A DOUBLE-CROSS-CORRELATION PROCESSOR FOR BLIND SAMPLING RATE OFFSET ESTIMATION IN ACOUSTIC SENSOR NETWORKS

Aleksej Chinaev, Philipp Thüne, Gerald Enzner

Ruhr-Universität Bochum, Department of Electrical Engineering and Information Technology, 44780 Bochum, Germany
Email: {aleksej.chinaev, philipp.thuene, gerald.enzner}@rub.de

ABSTRACT

Signal synchronization in wireless acoustic sensor networks requires an accurate estimation of the sampling rate offset (SRO) inevitably present in signals acquired by sensors of ad-hoc networks. Although some sophisticated methods for blind SRO estimation have been recently proposed in this very young field of research, there is still a need for the development of new ideas and concepts especially regarding robust approaches with low computational complexity. We therefore propose a novel time-domain method based on the calculation of a double-cross-correlation function in this contribution. Experimental evaluation of the introduced approach in a challenging acoustic environment and comparison with a state-of-the-art frequency-domain method confirms the high accuracy and low computational load of the proposed technique.

Index Terms— Wireless acoustic sensor networks, Blind sampling rate offset estimation, Wideband correlation processing

1. INTRODUCTION AND RELATION TO PRIOR WORK

Audio acquisition and signal processing via wireless acoustic sensor networks (WASN) may exhibit advantages compared to traditional microphone arrays since global sampling of the sound field generally results in a higher quality recordings [1]. The specific infrastructure of WASN for distributed sensing and classification of sound, however, requires new attention in order to meet the application-specific requirements of network self-calibration [2, 3], sound source localization [4–6], acoustic signal enhancement [7, 8], and acoustic scene classification [9–11]. Thus, sensor signals of ad-hoc networks have to be synchronized before their joint signal processing can be accomplished successfully. Commonly, time synchronization of acquired signals is performed by digital-to-digital arbitrary sampling rate conversion (ASRC) [12–15], which requires an estimation of SRO between sensor clocks.

A robust SRO estimation in WASN is a challenging task that often needs to be executed in a blind way by having access only to asynchronous signals without using any additional reference information [16]. Blind SRO estimation can be carried out either in the short-time Fourier transform (STFT) domain [16–21] or in the time domain [22, 23]. While approaches from [16, 19, 22] employ a time consuming exhaustive search over a grid of predefined SRO values, more efficient but still computationally intensive iterative procedures are used in [17, 21]. Further, ASRC methods are readily deployed in SRO estimation in order to obtain more precise estimates [16, 18, 19, 21]. While frequency-domain approaches often make use of the coherence function calculated from the cross power spectral density, estimators in the time domain employ a cross-correlation function in the estimation procedure. In our contribution, we develop a novel approach for blind SRO estimation by introducing a double-cross-correlation function that can be used for robust, precise and low-cost SRO estimation.

The contribution[1] is organized as follows: Buildung upon preliminaries introduced in Sec. 2, a double-cross-correlation processor is proposed in Sec. 3. After comprehensive experimental evaluation described in Sec. 4, conclusions are drawn in Sec. 5.

2. PRELIMINARIES

Assuming a coherent speech source signal $s(t)$ spread out in a reverberant setting and acquired by two sensors of an ad-hoc network, the respective microphone signals are given by

$$x_m(t) = h_m(t) * s(t) + v_m(t) \ , \tag{1}$$

where $m \in \{1, 2\}$ is the microphone index, t the continuous time, $h_m(t)$ an acoustic impulse response between the source and the m-th sensor, $v_m(t)$ a spatially uncorrelated noise of the m-th sensor and $*$ denotes linear convolution [23]. The discrete-time signals $x_m[n_m]$ result from time-sampling of the respective continuous-time signals $x_m(t)$ sampled at slightly different sampling rates $f_{s,m}$ via

$$x_m[n_m] = x_m(t_m)|_{t_m(n_m)=n_m \cdot T_m + d_m} \ , \tag{2}$$

where $n_m \in \mathbb{Z}$ is a discrete-time index, $T_m = 1/f_{s,m}$ a sampling time period and $d_m \in \mathbb{R}$ a real-valued delay in sampling start of the m-th sensor node. Considering microphone $m = 1$ without any loss of generality as the node with the reference sampling rate $f_{s,1} = f_s$, a sampling rate offset $\epsilon \in \mathbb{R}$ of the second sensor node can be defined as $\epsilon = f_{s,2}/f_s - 1$ and is assumed to be time-invariant for the scope of this publication.

As a result of asynchronous sampling in autonomous nodes, the sampling times $t_1(n_1)$ and $t_2(n_2)$ for the same indices (i.e., $n_1 = n_2 = n$) spread out increasingly across time. In order to describe this divergence across time, an accumulating time delay (ATD) can be defined as follows

$$\delta_{\mathrm{s}}(n) = \frac{t_1(n) - t_2(n)}{T_2} = \varepsilon \cdot n + \frac{d_1 - d_2}{T_2} \ , \tag{3}$$

where subscript 's' indicates sample-wise signal processing. According to (3), $\delta_{\mathrm{s}}(n)$ grows linearly with time index n for a time-invariant SRO ε, meaning that a difference of consecutive ATD values $\delta_{\mathrm{s}}(n) - \delta_{\mathrm{s}}(n-1) = \varepsilon$ can provide a foundation for the SRO estimation proposed in the following.

In frame-oriented signal processing, an ATD $\delta(\ell)$ for every signal frame $\ell \in \{1 \dots L\}$ can be defined as $\delta_s(n)$ evaluated at discrete

[1]This work has been supported by *Deutsche Forschungsgemeinschaft* (DFG) under grant EN 869/3-1 within the Research Unit FOR2457 "Acoustic Sensor Networks".

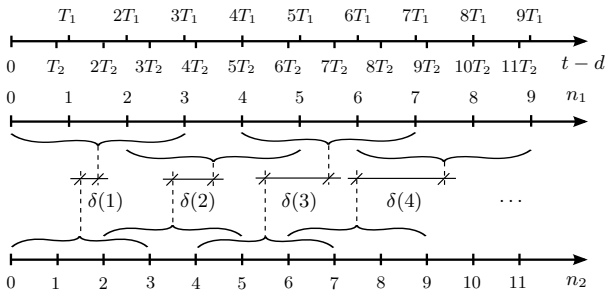

Fig. 1. Timing diagram with sampling times $t_m(n_m)$ for corresponding indices n_m from (2) and with frame ATDs $\delta(\ell)$ from (4) for $\varepsilon = 1/4$, $K = 4$, $B = 2$ and $d_1 = d_2 = d$.

time $n_c(\ell)$ corresponding to the center of the ℓ-th frame. Since an estimation of d_m is beyond the scope of this contribution, $d_1 = d_2 = d$ is assumed for further discussion. Under this assumption, the frame ATD $\delta(\ell) = \delta_s(n_c(\ell))$ can be calculated with (3) as

$$\delta(\ell) = \varepsilon \cdot n_c(\ell) = \varepsilon \cdot \left(B \cdot (\ell - 1) + \frac{K-1}{2} \right) , \qquad (4)$$

where K is the data frame size and B the accumulation time (also known as frame shift). A timing diagram with sampling times $t_m(n_m)$ and corresponding indices n_m from (2) and with respective frame ATD values $\delta(\ell)$ from (4) is shown in Fig. 1.

When asynchronous signals $x_1[n]$ and $x_2[n]$ are acquired, the presence of SRO effectively results in a temporal stretching of the wave form. By frame-oriented signal processing, the input signals $x_m[n]$ are split into overlapped windowed signal frames

$$x_{m,\ell}[k] = w[k] \cdot x_m[(\ell - 1) \cdot B + k - 1] , \qquad (5)$$

with an analysis window $w[k]$ and a subindex $k \in \{1 \dots K\}$. In this case, the aforementioned signal stretching produces a drift $\delta(\ell)$ between corresponding signal segments $x_{1,\ell}[k]$ and $x_{2,\ell}[k]$. To estimate $\delta(\ell)$ from $x_{1,\ell}[k]$ and $x_{2,\ell}[k]$ in time domain, the maximization of a cross-correlation function (CCF) is often recommended [24], which can be defined as

$$\phi_{12}(n, \upsilon) = \mathrm{E}\left\{ X_1[n] \cdot X_2[n + \upsilon] \right\} , \qquad (6)$$

where E denotes the expectation operator, $X_m[n]$ are random processes of the corresponding signals $x_m[n]$ and υ is a lag index. Note, $\phi_{12}(n, \upsilon)$ is modelled here as a time-variant statistic of the second order moving with time n over the lag-axis υ with a constant 'velocity' ε according to (3). Because of the finite observation time within one signal frame, $\phi_{12}(n, \upsilon)$ can only be estimated from signal segments of the ℓ-th frame via

$$\hat{\phi}_{12}(n_c(\ell), \upsilon) = \sum_{k=1}^{K} x_{1,\ell}[k] \cdot x_{2,\ell}[k + \upsilon] , \qquad (7)$$

where $\upsilon \in \{-\Upsilon, \dots, \Upsilon\}$ is a lag index of the CCF with a maximum lag index $0 < \Upsilon \leq K - 1$. However, a straightforward estimate of the frame ATD $\hat{\delta}(\ell)$ calculated from the CCF of a single frame $\hat{\phi}_{12}(n_c(\ell), \tau)$ via maximization seems to be a very challenging task, if speech is used as stimulus in a reverberant environment with a big distance between acoustic nodes – a scenario often encountered in wireless acoustic sensor networks.

3. A DOUBLE-CROSS-CORRELATION PROCESSOR

Similar to the sample-wise model (3), a difference of consecutive ATDs (called also ATD step) in the frame-oriented model (4) is time-invariant, i.e.,

$$\delta_\Delta = \delta(\ell) - \delta(\ell - 1) = B \cdot \varepsilon . \qquad (8)$$

As one can see, the ATD step δ_Δ is directly linked with the underlying SRO value and, consequently, is of great interest for blind estimation of time-invariant SRO assuming that an estimate of δ_Δ from the observed signals $x_m[n]$ is available.

The proposed double-cross-correlation processor (DXCP) aims at a robust estimation of δ_Δ based on a calculation of a second CCF defined on the subsequent CCFs from (6) as follows:

$$\psi_{12}(\lambda) = \mathrm{E}\left\{ \Phi_{12}(n, \upsilon) \cdot \Phi_{12}(n - B, \upsilon + \lambda) \right\} , \qquad (9)$$

where $\Phi_{12}(n, \upsilon)$ is a random process of $\phi_{12}(n, \upsilon)$ and λ a lag index of the second CCF. In contrast to the time-variant first CCF $\phi_{12}(n, \upsilon)$ from (6), $\psi_{12}(\lambda)$ is modelled as a time-invariant stationary statistic of the fourth order, since δ_Δ is assumed to be time-invariant. In an application, the second CCF $\psi_{12}(\lambda)$ can be estimated from realizations of the first CCF calculated via (7) as

$$\hat{\psi}_{12}(\ell, \lambda) = \sum_{\upsilon=-\Upsilon}^{\Upsilon} \hat{\phi}_{12}(n_c(\ell), \upsilon) \cdot \hat{\phi}_{12}(n_c(\ell - 1), \upsilon + \lambda) \qquad (10)$$

for lag index $\lambda \in \{-\Lambda, \dots, \Lambda\}$ with a maximum lag index $0 < \Lambda \leq 2\Upsilon$. Further, $\hat{\psi}_{12}(\ell, \lambda)$ is normalized to its maximum value over lag λ via

$$\tilde{\psi}_{12}(\ell, \lambda) = \frac{\hat{\psi}_{12}(\ell, \lambda)}{\max_\lambda \left(\hat{\psi}_{12}(\ell, \lambda) \right)} , \qquad (11)$$

which turns out to be useful for robust SRO estimation.

Since the second CCF $\psi_{12}(\lambda)$ is meant to be ergodic, its normalized estimates $\tilde{\psi}_{12}(\ell, \lambda)$ can be averaged. The average across all past estimates $\frac{1}{\ell-1} \sum_{\ell'=2}^{\ell} \tilde{\psi}_{12}(\ell', \lambda)$ can be stated recursively as

$$\bar{\psi}_{12}(\ell, \lambda) = \frac{1}{\ell - 1} \cdot \tilde{\psi}_{12}(\ell, \lambda) + \frac{\ell - 2}{\ell - 1} \cdot \bar{\psi}_{12}(\ell - 1, \lambda) \qquad (12)$$

for $\ell \geq 2$. Note, averaging (12) aims at achieving a more pronounced maximum in the resulting averaged normalized second CCF (ANS-CCF) and in consequence a better estimate of δ_Δ.

In order to estimate a real-valued δ_Δ from the averaged second CCF $\bar{\psi}_{12}(\ell, \lambda)$ calculated only for integer values of lag index λ, we propose to use a second-order polynomial

$$f(\lambda_p) = a(\ell) \cdot \lambda_p^2 + b(\ell) \cdot \lambda_p + c(\ell) , \qquad (13)$$

with a real-valued argument λ_p for interpolation through three supporting points $\bar{\psi}_{12}(\ell, \lambda_p^{sp})$ for $\lambda_p^{sp} \in \{\lambda_{\max}(\ell) - 1, \lambda_{\max}(\ell), \lambda_{\max}(\ell) + 1\}$ with an integer-valued

$$\lambda_{\max}(\ell) = \operatorname*{argmax}_\lambda \ \bar{\psi}_{12}(\ell, \lambda) . \qquad (14)$$

The single maximum point $\lambda_{p,\max}(\ell)$ of $f(\lambda_p)$ calculated in the ℓ-th frame is then a desired estimate of the ATD step $\hat{\delta}_\Delta(\ell)$, from which an SRO estimate $\hat{\varepsilon}(\ell)$ can be simply deduced via (8), i.e.,

$$\hat{\varepsilon}(\ell) = \frac{\hat{\delta}_\Delta(\ell)}{B} = -\frac{b(\ell)}{2 \cdot B \cdot a(\ell)} . \qquad (15)$$

The proposed DXCP for a blind SRO estimation has 4 parameters to be set, K, B, Υ, and Λ, and can be summarized in the following five processing steps executed for every signal frame $\ell \geq 1$:

1. Windowed framing of input data: $x_m[n] \rightarrow x_{m,\ell}[k]$ as in (5)
2. Calculation of first CCF $\hat{\phi}_{12}(\ell, \upsilon)$ with (7)
3. Saving $\hat{\phi}_{12}(\ell, \upsilon)$ as $\hat{\phi}_{12}(\ell - 1, \upsilon)$ for the next frame

For $\ell = 1$, skip steps 4 and 5

4. Computation of second CCF $\hat{\psi}_{12}(\ell, \lambda)$ with (10), its normalization $\tilde{\psi}_{12}(\ell, \lambda)$ with (11) and its recursive averaging with (12) resulting in $\bar{\psi}_{12}(\ell, \lambda)$
5. Parabolic interpolation (13) through points $\bar{\psi}_{12}(\ell, \lambda_p^{sp})$ for $\lambda_p^{sp} \in \{\lambda_{\max}(\ell) - 1, \lambda_{\max}(\ell), \lambda_{\max}(\ell) + 1\}$ with $\lambda_{\max}(\ell)$ from (14) and computation of $\hat{\delta}_\Delta(\ell)$ and $\hat{\varepsilon}(\ell)$ estimates with (15)

4. EXPERIMENTAL EVALUATION

The evaluation section is split into three parts. While in the first part some properties of the proposed DXCP method are illustrated, the second part presents optimization of the data frame size parameter K. In the third part, the performance of the DXCP method is compared to the state-of-the-art average coherence drift (ACD) approach for blind SRO estimation [18]. Note, the accumulation time of the proposed DXCP method was set in all experiments to $B = K/2$.

Data for experimental evaluation were generated by simulation of an acoustic enclosure of size $4\,\text{m} \times 5\,\text{m} \times 3\,\text{m}$ with different reverberation times $T_{60} = \{100, 500, 1000\}$ms. Assuming a single static speaker placed at position $(1, 2, 1.8)$ or $(1, 2.5, 1.8)$ and two microphones at positions $(3, 4, 1.5)$ and $(3, 1, 1.5)$ with a distance between microphones of 3 meters. Clean speech signals for male and female speakers are taken from the TIMIT database [25] and concatenated to a total length of one/three minutes each. The room impulse responses between the speakers and the microphones are generated using the image source method [26]. As spatially uncorrelated microphone noise, *white* and *babble* noise signals from the signal processing information base (SPIB) data [27] are used to generate noisy signals at different signal-to-noise ratio (SNR) values of $\{10, 20, 30\}$dB. In all experiments, the reference sampling rate $f_s = 16\,\text{kHz}$ is used. The second microphone signal is resampled for SRO values $\varepsilon \in \{0, 20, 40, 60, 80, 100\}$ppm by a highly accurate SINC method described in [28], which uses the Hann windowed sinc function of length $N_w = 513$ samples with signal-to-interpolation-noise ratio (SINR) of $\sim 110\,\text{dB}$ [29, 30].

4.1. Insight into functionality of the proposed DXCP

In order to give a deeper insight into the functionality of the proposed DXCP method, typical CCFs are presented in Fig. 2 calculated from input signals of 3 minutes length for the source position $(1, 2, 1.8)$. The signals are generated for $T_{60} = 1000\,\text{ms}$ and distorted by *babble* noise at 10 dB. The second microphone signal exhibits an SRO of $\varepsilon = 60$ ppm. And since the data frame size of the proposed DXCP method was set here to $K = 16 \cdot 10^4$, meaning calculation of the first CCF with (7) over 10 seconds of input data justified later and leading to an accumulation time of $B = 8 \cdot 10^4$, a true value of ATD according to (8) is in this experiment $\delta_\Delta = 4.8$. The remaining DXCP parameters were set here to $\Upsilon = 600$ and $\Lambda = 50$.

The time behavior of the estimated first and the second normalized cross-correlation functions $\hat{\phi}_{12}(n_c(\ell), \upsilon)$ and $\tilde{\psi}_{12}(\ell, \lambda)$ calculated by using (7) and (11) are depicted in Fig. 2 (a) and Fig. 2 (b), respectively. Fig. 2 (a) confirms the assumed linear time drift of the first CCF caused by the SRO, which produces wave pattern with

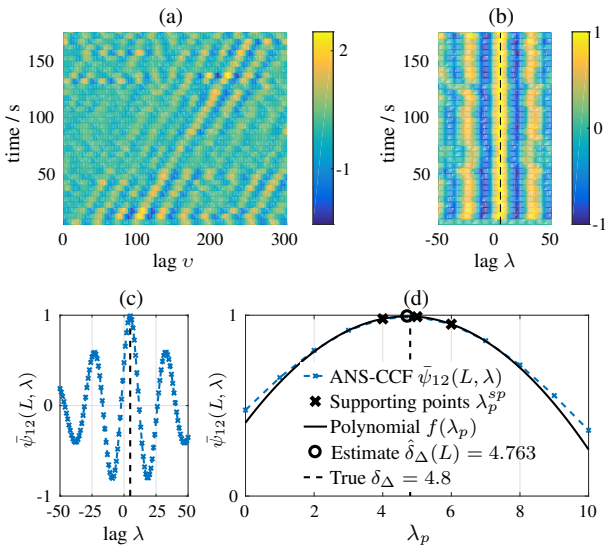

Fig. 2. Example of the proposed DXCP in a challenging acoustic environment: (a) Time drift of first CCF $\hat{\phi}_{12}(n_c(\ell), \upsilon)$ from (7), (b) Time course of normalized second CCF $\tilde{\psi}_{12}(\ell, \lambda)$ from (11), (c) Final averaged normalized second CCF $\bar{\psi}_{12}(L, \lambda)$ from (12), (d) Interpolation by polynomial (13) with resulting $\hat{\delta}_\Delta(L)$ estimate.

many more or less pronounced maxima. The latter move with constant velocity and accomplish an expected distance of 172.8 samples after 3 minutes. Since tracking of one of the maxima of the first CCF seems to be a very challenging task in such demanding acoustic environments as in our experiments, the second CCF is proposed to be used for an SRO estimation. As it is depicted in Fig. 2 (b), the normalized second CCF shows a better pronounced maximum compared to the first CCF and, even more important, does not exhibit any systematic time drift allowing averaging over time. The result of averaging (12) over $\tilde{\psi}_{12}(\ell, \lambda)$ calculated for all signal frames, $\bar{\psi}_{12}(L, \lambda)$, is depicted in Fig. 2 (c), whose maximum is located in the vicinity of the true ATD value. A zoom into the polynomial interpolation over the supporting points of $\bar{\psi}_{12}(L, \lambda)$ is shown in Fig. 2 (d) resulting in a good estimate $\hat{\delta}_\Delta(L) = 4.763$.

4.2. Optimization of data frame size parameter K

The data frame size K is a crucial DXCP parameter, which has to be set appropriately in order to support the SRO estimation. For this reason, a numerical optimization of K is carried out on a development data set generated for speaker position $(1, 2.5, 1.8)$ with $T_{60} = 500\,\text{ms}$. The development data set is created based on 100 clean speech signals of one minute length each distorted by *white* microphone noise at different SNRs for various SRO values resulting in 30 hours of two-channel speech material in total. While the parameter K is varied in the range $\{1\,\text{s}, \ldots, 19\,\text{s}\} \cdot f_s$ samples corresponding to signal segment lengths of $\{1, \ldots 19\}$ seconds, further parameters are set to $\Upsilon = K - 1$ and $\Lambda = 50$. As performance measures, a mean μ_ε and a standard deviation σ_ε of the estimation error $\hat{\varepsilon} - \varepsilon$ are used evaluated and depicted in Fig. 3.

According to Fig. 3 (a), the mean estimation error μ_ε decreases with increasing K values. Moreover, the absolute values of μ_ε grow with larger SRO values and the true SRO values are in general

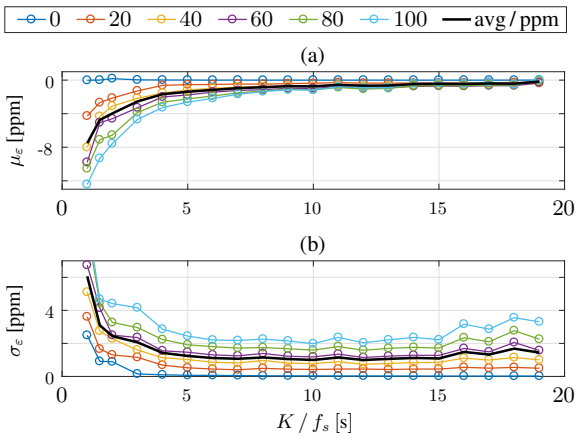

Fig. 3. Optimization of the data frame size K of the proposed DXCP from Sec. 3 on the development data set: (a) mean μ_e, (b) standard deviation σ_e of the estimation error $\hat{\varepsilon} - \varepsilon$.

slightly underestimated. Note, the latter property was registered also for some others SRO estimators [18, 21] and may be attributed to incoherent components available in both microphone signals. Calculation of CCFs over longer signal frame sizes ensures that these signal components are averaged out. In contrast to μ_e, the standard deviation in Fig. 3 (b) initially drops with K and slightly increases for bigger data frame sizes especially for larger SRO values. As observed in the experiments, the standard deviation of the proposed estimator achieves its minimum in the vicinity of $K = 10\,\mathrm{s} \cdot f_s$ corresponding to ten seconds of input data for every signal frame. Further, the evaluation reveals an important ability of the proposed estimator to benefit from lower SRO, which is proved by smaller values of μ_e and σ_e for $\varepsilon \to 0\,\mathrm{ppm}$, if the data frame size is set to values $K > 5\,\mathrm{s} \cdot f_s$. This feature enables the DXCP estimator to be deployed in an iterative multi-stage fashion – a procedure recently introduced in [16, 21]. The multi-stage technique is an iterative procedure, where the SRO-affected signal is resampled (for compensation of SRO value estimated in the previous iteration) before the 'remaining' SRO value is estimated in the next iteration leading in the course of the iteration process to a more precise final SRO estimate. The number of performed iterations I is equivalent here to the number of SRO estimations taking place.

As a result of the parameter optimization, $K_{\mathrm{opt}} = 10\,\mathrm{s} \cdot f_s$ is suggested as an appropriate choice for DXCP parameterization resulting in an averaged mean estimation error and standard deviation of 1 ppm for white microphone noise according to Fig. 3.

4.3. Comparison between the proposed DXCP and ACD

In order to compare the performance of the proposed DXCP method with the ACD approach, a test data set is generated for speaker position $(1, 2.5, 1.8)$ based on 40 speech signals (different from the development data) of one minute length each, distorted by *babble* microphone noise for different values of SNR and T_{60}, as mentioned before. The ACD approach from [18] is implemented with parameters recommended in [21]. Thus, the data frame size is set to $K_{\mathrm{ACD}} = 3 \cdot L_{\mathrm{W}}$ with a temporal distance between consecutive coherence functions of $P_{\mathrm{ACD}} = L_{\mathrm{W}}$ samples, where the Welch method is used for the estimation of the coherence function with FFT size $L_{\mathrm{W}} = 2^{13}$ and Welch shift $D_{\mathrm{W}} = 2^{10}$ [31]. The proposed DXCP is

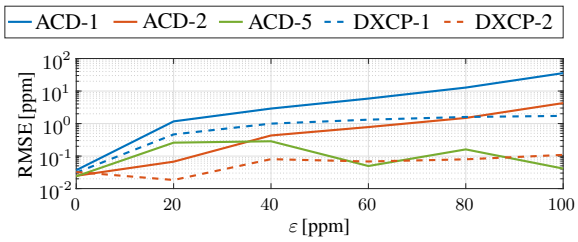

Fig. 4. RMSE values averaged over the test data set achieved by ACD and DXCP approaches implemented in single- and in multi-stage fashion for $I_{\mathrm{ACD}} \in \{1, 2, 5\}$ and $I_{\mathrm{DXCP}} = \{1, 2\}$, respectively.

implemented for $K = K_{\mathrm{opt}}$, $\Upsilon = K_{\mathrm{opt}}/4$ and $\Lambda = 50$. Every SRO estimator is implemented either in single- or in multi-stage fashion according to [21] in combination with SINC resampling, which uses the Hann-windowed sinc function of length $N_w = 33$.

While the precision of the estimators is measured on the test data by means of a root mean squared error (RMSE), a realtime factor (RF) is further considered to quantify computational complexity. The resulting RMSE and RF values averaged over the whole test data set are given in Tab. 1. The numbers reveal that the ACD method benefits strongly from the iterative multi-stage procedure, as it was already noticed in [21], and achieves its best performance on our test data for $I = 5$. In contrast to this, a single-stage implementation of the proposed DXCP obtains better accuracy than the corresponding ACD realization. And since it further benefits from the multi-stage technique, the proposed DXCP does not need more than one resampling step to reach much better estimates than ACD with 5 iterations (ACD-5). The RMSE values of both approaches plotted over SRO values are depicted in Fig. 4. It is striking here that estimation of large SRO values is a challenging task solved by the proposed DXCP with fewer iterations compared to the ACD method. Further, DXCP achieves better performance than the ACD approach for lower SRO values. However in the case of SRO absence, both approaches deliver similar RMSE values of $\sim 0.02\,\mathrm{ppm}$. The resulting RF values from Tab. 1 reveal a potential of the proposed approach to be implemented on portable devices with small computational power. Further, the proposed DXCP showed robustness towards different values of T_{60} and SNR.

Estimator	ACD				DXCP	
Iterations I	1	2	5	10	1	2
RMSE [ppm]	15.67	1.88	**0.17**	0.18	1.19	**0.07**
RF	0.005	0.128	**0.498**	1.115	0.006	**0.131**

Table 1. Averaged RMSE and RF values achieved on test data.

5. CONCLUSIONS

In this contribution, the double-cross-correlation processor DXCP has been proposed for blind, robust and accurate SRO estimation with reduced computational complexity. An experimental comparison of DXCP with a state-of-the-art ACD method in the STFT domain shows superiority of the proposed method in terms of better estimation accuracy at lower computational load. The latter is related to a smaller SRO underestimation bias per iteration, which in turn reduces the number of iterations required for a desired accuracy.

6. REFERENCES

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

[2] P. Pertilä, M. S. Hämäläinen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[3] M. Parviainen, P. Pertilä, and M. S. Hämäläinen, "Self-localization of wireless acoustic sensors in meeting rooms," in *IEEE Joint Workshop on Hands-free Speech Commun. and Microphone Arrays*, 2014, pp. 152–156.

[4] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process. Elsevier*, vol. 107, pp. 54–67, 2015.

[5] A. Brendel and W. Kellermann, "Localization of multiple simultaneously active sources in acoustic sensor networks using ADP," in *Proc. of Int. Workshop on Comput. Adv. in Multi-Sensor Adapt. Process.*, Dec. 2017, pp. 1–5.

[6] A. Brendel and W. Kellermann, "Learning-based acoustic source-microphone distance estimation using the coherent-to-diffuse power ratio," in *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Process.*, Apr. 2018.

[7] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP J. on Adv. in Signal Process.*, 2009.

[8] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process. Elsevier*, vol. 107, pp. 4–20, 2015.

[9] S. Gergen, A. M. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Process. Elsevier*, vol. 107, pp. 21–32, 2015.

[10] P. Arora and R. Haeb-Umbach, "A study on transfer learning for acoustic event detection in a real life scenario," in *Proc. of IEEE Int. Workshop on Multimedia Signal Process.*, Oct. 2017.

[11] J. Ebbers, A. Nelus, R. Martin, and R. Haeb-Umbach, "Evaluation of modulation-MFCC features and DNN classification for acoustic event detection," in *Proc. of Jahrestagung für Akustik (DAGA)*, Mar. 2018.

[12] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications (3. ed.)*, Prentice-Hall Int. Corp., 1996.

[13] G. Evangelista, "Design of digital systems for arbitrary sampling rate conversion," *Signal Process. Elsevier*, vol. 83, no. 2, pp. 377–387, Feb. 2003.

[14] H. G. Göckler and A. Groth, *Multiratensysteme: Abtastratenumsetzung und digitale Filterbänke*, Schlembach Fachverlag, 2004.

[15] P. Prandoni and M. Vetterli, *Signal Processing for Communications*, EPFL Press, 2008.

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

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

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

[19] 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. Elsevier*, vol. 107, pp. 185 – 196, Sept. 2015.

[20] 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. on Audio, Speech and Lang. Process.*, vol. 25, no. 3, pp. 674–686, Mar. 2017.

[21] 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. of Int. Workshop on Multimedia Signal Process.*, Oct. 2017.

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

[23] D. Cherkassky and S. Gannot, "Blind Synchronization in Wireless Acoustic Sensor Networks," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

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

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," Feb. 1993.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The J. of the Acoustic. Society Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[27] D. Johnson and P. N. Shami, "The signal processing information base," in *IEEE Signal Process. Mag.*, Oct. 1993, vol. 10, pp. 36–43.

[28] L. R Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Englewood Cliffs, N.J., Prentice-Hall, 1975.

[29] A. Chinaev, P. Thüne, and G. Enzner, "Low-Rate Farrow Structure with Discrete-Lowpass and Polynomial Support for Audio Resampling," in *Proc. of European Signal Process. Conf. (EUSIPCO)*, Sept. 2018.

[30] 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. of ITG Conference on Speech Communication*, Oct. 2018.

[31] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoustics*, vol. 15, no. 2, pp. 70–73, June 1967.