\{ \sum_{k=1}^{K_{\max}} \left|P^{\text{WACD}}(k)\right| \exp\left(\frac{jN\angle\left\{P^{\text{WACD}}(k)\right\}}{2pBk\epsilon_{\max}}\right)\right\}, \tag{18}$$

where $\angle\{\}$ denotes the phase. Eq. (18) averages first across the utterance in the time domain and subsequently projects the result into the complex plane for averaging across the frequency bins. Given an assumed maximum SRO of $\pm\epsilon_{\max}$, the range of the normalized angles is limited to $\pm\pi$.

Here $K_{\max} = N/(2pB\epsilon_{\max})$ is the maximum frequency bin index without phase ambiguity, if the maximum SRO $\epsilon_{\max}$ occurs [3].

B. Multi-Stage SRO estimation

Coherence drift based SRO estimation suffers from the assumption that $\phi_{R,S}(l,k)$ (Eq. (9)) is the only phase contributing term, while all other terms, e.g., the product (WACD) or the ratio (ACD) of weighting terms $W_{R,S}(l,k)$, are real-valued. This shortcoming can be addressed by a resampling step reducing the inter-channel SRO, since with $\epsilon_{RS} \to 0$ all phase terms $e^{+j\frac{2\pi}{N}\kappa Bk\epsilon_{RS}}$ in Eq. (8) tend to be one and $W_{R,S}(l,k)$ becomes real-valued.

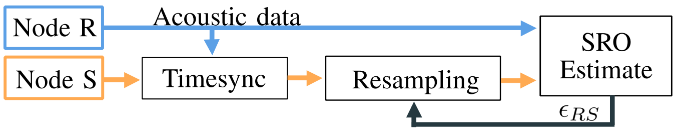

Fig. 2. Multi-Stage SRO estimation with initial GCC-based rough time synchronization and subsequent iterative SRO estimation and resampling.

Consider the two node example depicted in Fig. 2. At first a rough synchronization between the audio streams of node R and node S is conducted, where the initial (fixed) delay τ_{RS} is estimated with a GCC-PHAT-based method [18] and all leading zeros are dropped via an energy based Voice Activity Detection (VAD). During the first iteration (Stage

1) the resampling step is skipped and the SRO is directly estimated. This estimate is used to resample the data of node S such that the SRO is reduced. Subsequently, a new SRO estimate is calculated between the resampled signal of S and the signal of R.

Signals are resampled using a sinc-interpolation where the temporally adjacent values within a window size of $(2 \cdot L_{\mathrm{sinc}} + 1)$ values are utilized with

$$x'_S(m) \approx \sum_{n=n'-L_{\mathrm{sinc}}}^{n'+L_{\mathrm{sinc}}} x_R(n) \; \operatorname{sinc}\left((1+\epsilon_{SR})m - n\right) \tag{19}$$

where n' is the index of the temporally closest sample in $x_R(n)$ to the newly interpolated m-th sample in $x'_S(m)$.

V. Beamforming

To extract the target signal, we employ a GEV beamformer [14]. The GEV beamformer can be derived by maximizing the SNR at the beamformer output for each frequency bin independently. This leads to the generalized eigenvalue problem [14]:

$$\mathbf{\Phi}_{\mathbf{XX}}(k)\, \mathbf{F}(k) = \lambda\, \mathbf{\Phi}_{\mathbf{NN}}(k)\, \mathbf{F}(k), \tag{20}$$

where the spatial correlation matrices $\mathbf{\Phi}_{\mathbf{XX}}(k)$ and $\mathbf{\Phi}_{\mathbf{NN}}(k)$ are estimated using time-frequency masks generated by a neural network [19], [20]. The mask estimation is rather unaffected by the sampling rate deviation, since it operates on each channel independently and does not make use of phases or phase differences. The network configuration and its training procedure is the same as described in [19].

The solution to Eq. (20) yields our desired beamforming vector $\mathbf{F}_{\mathrm{GEV}}(k)$, up to an arbitrary complex scale factor: Any magnitude and phase factor still solves Eq. (20), which is why the GEV beamforming approach is often said to introduce arbitrary distortions.

We therefore carefully addressed the magnitude degree of freedom by employing Blind Analytic Normalization (BAN) [14]. In earlier work we demonstrated that BAN has a great influence on perceptual quality and depending on the setup may effect Automatic Speech Recognition (ASR) performance as well [19] [21].

Finding a good constraint for the phase ambiguity is a bit more intricate. A first shot is to set the phase of a reference microphone $\tilde{d} \in \{1, \ldots, D\}$ to zero and adjust the other phases accordingly on each frequency independently:

$$F_d^{'}(k) = F_d(k) \cdot \exp(-\mathrm{j} \angle \{F_{\tilde{d}}(k)\}), \tag{21}$$

with $\mathbf{F}(k) = (F_1(k), \cdots F_D(k))^{\mathrm{T}}$. Intuitively, this is already much better than multiplying with a random phase.

An alternative is to minimize the group delay (rate of change between phase of subsequent frequency bins) introduced by the filter. To achieve this, we subtract the mean phase difference between two subsequent frequencies. To account for phase wrap, it is easier to do this by a multiplication with the complex conjugate of the phase factor corresponding to the phase difference:

$$\mathbf{F}^{'}(k) = \mathbf{F}(k) \cdot \exp\left(-\mathrm{j} \angle \left\{\mathbf{F}^{\mathrm{H}}(k-1)\mathbf{F}(k)\right\}\right). \tag{22}$$

VI. ASR Backend

The acoustic model is based on a Wide Residual Network [15]. It is the same network described in the context of the CHiME challenge [16] (network WRBN+BN). We omit the (speaker) adaptation due to restricted computational resources.

Previous papers also used a strong language model with RNN rescoring and tuning of language model weights and insertion penalties. For the given paper, however, we only use the 3-gram language model provided by the WSJ corpus [22] with a fixed language model weight and refrain from RNN rescoring. The motivation is as follows: We are mostly interested in the impact on the acoustic model, and a strong language model may obfuscate or hide possible influences of the evaluation at hand on word error rates.

VII. Experiments

We conducted experiments on two different datasets. The first is a self-compiled dataset using special recording hardware (see [5] for a detailed description). In this setup two sensor nodes were connected to a common input signal and the hardware synchronization was modified in a way such that the sampling clock signal generator kept a predefined SRO between the nodes with an error of less than $0.15\,\mathrm{ppm}$. Utterances from the TIMIT corpus [23] were played and recorded with predefined SROs between $0\,\mathrm{ppm}$ and $100\,\mathrm{ppm}$. Subsequently, we added uncorrelated noise to the channels to realize different SNRs. We will refer to this data with the term "HW-Dataset" (1100 files per SNR, each of $30\,\mathrm{s}$ duration).

Secondly, we used the resampling techniques from Eq. (19) to modify recordings from the CHiME dataset. Each file and channel was resampled with a randomly chosen, individual SRO drawn from a uniform distribution in the range of $\pm 50\,\mathrm{ppm}$. This scenario simulates a spontaneous recording using an ad-hoc network of nodes, e.g., smartphones on a table. Thus the algorithms had to cope with short and medium length utterances ($1.2\,\mathrm{s}$ - $13\,\mathrm{s}$) in noisy environments ($\approx 3\,\mathrm{dB}$ SNR), without the possibility to learn from consecutive files.

To characterize the degree of distortion with respect to the inter channel SROs we calculated the standard deviation of SROs σ_{SRO} in ppm between all six channels.

In our experiments we used a FFT size of $N = 8192$ with a shift in the Welch method of 1024 samples, and a temporal distance of $64\,\mathrm{ms}$ between the coherence functions.

A. HW-Dataset

In Fig. 3 sample results for the SRO estimation procedure are shown. The initial estimate of $77.16\,\mathrm{ppm}$ (1st stage) is used in the 2nd stage to resample the signal and perform a new estimate (new estimate is $95.29\,\mathrm{ppm}$). Each stage has a less ascending phase, since we reduce the SRO between the channels by resampling one of them. Additively combining

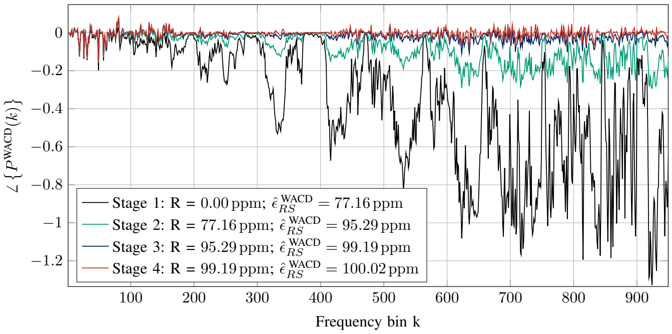

Fig. 3. HW-Dataset example: Multi-Stage SRO estimation experiment showing the phase estimates for different stages (ground truth 100 ppm); for each stage the resample factor R and the newly estimated SRO $\hat{\epsilon}_{RS}^{\mathrm{WACD}}$ are given.

the resampling factor and the current SRO estimate gives the final SRO estimate at each stage. Furthermore, we can observe that the variance of the phase estimate is reduced iteratively.

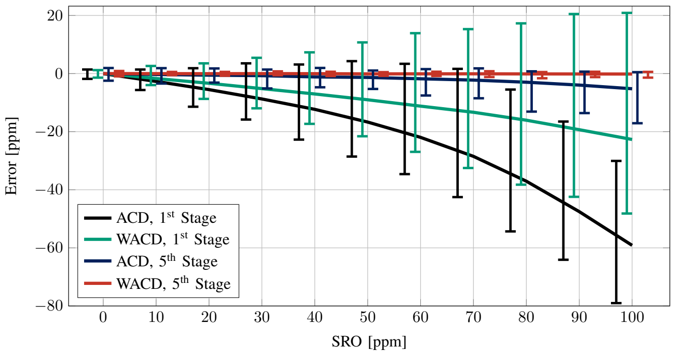

Fig. 4. Average SRO estimation errors for ACD and WACD at 1st and 5th stage with respect to SRO in recording (HW-Dataset, SNR 20 dB). Spans show minimum and maximum error.

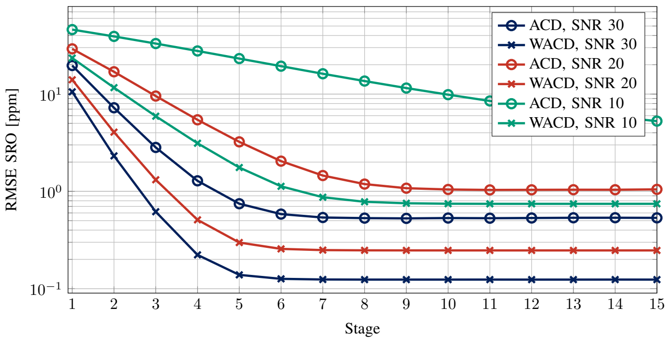

Fig. 5. Multi-Stage SRO estimation on HW-dataset database: Comparison of ACD and WACD for different SNR conditions with ground truth SROs between 0 and 100 ppm.

Eq. (24) and Eq. (25) indicate that the SRO estimation errors of the ACD and the WACD approach depend on the value of the SRO. This dependency can be seen in Fig. 4 for recordings with an SNR of 20 dB. Each stage reduces the SRO by resampling, which in turn reduces the bias and the variance of the estimator until the error remains on an equal level for all SROs. This lower limit is determined by the SNR and approximately independent of the SRO.

Fig. 5 shows the Root Mean Square Error (RMSE) for ACD and WACD on the HW-dataset for SNRs between 10 dB and 30 dB. Both approaches benefit from multi-stage resampling, but WACD constantly outperforms ACD in terms of RMSE.

B. CHiME Dataset

In Tab. I the WERs on the CHiME 6-channel real data evaluation test set for different SRO estimators are shown. Due to the fact that the SNR of the recordings is low (≈ 3 dB), the multi-stage approach is limited to a certain extent, however, it gains some remarkable improvements in the first 6 stages (see Fig. 6). The best results are achieved with the CORR approach from [8] using a fine-granular grid search.

TABLE I
WERs in [%] on CHiME database (eval. test set, real data) for different SRO estimators. Reference methods are our implementations of ACD from [3] and CORR from [8]

Beamformer Normalization	GEV-BAN -	 Phase	 RefMic	MVDR -	 Phase	 RefMic	σ_{SRO} [ppm]
No Sync.	9.57	9.26	10.02	9.44	8.87	9.68	25.68
ACD,1st Stage	8.45	7.93	8.46	8.49	7.80	8.17	18.34
ACD,10th Stage	7.17	6.65	6.88	7.41	6.73	7.02	7.63
ACD,15th Stage	7.26	6.70	6.87	7.36	6.73	7.05	7.35
WACD,1st Stage	7.55	7.14	7.65	7.81	7.06	7.45	13.81
WACD,10th Stage	7.30	6.71	7.08	7.43	6.72	6.99	6.71
WACD,15th Stage	7.03	6.56	6.77	7.40	6.61	6.92	6.63
CORR	6.80	6.38	6.62	7.28	6.52	6.62	6.29
No Offset	6.92	6.38	6.77	7.24	6.45	6.84	0

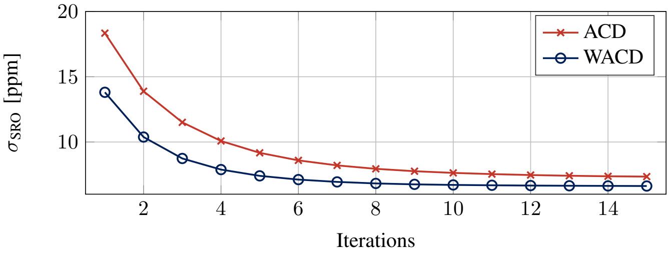

Fig. 6. Multi-Stage SRO estimation on CHiME database: Average standard deviation of SROs between resampled data streams.

The newly proposed phase normalization technique ("Phase"), which reduces the group delay, shows the best results under all SRO conditions and for both beamformers. The normalization according to a reference microphone ("RefMic") also improves the results compared to no phase normalization ("-"), but it is less effective than reducing the group delay. A non-zero SRO distracts the beamformer and causes higher WERs. ACD and WACD both reduce the inter-channel SROs, but again WACD delivers significantly better results.

VIII. Conclusions

We considered coherence drift based SRO estimation for WASN scenarios and advanced the existing ACD approach towards a Matched-filter like technique. Furthermore, the shortcomings of a key assumption in the derivation of the estimators are discussed and a multi-stage processing is proposed for mitigating its detrimental effects. Experiments on two databases show the effectiveness of the new approach in terms of SRO estimation precision and WERs. Additionally, we proposed a new phase normalization technique which is applicable to beamformers computing an ATF via eigenvector decomposition, such as the GEV and MVDR beamformer. It improves the WERs on the CHiME corpus significantly, both on synchronized recordings and on SRO distorted ones.

Acknowledgment

This work was supported by DFG under contract no <SCHM 3301/1-1> within the framework of the Research Unit FOR2457 "Acoustic Sensor Networks". We thank our students J. Deuse-Kleinsteuber, S. Fard, P. Hanebrink and S. Kosfeld for their practical work supporting this paper.

$$\Gamma_{R,S}(l,k) = \frac{\frac{H_R(k)H_S^*(k)}{\sqrt{|H_R(k)|^2}\cdot\sqrt{|H_S(k)|^2}}\left(\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa,k)|^2 \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}\right)\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}[\tau_{RS}+(\frac{N}{2}+lB)\epsilon_{RS}]k}}{\sqrt{\left(\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa,k)|^2+\frac{|V_R(l+\kappa,k))|^2}{|H_R(k)|^2}\right)\cdot\left(\sum_{\kappa=0}^{N_W-1}|S_S(l{+}\kappa,k)|^2+\frac{|V_S(l+\kappa,k))|^2}{|H_S(k)|^2}\right)}} \tag{23}$$

$$\frac{\Gamma_{R,S}(l{+}p,k)}{\Gamma_{R,S}(l,k)} = \frac{\overline{|X_{RS}(l,k)|^2}}{\overline{|X_{RS}(l{+}p,k)|^2}}\cdot\frac{\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa,k)|^2\ \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa{+}p,k)|^2\ \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(pBk)\epsilon_{RS}} \tag{24}$$

$$\Gamma_{R,S}(l{+}p,k)\cdot\Gamma_{R,S}^*(l,k) = \frac{\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa{+}p,k)|^2\ \mathrm{e}^{+\mathrm{j}\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\overline{|X_{RS}(l{+}p,k)|^2}}\frac{\sum_{\kappa=0}^{N_W-1}|S_R(l{+}\kappa,k)|^2\ \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\overline{|X_{RS}(l,k)|^2}}\mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(pBk)\epsilon_{RS}} \tag{25}$$

References

[1] 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, 2010.

[2] 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.

[3] 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.

[4] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A gossiping aproach to sampling clock synchronization in wireless acoustic sensor networks," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2014)*, 2014, pp. 7625–7629.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] M. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015.

[11] B. Van Veen and K. Buckley, "Beamforming techniques for spatial filtering," *Digital Signal Processing Handbook*, pp. 61–1, 1997.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[13] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 5210–5214, 2016.

[14] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, july 2007.

[15] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," *CoRR*, vol. abs/1605.07146, 2016.

[16] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition," in *Computer Speech and Language*, 2016.

[17] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[18] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[19] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2015.

[20] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[21] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, 2017.

[22] J. Garofalo *et al.*, "CSR-I (WSJ0) complete," 2007.

[23] DARPA, "Timit, acoustic-phonetic continuous speech corpus," 1990.