Online Estimation of Sampling Rate Offsets in Wireless Acoustic Sensor Networks with Packet Loss

Aleksej Chinaev and Gerald Enzner
Ruhr-Universität Bochum, Germany
Institute of Communication Acoustics
{aleksej.chinaev, gerald.enzner}@rub.de

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

***Abstract*—Acoustic sensor networks with ad-hoc topology currently seek application for acoustic signal processing in place of traditionally compact microphone arrays. Wireless configuration of the network will, however, impose additional challenges of which the synchronization of autonomous clocks has already received considerable attention under the premise of lossless transmissions. Further constraining to low latency, a serious impact of packet loss is expected as a result of UDP protocol operating the network. Such packet loss will be detrimental not only to the function of multi-channel acoustic signal processing over the network, but may deteriorate already the necessary estimation of sampling rate offset (SRO) between ad-hoc sensor nodes from incomplete waveforms. To study this aspect, the paper firstly revisits formal statistical packet loss models based on first-order Markov chains and, secondly, introduces online-capable methods for blind SRO estimation including the recursive band-limited interpolation (RBI), an online form of weighted averaged coherence drift (WACD) and a double-cross-correlation processor (DXCP) in FFT-domain. For experimental investigation, we then embed statistical simulation of burst-packet loss in a range of packet-loss rates into otherwise real acoustic data. In this setting, WACD and DXCP methods with proposed modifications demonstrate effectiveness in terms of SRO estimation accuracy and synchronization performance.**

***Index Terms*—Wireless acoustic sensor network; sampling rate offset estimation; sampling time synchronization**

I. Introduction

In times of rising demand for digital communication and ongoing spread of wearable electronic devices, wireless multimedia sensor networks have attracted an elevated attention from the interdisciplinary research community [1]. If individual sensor nodes are equipped with microphones, such a network is usually referred to as wireless acoustic sensor network (WASN) [2]. Autonomous nature of sensor devices with independent clocks makes a rigid time synchronization of sensor signals necessary for diverse WASN applications aiming at coherent signal processing [3]. An essential component of clock synchronization is estimation of sampling-rate offset (SRO) existing between signals of any two ad-hoc sensor nodes [4]. Approaches for SRO estimation, which merely utilize the audio signals already available at the processing node for coherent signal processing, are typically categorized as blind methods [5]. To the best of our knowledge, all approaches for blind SRO estimation published so far were developed under the assumption of idealized lossless data transmission in WASN [5]–[20].

In WASN applications for speech communication, the acquired audio signals have to be broadcast over the network in real-time with a limited end-to-end delay. In such low-latency communication systems, the message-oriented User Datagram Protocol (UDP) is often used in the transport layer for fast signal delivery without any retransmission of lost data packets [21] due to various transmission impairments such as path loss, multipath distortion or RF interference. As reported in [22], wireless communication channels can suffer from a considerable amount of packet-loss rates even up to $50\,\%$. The lost signal may easily become harmful not only for speech enhancement but also for blind SRO estimation. Furthermore, the latter has to be executed in an online fashion. From all methods for blind SRO estimation, our study involves three methods suitable for online processing: recursive band-limited interpolation (RBI) [15], an online form of weighted averaged coherence drift (WACD) [17] and an online double-cross-correlation processor (DXCP) [20].

In our study of robustness of online SRO estimators against packet loss, we shall take into account an aspect reported in previous work [10], and confirmed by own experiments, that RBI shows better performance on real-world acoustic data in contrast to just simulation. We therefore rely on recordings in a real acoustic environment accomplished with a Raspberry-Pi network to deliver complete data. Controlled packet loss is then introduced artificially into the data by applying a generalized Gilbert-Elliott (GE) channel model [23], [24], which proved to be well applicable for wireless transmission channels with burst-loss characteristics [22], [25].

The remainder of this paper is organized as follows: In Sec. II, a prototype WASN with signal packet loss is expressed by means of the generalized GE model, which is revisited. Then, online-capable methods for blind SRO estimation, the time-domain RBI [15], an online form of the WACD algorithm [17] and the online DXCP in FFT domain [20], are briefly introduced in Sec. III. Sec. IV describes our acquisition of real-world acoustic data captured by WASN based on Raspberry Pi computers. The resulting performance of online SRO estimators under burst-packet loss are compared and discussed in Sec. V, before conclusions are drawn in Sec. VI.

II. Simulation of WASN Signals with Packet Loss

Fig. 1 depicts a two-sensor prototype WASN. The i-th sensor acquires a continuous-time microphone signal $y_i(t) = x_i(t) + v_i(t)$, $i \in \{1, 2\}$, to be sampled by an asynchronous analog-digital converter (ADC), where $v_i(t)$ is uncorrelated self-noise of respective microphones. Without loss of generality, node 1 is assumed to have a perfect ADC, which starts its recording immediately upon request and delivers a discrete-time signal $y_1[n] = y_1(T_1 \cdot n)$ with the nominal sampling-time interval $T_1 = 1/f_\mathrm{s}$ as of the nominal sampling rate f_s. In contrast, node 2 is equipped with an imperfect ADC which starts its operation with time offset $d \cdot T_1$ represented by sampling-time offset (STO) $d \in \mathbb{R}$ and has a slightly different sampling-time interval $T_2 = (1+\varepsilon) \cdot T_1$ characterized by small real-valued SRO[1] $\varepsilon \in \mathbb{R}$ and, thus, delivers a discrete-time signal $y_2[n] = y_2(T_2 \cdot n + T_1 \cdot d)$. Further, the asynchronous signals $y_i[n]$ are transmitted via unreliable traffic channels (ch-1 and ch-2) to a processing node 3 for SRO estimation and potentially further cooperative signals processing.

Funded by German Research Foundation (DFG) - Project 282835863.

[1] Note that ε is here termed SRO, since it similarly relates sampling frequencies as $f_2 = 1/T_2 = (1 - \varepsilon/(1+\varepsilon)) \cdot f_\mathrm{s} \approx (1-\varepsilon) \cdot f_\mathrm{s}$, if $|\varepsilon| \ll 1$.

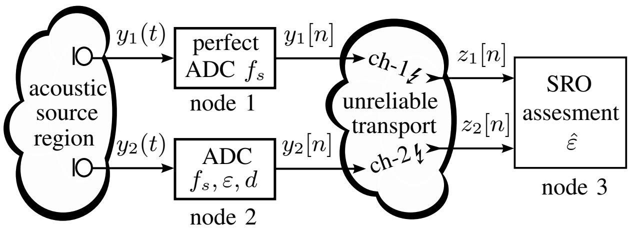


Fig. 1. Block diagram of prototype WASN under signal packet loss.

Due to the packet-based data transmission and randomly occurred impairments in both channels, some signal packets get lost in the signals $z_i[n]$ modelled with a binary mask here as follows:

$$z_i[\ell_p \cdot N_p + n_p] = m_i[\ell_p] \cdot y_i[\ell_p \cdot N_p + n_p]\,, \tag{1}$$

where $\ell_p \in \mathbb{Z}$ is the packet index, $n_p \in \{0, \ldots, N_p - 1\}$ is a subindex within one packet of length N_p samples, and $m_i(\ell_p) \in \{0, 1\}$ are the realizations of a binary random process $M_i(\ell_p)$. While $m_i(\ell_p) = 1$ implements a secure packet delivery over the i-th channel, a signal packet-loss is modelled with $m_i(\ell_p) = 0$ and no other packet-loss concealment takes place in our system.

The packet loss process $M(\ell_p)$ is often of bursty nature, i.e., it shows a mixture of variable-length packet loss, where one burst is defined as the loss of one or more consecutive data packets [22]. The packet loss distribution in such traffic channels is usually simulated using the generalized GE model [22]–[25], i.e., a first-order two-state hidden Markov chain as shown by Fig. 2. Specifically, the transition probabilities between the good (g) and bad (b) states of a hidden process $S_{\ell_p} \in \{g, b\}$ are defined as

$$p_{\text{gb}} = \Pr(S_{\ell_p} = \text{b} \,|\, S_{\ell_p - 1} = \text{g}) \tag{2}$$

$$p_{\text{bg}} = \Pr(S_{\ell_p} = \text{g} \,|\, S_{\ell_p - 1} = \text{b})\,, \tag{3}$$

while the conditional emission probabilities of packet loss in the good and bad states, respectively, can be written as

$$p_{\text{g}} = \Pr(M(\ell_p) = 0 \,|\, S_{\ell_p} = \text{g}) \tag{4}$$

$$p_{\text{b}} = \Pr(M(\ell_p) = 0 \,|\, S_{\ell_p} = \text{b}) \gg p_{\text{g}}\,. \tag{5}$$

Such GE model is thus controlled by four parameters and delivers a total probability of packet loss (packet loss rate) given by

$$p_{\text{pl}} = \Pr(M(\ell_p) = 0) = p_{\text{g}} \cdot \underbrace{\frac{p_{\text{bg}}}{p_{\text{gb}} + p_{\text{bg}}}}_{\Pr(S_{\ell_p} = \text{g})} + p_{\text{b}} \cdot \underbrace{\frac{p_{\text{gb}}}{p_{\text{gb}} + p_{\text{bg}}}}_{\Pr(S_{\ell_p} = \text{b})}\,. \tag{6}$$

In our investigations, Gilbert's assumptions of an error-free good state, $p_{\text{g}} = 0$ (no packet lost in the good state) and $p_{\text{b}} = 1$ (all packets lost in the bad state), result in the simplified Gilbert model controlled by merely two parameters p_{gb} and p_{bg}. The latter can then be calculated from two more intuitive parameters, i.e., a desired packet loss rate p_{pl} and a mean burst length[2] μ_{b} measured in seconds, via

$$p_{\text{bg}} = \frac{N_p}{f_{\text{s}} \cdot \mu_{\text{b}}} \quad \text{and} \quad p_{\text{gb}} = p_{\text{bg}} \cdot \frac{p_{\text{pl}}}{1 - p_{\text{pl}}}. \tag{7}$$

Note that the generalized GE model is adapted to variable packet sizes in [22] and its second order statistics are investigated in [25].

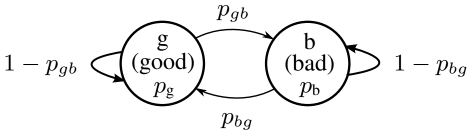


Fig. 2. The generalized Gilbert-Elliott Markov-channel model.

III. Online Methods for Blind SRO Estimation

Of the available methods for blind SRO assessment, the time-domain RBI [15] and the online (open-loop) DXCP in the FFT-domain [20] were developed for online processing. In order to improve robustness of the RBI algorithm and precision of the DXCP approach, both methods are modified in some processing steps compared to [15] and [20], respectively. The WACD method has been introduced in [17] as an offline estimator and is thus reformulated here for online processing on continuous audio streams.

A. Time-domain recursive band-limited interpolation (RBI)

In the RBI approach [15], maximization of an ambiguity function,

$$\hat{\alpha}^{\text{rbi}} = \underset{\alpha}{\text{argmax}} \sum_{\forall n} \widehat{Q}_n(\alpha, \tau)\,, \tag{8}$$

has to be solved, where $\alpha = 1 + \varepsilon$ denotes a scaling factor connected with the sought SRO ε, while τ is linked to the STO d. The function $\sum_{\forall n} \widehat{Q}_n$ is nothing but the discrete-time broadband correlation between two signals $z_i[n]$, where band-limited interpolation acts upon one of the signals $z_1[n]$,

$$\widehat{Q}_n(\alpha, \tau) = z_2[n] \sum_{m=p-L_0}^{p+L_0} z_1[m] \cdot \text{sinc}(\alpha n - \tau f_s - m) \tag{9}$$

with an anchor index $p = \lfloor \alpha n - \tau f_s \rfloor$, an interpolation window of size L_0, and the sinc-function $\text{sinc}(x) = \sin(\pi x)/(\pi x)$. Here, $\lfloor . \rfloor$ denotes to round down.

Given (9), the first and second order derivatives $q' = \partial \sum_{\forall n} \widehat{Q}_n / \partial \alpha$ and $q'' = \partial^2 \sum_{\forall n} \widehat{Q}_n / \partial \alpha^2$ can be analytically formed and a Newton-Raphson algorithm can be applied for a global maximum search. Furthermore, a stochastic approximation method can be utilized to obtain a sequential estimator for [15] with

$$\hat{\alpha}^{\text{rbi}}[n] = \hat{\alpha}^{\text{rbi}}[n-1] + \gamma \cdot (\widetilde{q''}[n])^{-1} \cdot \widetilde{q'}[n]\,, \tag{10}$$

$$\widetilde{q'}[n] = \beta \cdot \widetilde{q'}[n-1] + \widehat{q}'[n]\,, \tag{11}$$

$$\widetilde{q''}[n] = \beta \cdot \widetilde{q''}[n-1] + \widehat{q}''[n] \tag{12}$$

with stepsize γ=1, forgetting factor β, $\widehat{q}' = \partial \widehat{Q}_n / \partial \alpha$ and $\widehat{q}'' = \partial^2 \widehat{Q}_n / \partial \alpha^2$ as instantaneous derivatives at $\hat{\alpha}^{\text{rbi}}[n-1]$. From $\hat{\alpha}^{\text{rbi}}[n]$, the SRO estimate is obtained as $\hat{\varepsilon}^{\text{rbi}}[n] = \hat{\alpha}^{\text{rbi}}[n] - 1$.

In order to stabilize the convergence of (10), a damped Newton-Raphson method is used in our investigation with constant stepsize γ in contrast to [15]. Further, the update equation (10) is executed only if $|\gamma \cdot (\widetilde{q''}[n])^{-1} \cdot \widetilde{q'}[n]| < \varepsilon_{\text{ubi}}$ for limiting the corresponding SRO update to an upper bound per iteration (ubi) of ε_{ubi}.

It is important to mention that the RBI method has to be initialized in close proximity to the global maximum of the ambiguity function $\sum_{\forall n} \widehat{Q}_n(\alpha, \tau)$ by some good initial estimates $\hat{\varepsilon}^{\text{rbi}}[0]$ and $\hat{\tau}^{\text{rbi}}[0]$ as reported in [15]. Moreover, $\hat{\tau}^{\text{rbi}}[0]$ effectively comprises the aforementioned time offset $\hat{d} \cdot T_{\text{s}}$ and a time-difference of arrival (TDOA) τ_h between signals $z_i[n]$, the separation of which is not trivial and essentially requires external information of the TDOA.

In experimental evaluation, the following RBI parameters are used: $L_0 = 10$, $\beta = 0.999$, $\gamma = 10^{-6}$ and $\varepsilon_{\text{ubi}} = 1\,\text{ppm}$.

[2] The mean burst length μ_{b} in (7) is denoted by $\bar{b}$ in [22] and as the averaged burst error length (ABEL) parameter in [25].

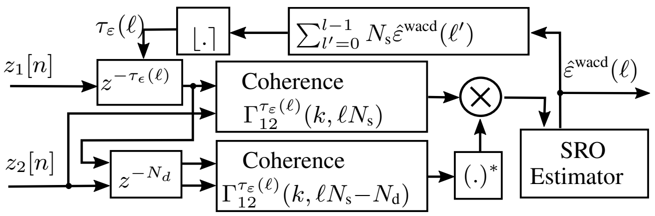

$\tau_\varepsilon(\ell)$
$\lfloor . \rceil$
$\sum_{l'=0}^{l-1} N_{\mathrm{s}}\hat{\varepsilon}^{\mathrm{wacd}}(\ell')$
$z_1[n]$
$z^{-\tau_\epsilon(\ell)}$
Coherence
$\Gamma_{12}^{\tau_\varepsilon(\ell)}(k,\ell N_{\mathrm{s}})$
$\hat{\varepsilon}^{\mathrm{wacd}}(\ell)$
$z_2[n]$
z^{-N_d}
Coherence
$\Gamma_{12}^{\tau_\varepsilon(\ell)}(k,\ell N_{\mathrm{s}}-N_{\mathrm{d}})$
$(.)^*$
SRO Estimator

Fig. 3. Online WACD implementation with tracking analysis window.

B. Online weighted averaged coherence drift (WACD)

In [17] we proposed a multi-stage coherence drift based sampling rate synchronization method (called WACD) that alternatingly estimates the SRO between two signals and resamples one of the signals until a convergence criteria is met. Here, we extend the basic idea towards an online method by incorporating the proposed stream processing from [26]. The key idea is to track the SRO-induced delay τ_ε between the signals and to adjust the analysis windows of the coherence estimators such that τ_ε is compensated to a large extent. This enables a continuous SRO estimation from data streams without the need for a resampling stage (see Fig. 3).

To this end, we have to change our approach from [17], i.e., a frame-oriented data addressing and processing, to a sample precise addressing and add an adjustable shift parameter τ. The required cross and auto power spectral densities (PSD) $\Phi_{ij}^{\tau}(k,n)$ with $i,j \in \{1,2\}$ are estimated via a Welch method, where k denotes the frequency bin index and n defines the index of the first used sample. On total for the PSD $\Phi_{ij}^{\tau}(k,n)$ N_W samples from each stream, i.e., $z_i[n{-}\tau]\ldots z_i[n{+}N_W{-}1{-}\tau]$ and $z_j[n]\ldots z_j[n{+}N_W{-}1]$, are used:

$$\Phi_{ij}^{\tau}(k,n) = \frac{1}{\nu_w}\sum_{\kappa=0}^{\nu_w-1} Z_i(k,n{+}\kappa N_w{-}\tau)\cdot Z_j(k,n{+}\kappa N_w)^*. \quad (13)$$

Hereby, $Z_j(k,n{+}\kappa N_w)$ denotes the N-point short-time Fourier transform (STFT) of the samples $z_j[n{+}\kappa N_w]\ldots z_j[n{+}\kappa N_w{+}N{-}1]$ and $(.)^*$ is the complex conjugate. So the method averages across $\nu_w{=}\lfloor(N_W{-}N{+}N_w)/N_w\rfloor$ STFT subframes for a Welch shift N_w.

Given the signal streams $z_1[n]$ and $z_2[n]$ every N_{s} samples the coherence function product $P(k,\ell N_{\mathrm{s}},\tau)$ for every ℓ-th frame is calculated via

$$P(k,\ell N_{\mathrm{s}},\tau) = \Gamma_{12}^{\tau}(k,\ell N_{\mathrm{s}})\cdot\Gamma_{12}^{\tau}(k,\ell N_{\mathrm{s}}{-}N_{\mathrm{d}})^* \quad (14)$$

with a temporal distance of N_{d} samples between the coherence functions that are defined by:

$$\Gamma_{12}^{\tau}(k,n) = \Phi_{12}^{\tau}(k,n)/\sqrt{\Phi_{11}^{0}(k,n-\tau)\cdot\Phi_{22}^{0}(k,n)}. \quad (15)$$

The temporal shift τ in (15) should follow the SRO induced delay τ_ε between the streams, which can be approximated by accumulating the SRO estimates $\hat{\varepsilon}^{\mathrm{wacd}}(\ell)$ over time. Since $\hat{\varepsilon}^{\mathrm{wacd}}(\ell)$ is estimated every ℓN_{s} samples also $\tau_\varepsilon(\ell)$ is estimated frame-wise via

$$\tau_\varepsilon(\ell) = \left\lfloor \sum_{\ell'=0}^{\ell-1} N_{\mathrm{s}}\,\hat{\varepsilon}^{\mathrm{wacd}}(\ell') \right\rceil, \quad (16)$$

where $\lfloor . \rceil$ denotes rounding to the next integer value.

SRO estimates $\hat{\varepsilon}^{\mathrm{wacd}}(\ell)$ are gathered by considering the SRO-induced delays $\tau_\varepsilon(\ell)$ and accumulating $P(k,\ell N_{\mathrm{s}},\tau)$ from a synchronization time instance ($\ell'=0$), i.e., the time after compensating the initial time offset $d\cdot T_1$, to the current time ($\ell'=\ell$) with

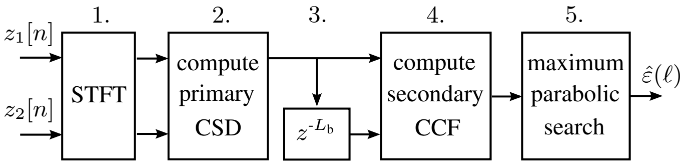


Fig. 4. Online DXCP in the FFT domain for open-loop architecture [20].

$$P^{\mathrm{wacd}}(k,\ell N_{\mathrm{s}}) = \frac{1}{\ell}\sum_{\ell'=0}^{\ell-1} P(k,\ell' N_{\mathrm{s}},\tau_\varepsilon(\ell')). \quad (17)$$

If rapid SRO changes are expected, the averaging period has to be reduced appropriately.

The slope of the phase from (17) can be used to estimate the SRO in different ways. As proposed in [17] we project the values $P^{\mathrm{wacd}}(k,\ell N_{\mathrm{s}})$ into the complex plane for averaging across the frequency bins and subsequently estimating $\hat{\varepsilon}^{\mathrm{wacd}}(\ell)$ via

$$\hat{\varepsilon}^{\mathrm{wacd}}(\ell){=}\frac{\varepsilon_{\max}}{\pi}\angle\sum_{k=K_{\min}}^{K_{\max}}\left|P^{\mathrm{wacd}}(k,\ell N_{\mathrm{s}})\right|\exp\left(\frac{jN\angle P^{\mathrm{wacd}}(k,\ell N_{\mathrm{s}})}{2N_d k\varepsilon_{\max}}\right). \quad (18)$$

Here, $K_{\min}\ldots K_{\max}$ defines the regarded frequency bin range and $\varepsilon_{\max}$ is the maximum expected SRO.

In experiments, the parameters are set as follows: $N=2^{13}$, STFT with periodic Blackman window, $N_W=2^{14}$, $N_w=2^{10}$, $N_{\mathrm{s}}=2^8$, $N_d=2^{14}$, $\varepsilon_{\max}=400$ ppm, $K_{\min}=100$, $K_{\max}=1800$.

C. Online (open-loop) double-cross-correlation processor (DXCP)

The core of DXCP-based SRO estimation [5], [20] is a calculation of the so-called secondary cross-correlation function (CCF) denoted here by $\widetilde{\psi}_{12}(\ell,\lambda)$ at signal frame index ℓ and lag λ and defined as cross-correlation of two primary CCFs at a certain time distance.

For an online form of the secondary CCF $\widetilde{\psi}_{12}(\ell,\lambda)$ from [20], the STFT coefficients $Z_i(k,\ell)$ at frequency bin k are calculated with frame size N and frame shift N_{s} from asynchronous signals $z_i[n]$ in the first step as shown by Fig. 4. Next, a primary cross-spectral density (CSD) $\widetilde{\Phi}_{12}(k,\ell)$ is calculated recursively,

$$\widetilde{\Phi}_{12}(k,\ell){=}\alpha_1{\cdot}\widetilde{\Phi}_{12}(k,\ell{-}1){+}(1{-}\alpha_1){\cdot}Z_1(k,\ell){\cdot}Z_2^*(k,\ell)\ , \quad (19)$$

where α_1 is a smoothing constant and $\widetilde{\Phi}_{12}(k,0)=0$. With the most recent values of $\widetilde{\Phi}_{z_1 z_2}(k,\ell)$ stored in a circular buffer, the desired secondary CCF $\widetilde{\psi}_{12}(\ell,\lambda)$ is computed as N-point inverse FFT (IFFT) of a secondary CSD $\widetilde{\Psi}_{12}(k,\ell)$ for frames $\ell \geq L_{\mathrm{b}}$+1,

$$\begin{aligned}\widetilde{\psi}_{12}(\ell,\lambda)\ \overset{\text{IFFT}}{\circ\!\!-\!\!\bullet}\ \widetilde{\Psi}_{12}(k,\ell) &= \alpha_2\cdot\widetilde{\Psi}_{12}(k,\ell-1)+ \quad (20)\\ &+(1-\alpha_2)\cdot\widetilde{\Phi}_{12}(k,\ell)\cdot\widetilde{\Phi}_{12}^*(k,\ell-L_{\mathrm{b}})\ ,\end{aligned}$$

where α_2 is a second smoothing constant and $\widetilde{\Psi}_{12}(k,L_{\mathrm{b}})=0$. Finally, a real-valued maximum $\lambda_{\mathrm{p}}^{\max}(\ell)$ is obtained by parabolic interpolation over the secondary CCF $\widetilde{\psi}_{12}(\ell,\lambda)$ as in [5] to deliver the desired SRO estimate for frames $\ell \geq L_{\mathrm{b}}$+L_{c}+1 obtained via

$$\hat{\varepsilon}^{\,\mathrm{dxcp}}(\ell) = \frac{\lambda_{\mathrm{p}}^{\max}(\ell)}{N_{\mathrm{s}}\cdot L_{\mathrm{b}}}\ . \quad (21)$$

In experimental evaluation, the parameters of FFT-DXCP are set to $N=2^{13}$, $N_s=2^{11}$, $L_b=L_c=37$, $\alpha_1=0.5$ and $\alpha_2=0.99$.

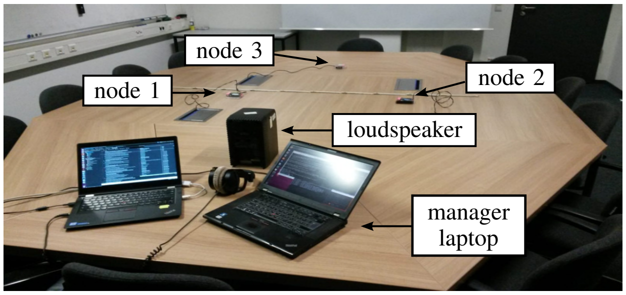


Fig. 5. Geometrical setup of WASN based on Raspberry Pi.

IV. Real WASN Data with Simulated Packet Loss

The WASN of Fig. 1 is implemented (with lossless transmission) on three Raspberry-Pi[3] (R-Pi) computers of model B+ by using a framework for networked signal processing called MARVELO [27]. As shown by Fig. 5, the R-Pi network is placed in a typical meeting room of size $(7\times5\times2.5)$ m with reverberation time $T_{60} \approx 700\,$ms. The network is controlled by an external laptop computer operating as manager. As acoustic excitation, speech signals are played back by a loudspeaker. Both node 1 and 2 are equipped with a ReSpeaker soundcard with a 4-microphone array of which only one microphone is used for signal acquisition at a nominal sampling rate $f_\mathrm{s} = 16\,$kHz. For purposes of investigation, we provide on node 2 a fast STFT resampling method [12] with FFT size 2^{10} as in [28] in order to create reproducible SROs $\varepsilon_p = (p-1)\cdot 20\,$ppm for $i \in \{1,\dots,P\}$ with $P = 6$. Single-microphone signals of node 1 and 2 are losslessly sent via TCP/IP over WLAN to node 3 and recorded there as asynchronous two-channel signals for further analysis on an extra PC. For every p-th SRO value, $R = 5$ different one-minute speech signals enumerated by r are acquired, resulting in overall 30 recordings.

Offline evaluation of lossless asynchronous recordings: In order to estimate an inherent SRO ε_inh between node 1 and 2 (common for all recordings) and the STO d_{pr} (specific for every recording) both required for evaluation of online methods for SRO estimation, the acquired signals are first analysed by the offline time domain DXCP (TD-DXCP) method implemented in multi-stage (MS) fashion [5]. In contrast to [5], TD-DXCP is extended here by STO estimation based on the real-valued maximum lag $\upsilon_{pr}^\mathrm{max}$ of an SRO-compensated primary CCF $\phi_{12}(\ell,\upsilon)$. Relying on symmetrical geometrical setups with TDOA $\tau_\mathrm{h} = 0$ in all recordings, the offline STO estimates are obtained as $\hat{d}_{pr}^\mathrm{off} = \upsilon_{pr}^\mathrm{max} - \tau_\mathrm{h} = \upsilon_{pr}^\mathrm{max}$. With TD-DXCP parameters set to $B{=}5f_\mathrm{s}$, $K{=}2B$, $\Upsilon{=}K-1$, $\Lambda{=}50$, periodic Blackman analysis window, and using the aforementioned STFT-resampler [12], the SRO and STO estimates $\hat{\varepsilon}_{pr}^\mathrm{off}$ and $\hat{d}_{pr}^\mathrm{off}$ of each recording are drawn after the 5-th MS iteration. From $\hat{\varepsilon}_{pr}^\mathrm{off}$, an estimate of the inherent SRO is obtained as $\hat{\varepsilon}_\mathrm{inh} = (\sum_{p=1}^{P}\sum_{r=1}^{R}\hat{\varepsilon}_{pr}^\mathrm{off} - \varepsilon_p)/(P\cdot R) = 3.056\,$ppm. Assessment of $\hat{d}_{pr}^\mathrm{off}$ reveals a large STO range of $[1300;4000]\,$smp with a mean of $2263.32\,$smp and a standard deviation of $708\,$smp.

Generation of burst-like packet loss: In our investigations, the signal packet length is set to $N_p = 256\,$smp ($16\,$ms for $f_\mathrm{s} = 16\,$kHz) typical for packed-based transmission of speech or generally audio signals [22]. The mean burst length of the GE model is chosen to $\mu_\mathrm{b} = 32\,$ms resulting in the transition probability $p_\mathrm{bg} = 0.5$ according to (7). The transition probabilities p_gb are further calculated via (7) from tentative packet loss probabilities in the range $p_\mathrm{pl} = (\upsilon-1)\cdot 5\%$ for $\upsilon \in \{1,\dots,V\}$ with $V = 11$. The burst-like packet loss is then simulated on both channels with independent loss patterns.

V. Experimental Evaluation

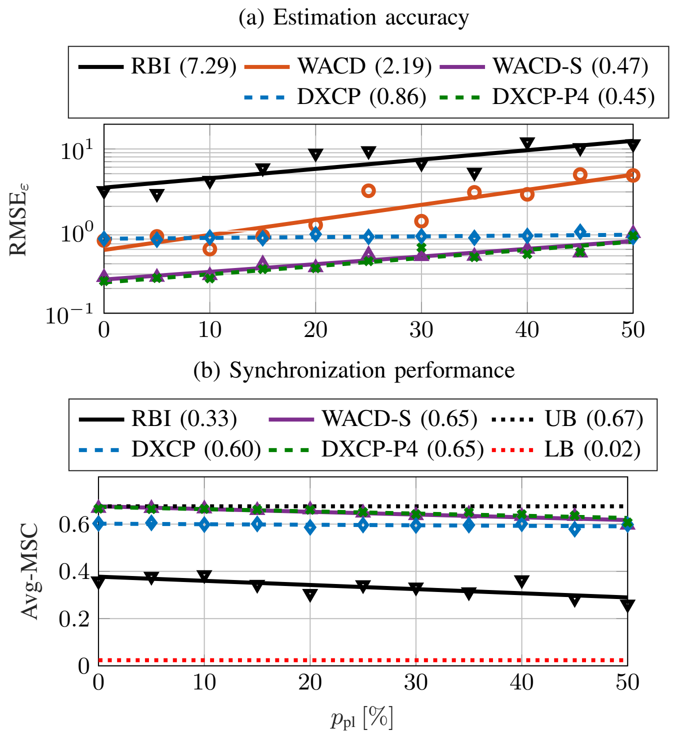
The online methods for blind SRO estimation of Sec. III take part in our evaluation and deliver SRO estimates $\hat{\varepsilon}_{pr}$ specific to each SRO p and recording r. The RBI method is supported oracle with an offline STO estimate $\hat{d}_{pr}^\mathrm{off}$ obtained on lossless recordings and initialized with a good SRO estimate of online DXCP available after $10\,$sec. The WACD approach makes use of GCC-PHAT [29] on the first N_W samples of the recordings to estimate the STO. To show the impact of STO synchonization on WACD, the experiments include results for WACD without and with STO compensation denoted by WACD and WACD-S, respectively. Further, a modification of the DXCP method (denoted by DXCP-P4) is implemented by using phase transform (PHAT) normalization from [29] in (19) followed by an additional 4-fold upsamling of the secondary CCF $\widetilde{\psi}_{12}(\ell,\lambda)$ prior to parabolic interpolation used for maximum search.

Fig. 6 (a) then shows SRO root mean-square error (RMSE) values

$$\mathrm{RMSE}_\varepsilon = \sqrt{\frac{1}{PR}\sum_{p,r}\left(\hat{\varepsilon}_{pr}(\infty) - \varepsilon_p - \hat{\varepsilon}_\mathrm{inh}\right)^2} \tag{22}$$

of online estimations taken at the end of the signals [30]. The DXCP method seems to be robust to packet loss in the considered range. The RMSE values of other methods grow with increasing packet-loss rates p_pl. Obviously, WACD-S and DXCP-P4 achieve the best estimation accuracy and deliver the lowest global RMSE values of $0.47\,$ppm and $0.45\,$ppm, respectively, as shown by the legend.

In order to evaluate the synchronization performance, the online SRO estimates and the offline STO estimates $\hat{d}_{pr}^\mathrm{off}$ are used for STO compensation and resampling [31]–[33] of the lossless recordings $y_2[n]$ resulting in a signal $y_{2,\mathrm{sync}}[n]$ synchronized to $y_1[n]$. The resulting values of an averaged magnitude-squared coherence (MSC) obtained in the style of a Welch average across the entire synchro-


(a) Estimation accuracy

(b) Synchronization performance

Fig. 6. Performance evaluation of online SRO estimation: (a) SRO estimation accuracy, (b) Synchronization performance on signals $y_1[n]$ and $y_{2,\mathrm{sync}}[n]$.

nized signals $y_1[n]$ and $y_{2,\text{sync}}[n]$ are depicted in Fig. 6 (b), where results of WACD are omitted. Although all online methods improve the raw MSC of the asynchronous signals $y_1[n]$ and $y_2[n]$ occurring here as a lower bound (LB), the two refined methods WACD-S and DXCP-P4 perform best among all contenders in restoring or approaching a coherence upper bound (UB) based on resampling with oracle SRO and STO.

VI. Conclusions

In this paper we have described three online methods for blind SRO estimation to extend the previously published time-domain RBI method [15], the WACD algorithm [17] and the open-loop DXCP approach in FFT-domain [20]. In order to approach realistic conditions, our evaluation takes place on real-world recordings with simulated burst-packet loss typical for low-latency wireless acoustic sensor networks. To this end, appropriate signal models were introduced including the generalized Gilbert-Elliott channel model for controlled simulation of packet loss. It turns out that WACD with STO pre-synchronization of the input signals and DXCP with additional PHAT normalization of the involved primary cross-spectral density achieve the best performance in terms of estimation accuracy and synchronization performance under packet loss.

VII. Acknowledgement

The authors are grateful to member H. Afifi of the DFG research unit FOR-2457 "Acoustic Sensor Networks" for software support in form the MARVELO implementation.

References

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer networks*, vol. 51, no. 4, pp. 921–960, 2007.

[2] A. Bertrand, S. Doclo, S. Gannot, N. Ono, and T. van Waterschoot, *Special Issue on Wireless Acoustic Sensor Networks and Ad Hoc Microphone Arrays*, vol. 107-C, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, Feb. 2015.

[3] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symp. Commun. Veh. Technol.*, Nov. 2011, pp. 1–6.

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

[5] A. Chinaev, P. Thüne, and G. Enzner, "A double-cross-correlation processor for blind sampling rate offset estimation in acoustic sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 641–645.

[6] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Sept. 2008, pp. 1–4.

[7] 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. Int. Workshop on Acoustic Signal Enhancement*, Sept. 2012, pp. 1–4.

[8] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 674–678.

[9] S. Miyabe, N. Ono, and S. Makino, "Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.

[10] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Proc. Int. Workshop on Acoustic Signal Enhancement*, Sept. 2014, pp. 183–187.

[11] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization," in *Proc. of Eur. Signal Process. Conf.*, Aug. 2015, pp. 245–249.

[12] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, no. Suppl. C, pp. 185 – 196, Sept. 2015.

[13] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks," in *Proc. Eur. Signal Process. Conf.*, Aug 2015, pp. 2281–2285.

[14] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 571–582, Mar. 2016.

[15] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

[16] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 674–686, Mar. 2017.

[17] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *Proc. Int. Workshop on Multimedia Signal Process.*, Oct. 2017, pp. 1–6.

[18] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 785–789.

[19] K. Itoyama and K. Nakadai, "Synchronization of microphones based on rank minimization of warped spectrum for asynchronous distributed recording," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Oct. 2020, pp. 4842–4847.

[20] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, June 2021, pp. 801–805.

[21] Internet Engineering Taskforce, "RFC 768 User Datagram Protocol," 1980.

[22] F. Mertz, *Efficient audio communication over heterogeneous packet networks with wireless access*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule (RTWH) Aachen, 2011.

[23] E. N. Gilbert, "Capacity of a burst-noise channel," *The Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, Mar. 1960.

[24] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, Apr. 1963.

[25] G. Haßlinger and O. Hohlfeld, "The Gilbert-Elliott model for packet loss in real time services on the internet," in *Proc. of GI/ITG Conf.-Measurement, Modelling and Evalutation of Computer and Comm. Systems*. VDE, Mar. 2008, pp. 1–15.

[26] J. Schmalenstroeer and R. Haeb-Umbach, "Insights into the interplay of sampling rate offsets and MVDR beamforming," in *Proc. ITG Conf. Speech Commun.*, Oct. 2018, pp. 246–250.

[27] H. Afifi, J. Schmalenstroeer, J. Ullmann, R. Haeb-Umbach, and H. Karl, "MARVELO - A framework for signal processing in wireless acoustic sensor networks," in *13th ITG Symp. on Speech Commun.*, Oct. 2018, pp. 311–315.

[28] A. Chinaev, G. Enzner, and J. Schmalenstroeer, "Fast and accurate audio resampling for acoustic sensor networks by polyphase-Farrow filters with FFT realization," in *Proc. ITG Conf. Speech Commun.*, Oct. 2018, pp. 96–100.

[29] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[30] P. Thüne and G. Enzner, "Tracking theory of adaptive filters with input-output sampling rate offset," in *Proc. of Eur. Signal Process. Conf.*, Sept. 2019, pp. 1–5.

[31] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, second edition, 1999.

[32] A. Chinaev, P. Thüne, and G. Enzner, "Low-rate Farrow structure with discrete-lowpass and polynomial support for audio resampling," in *Proc. Eur. Signal Process. Conf.*, Sept. 2018, pp. 475–479.

[33] J. Schmalenstroeer and R. Haeb-Umbach, "Efficient sampling rate offset compensation - an Overlap-Save based approach," in *Proc. of Eur. Signal Process. Conf.*, Sept. 2018, pp. 499–503.