Blind Sampling Rate Offset Estimation for Wireless Acoustic Sensor Networks through Weighted Least-Squares Coherence Drift Estimation

Mohamad Hasan Bahari, Member, IEEE Alexander Bertrand, Member, IEEE Marc Moonen, Fellow, IEEE

Abstract

Microphone arrays allow to exploit the spatial coherence between simultaneously recorded microphone signals, e.g., to perform speech enhancement, i.e. to extract a speech signal and reduce background noise. However, in systems where the microphones are not sampled in a synchronous fashion, as it is often the case in wireless acoustic sensor networks, a sampling rate offset (SRO) exists between signals recorded in different nodes, which severely affects the speech enhancement performance. To avoid this

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC), KU Leuven Research Council Bilateral Scientific Cooperation Project Tsinghua University 2012-2014 (BIL 11/21T), the Interuniversity Attractive Poles Programme initiated by the Belgian Science Policy Office: IUAP P7/23 'Belgian network on stochastic modeling analysis design and optimization of communication systems' (BESTCOM) 2012-2017, Research Project FWO nr. G.0763.12 'Wireless Acoustic Sensor Networks for Extended Auditory Communication', Research Project FWO nr. G.0931.14 'Design of distributed signal processing algorithms and scalable hardware platforms for energy-vs-performance adaptive wireless acoustic sensor networks', the FP7-ICT FET-Open Project Heterogeneous Ad-hoc Networks for Distributed, Cooperative and Adaptive Multimedia Signal Processing (HANDiCAMS)', funded by the European Commission under Grant Agreement no. 323944, BOF/STG-14-005, iMinds Medical Information Technologies: SBO 2015. The scientific responsibility is assumed by its authors. A conference precursor of this manuscript has been published in [1]

M. H. Bahari is with the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT) and also with Sensifai, Brussels, Belgium (e-mail: bahari@sensifai.com).

A. Bertrand and M. Moonen are with the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven (e-mail: alexander.bertrand@esat.kuleuven.be;marc.moonen@esat.kuleuven.be).

performance reduction, the SRO should be estimated and compensated for. In this paper, we propose a new approach to blind SRO estimation for an asynchronous wireless acoustic sensor network, which exploits the phase-drift of the coherence between the asynchronous microphones signals. We utilize the fact that the SRO causes a linearly increasing time-delay between two signals and hence a linearly increasing phase-shift in the short-time Fourier transform domain. The increasing phase-shift, observed as a phase-drift of the coherence between the signals, is used in a weighted least-squares framework to estimate the SRO. This method is referred to as least-squares coherence drift (LCD). Experimental results in different real-world recording and simulated scenarios show the effectiveness of LCD compared to different benchmark methods. The LCD is effective even for short signal segments. We finally demonstrate that the use of the LCD within a conventional compensation approach eliminates the performance-loss due to SRO in a speech enhancement algorithm based on the multi-channel Wiener filter.

I. Introduction

Technological advances in micro-electronics and communications have paved the way towards novel acoustic sensing platforms, such as, e.g., wireless acoustic sensor networks (WASNs). WASNs consist of a multitude of wireless microphone nodes — each containing a single microphone or a small microphone array— distributed randomly over the environment. WASNs can be applied, e.g., for speech enhancement or to localize sound sources and extract spatial properties of the acoustic scenario in many applications such as teleconferencing, hands-free telephony, automatic speech recognition, monitoring and surveillance, video games and hearing aids [2]–[7]. However, the design of signal processing algorithms is more challenging for WASNs compared to traditional (wired) microphone arrays. It involves many different aspects such as dealing with unknown array geometries, routing, topology selection, synchronization and distributed processing [6], [8].

In a WASN, the fusion of microphone signals recorded in different nodes is a difficult task since each node utilizes an individual clock. Due to small imperfections in each clock's oscillator, sampling rate offsets (SROs) between signals recorded in different nodes are unavoidable [9], [10]. It has been shown that the existence of SROs severely degrades the performance of signal processing algorithms for Direction-of-arrival (DOA) estimation, speech enhancement and blind source separation [9]–[14]. In this paper, we only consider SRO estimation and compensation. However, it is noted that the use of a local clock at each node also results in sampling phase offsets, i.e., differences in the sampling time points at different nodes, or clock offsets, i.e., differences between the current time of the local clocks compared to a reference clock. Obviously, both of these phenomena are also influenced by the SRO, but they should be estimated in addition, e.g., to perform source localization.

The first step toward compensating for the effect of SRO and re-gaining the performance-loss consists of estimating the SRO. Two general approaches have been suggested to estimate the SRO. First, the SRO can be estimated based on a broadcasting of specific reference signals [11], [14]–[19]. For example, [15] addressed the time synchronization problem in wireless sensor networks (WSNs) by using a reference-broadcast synchronization algorithm to synchronize the clocks. The SRO estimation problem for acoustic beamforming in particular was tackled by [14], using a modulated radio frequency (RF) reference signal that is broadcast to each device. [16] used a reference signal to estimate the SRO between input and output channels in an echo cancellation system. However, SRO estimation based on a broadcasting of reference signals requires dedicated hardware, protocols, and/or communication channels. An alternative approach consists in using a reference-free ('blind') technique, where the SRO is directly estimated from the recorded microphone signals without using any reference signals. For example, [20] suggested a SRO estimation technique based on independent component analysis (ICA). In this method, it is assumed that ICA yields uncorrelated sources only when the SRO is perfectly compensated. However, to extract the independent components, the number of sources and microphones should be the same in this method. [10], [12] developed a method based on a maximum likelihood estimation of the SRO in the short-time Fourier transform (STFT) domain. In this method, the SRO is assumed to cause a linear phase-shift in the STFT domain and a likelihood function is derived to evaluate the compensation of the SRO. Then an exhaustive search method is applied to maximize the likelihood function to extract the SRO. This method is accurate and robust against environmental noise and can be applied in multiple source scenarios. However, it requires a stationary time-difference-of-arrival (TDOA) over long signal segments to yield an accurate SRO estimation, hence it is less applicable in turn-taking source scenarios. [13] used the link between SRO and the Doppler effect and applied a wideband correlation processor for blind SRO estimation. This method involves an exhaustive search over the SRO to maximize the wideband correlation processor and for each SRO in the applied search algorithm the signals are re-sampled in the time-domain.

[9] have also tackled the same problem using a voice activity detector (VAD) and phase-drift of the coherence of the noise-only segments in the signals, assuming the availability of a coherent noise source. The advantage of this method is its low computational complexity, i.e. the SRO is estimated without exhaustive search. However, it has a limited accuracy, it suffers from robustness issues and requires a VAD.

In this paper, we propose a new approach to blind SRO estimation without the need for a VAD and exhaustive search. Similar to [9] the proposed approach exploits the coherence phase-drift of the signals

and then applies a robust SRO estimation technique in a weighted least-squares (WLS) framework. The combination of the WLS and an outlier removal procedure allows to estimate the SRO even over signal segments with multiple active sources and scenarios with turn-taking sources. This paper extends our preliminary work [1] by (1) proposing a novel weighting (WG) scheme to emphasize on useful frequency bins, (2) evaluating the results over static and turn-taking source scenarios and (3) proving that the proposed method yields more accurate results compared to [9] in an equal condition.

Once the SRO is estimated with sufficient accuracy, the estimate can be used to synchronize the microphone signals. In [9], [13], after the estimation of the SRO, the signal is re-sampled in the time-domain using the Lagrange polynomials interpolation method [21]. While effective, this method is computationally expensive. In [10], [12] an explicit time-domain re-sampling is avoided, and instead a compensation for the SRO in the STFT domain is applied, assuming further processing is also applied in the STFT domain. We use a similar approach and validate our SRO estimation and compensation approach in a multi-channel Wiener-filter (MWF) based speech enhancement algorithm, where STFT-domain processing is used [22].

The rest of this paper is organized as follows. In Section II we formulate the SRO estimation problem. In Section III, we describe the proposed SRO estimation approach. In Section IV we briefly describe the applied SRO compensation approach. In Section V, we evaluate our approach in different real-world scenarios, and benchmark it against existing methods. Conclusions are drawn in Section VI.

II. Problem formulation

Without loss of generality (w.l.o.g.), we assume that each microphone belongs to a different node of the WASN, and hence there is an SRO between any microphone signal pair. The sound pressure of the i^{th} microphone and its corresponding discrete-time signal are written as $x_i(t)$ and $x_i[n]$, respectively, where t denotes the continuous time and n denotes the discrete time. The sampling rate of the i^{th} microphone is equal to

$$f_{s,i} = (1+\epsilon_i) f_s^{\text{ref}}, \tag{1}$$

where the parameter $|\epsilon_i| \ll 1$ is the relative SRO with respect to the reference sampling rate f_s^{ref} at an arbitrarily chosen reference node. W.l.o.g. we assume that the first node is the reference node, i.e. $f_{s,1} = f_s^{\text{ref}}$ and hence $\epsilon_1 = 0$. It is assumed that nodes i and node 1 are exchanging locally recorded signals, e.g., to perform multi-channel speech enhancement using MWF.

The goal is to estimate ϵ_i for a given microphone signal $x_i[n]$, and to compensate for its effect, e.g., within the computation of the MWF-based speech enhancement. The MWF is typically conducted in

the STFT domain to reduce the computational load, hence we aim for SRO compensation in the STFT domain. The ι^{th} segment $X_i^\iota[k]$ of the STFT of $x_i\,[n]$ is obtained as follows:

$$X_i^\iota[k] = \sum_{l=0}^{K-1} w[l] x_i \left[\iota P + l - \frac{K}{2}\right] \exp\left(-\frac{2\pi k l}{K} j\right), \tag{2}$$

where $j = \sqrt{-1}$, K is the STFT segment length, P is the STFT segment shift, $w[l]$ is a user-defined window function, and k is the discrete frequency index ranging from 0 to $K-1$.

Assuming $S_z^\iota[k]$ is the ι^{th} segment of the z^{th} source signal in the STFT domain, the i^{th} microphone signal $X_i^\iota[k]$ can be modelled as

$$X_i^\iota[k] = \sum_{z=1}^{Z} H_{i,z}^\iota[k] S_z^\iota[k] + n_i^\iota[k], \tag{3}$$

where $H_{i,z}^\iota[k]$ is the STFT domain transfer function from the z^{th} source to the i^{th} microphone in the ι^{th} segment, Z is the total number of coherent sources and $n_i^\iota[k]$ is the spatially uncorrelated noise component with $E\left[|n_i^\iota[k]|^2\right] = \sigma_i^2$. The coherent sources can be speech sources and/or (stationary) noise sources.

III. Least-squares coherence drift SRO estimation

In this section, we describe a new SRO estimation method, which is referred to as least-squares coherence drift (LCD).

A. Coherence

Consider the reference microphone signal $x_1\,[n]$ and the i^{th} microphone signal $x_i\,[n]$. The coherence of these signals within frame[1] m of length $\Gamma > K$ is obtained as

$$\Phi_{1,i}^m[k] = \frac{\Psi_{1,i}^m[k]}{\sqrt{\Psi_{1,1}^m[k]\Psi_{i,i}^m[k]}}, \tag{4}$$

where $\Psi_{1,i}^m[k]$ is the cross-spectrum between the microphone signals 1 and i, $\Psi_{1,1}^m[k]$ and $\Psi_{i,i}^m[k]$ denote the auto-spectrum of microphone signals 1 and i. We define m as the discrete time index of the mid-frame sample of the frame that is used to compute $\Phi_{1,i}^m[k]$. This means that $\Phi_{1,i}^m[k]$ and $\Phi_{1,i}^{m+1}[k]$ are defined over frames of length Γ that are shifted by only 1 sample. This is merely for the sake of notational convenience. In practice however, m will be incremented by $\Lambda >> 1$ samples to reduce the computational complexity.

[1]It is noted that a coherence frame is not the same as an STFT segment in (2).

The $\Psi_{q,p}^m$ can be estimated using the Welch method [23], which is a common method to estimate power spectral densities. To estimate $\Psi_{q,p}^m$, the Welch method chunks the m^{th} frame of length Γ into several overlapping segments of length K, and then takes the average of the cross-correlated STFT coefficients over the different segments. More specifically,

$$\Psi_{q,p}^m[k] = \frac{1}{N_{K\Gamma}} \sum_{\iota=1}^{N_{K\Gamma}} (X_q^\iota[m;k] X_p^\iota[m;k]^*), \tag{5}$$

where $X_i^\iota[m;k]$ is the STFT of the ι^{th} segment of signal $x_i[n]$ in the m^{th} frame, $(\cdot)^*$ denotes the conjugate transpose, and where $N_{K\Gamma}$ is the total number of overlapping segments of length K within a frame of length Γ.

By inserting (3) into (5) and assuming all sources are independent, we can write $\Psi_{1,i}^m[k]$ as

$$\Psi_{1,i}^m[k] = \sum_z \Psi_{1,i,z}^m[k], \tag{6}$$

where

$$\Psi_{1,i,z}^m[k] = \frac{1}{N_{K\Gamma}} \sum_{\iota=1}^{N_{K\Gamma}} (H_{1,z}[k])(H_{i,z}[k])^* \left|S_z^\iota[m;k]\right|^2, \tag{7}$$

where $S_z^\iota[m;k]$ is the ι^{th} STFT segment of the z^{th} source signal in the m^{th} coherence frame and $|\cdot|$ denotes absolute value.

It is noted that the acoustic transfer functions between the sources and the microphones are assumed to remain fixed over at least Γ samples, i.e., over the frame over which the cross-spectrum is computed, hence superscript ι is not used for $H_{z,1}[k]$ in (7).

B. Least-squares estimation

We exploit the phase-drift of the coherence over different frames to estimate the SRO. For the sake of an easy exposition, we first develop a least-squares (LS) estimation framework for a single source scenario and later we extend it to a multiple source scenario through a WLS estimation.

1) Single source scenario: The coherence $\Phi_{1,i}^m[k]$ is calculated by inserting (6) in (4) as follows

$$\Phi_{1,i}^m[k] = \frac{\sum_z \Psi_{1,i,z}^m[k]}{\sqrt{\sum_z \Psi_{1,1,z}^m[k] + \sigma_1^2}\sqrt{\sum_z \Psi_{i,i,z}^m[k] + \sigma_i^2}}, \tag{8}$$

For a single source scenario, we can replace $\sum_z \Psi_{1,i,z}^m[k]$ in (8) by $\Psi_{1,i,z}^m[k]$, i.e.

$$\Phi_{1,i}^m[k] = \frac{\Psi_{1,i,z}^m[k]}{\sqrt{\Psi_{1,1,z}^m[k] + \sigma_1^2}\sqrt{\Psi_{i,i,z}^m[k] + \sigma_i^2}}. \tag{9}$$

by expanding (9) using (7) and assuming $\Psi^m_{i,i,z}[k] >> \sigma_i^2$ we obtain

$$\Phi^m_{1,i}[k] = \frac{H_{1,z}[k](H_{i,z}[k])^* \sum_\iota |S^\iota_z[m;k]|^2}{|H_{1,z}[k]|\,|H_{i,z}[k]| \sum_\iota |S^\iota_z[m;k]|^2}, \tag{10}$$

or

$$\Phi^m_{1,i}[k] = \frac{H_{1,z}[k](H_{i,z}[k])^*}{|H_{1,z}[k]|\,|H_{i,z}[k]|}. \tag{11}$$

Assuming the transfer functions between the source and microphones remain unchanged, a fixed delay of ϱ_i samples to $x_i\,[n]$, affects the coherence phase as

$$\Phi^m_{1,i}[k;\varrho_i] = \frac{H_{1,z}[k](H_{i,z}[k])^* \exp\left(\frac{2\pi k \varrho_i}{K} j\right)}{|H_{1,z}[k]|\,|H_{i,z}[k]|}, \tag{12}$$

or

$$\Phi^m_{1,i}[k;\varrho_i] = \Phi^m_{1,i}[k] \exp\left(\frac{2\pi k \varrho_i}{K} j\right), \tag{13}$$

where $\Phi^m_{1,i}[k;\varrho_i]$ is the coherence between $x_1\,[n]$ and $x_i\,[n]$ after the latter is delayed by ϱ_i samples. Such a fixed delay usually occurs due to acoustic propagation delays, e.g. when the microphones are not equidistant from the source. However, note that these fixed delays are assumed to be unknown and are in principle absorbed within the two transfer functions $H_{1,z}[k]$ and $H_{i,z}[k]$.

An SRO between the microphone signals 1 and i also causes a linearly increasing delay in the time-domain, and hence a linearly increasing phase-shift in the coherence. The sample delay of the i^{th} microphone signal in the mid-frame sample (m) caused by the SRO ($\epsilon_i << 1$) w.r.t. microphone signal 1 is denoted as ρ_i^m, and can be computed as

$$\rho_i^m \;\; = f_s^{\text{ref}} \left[\tfrac{m}{f_s^{\text{ref}}} - \tfrac{m}{(1+\epsilon_i) f_s^{\text{ref}}}\right] \approx m\epsilon_i. \tag{14}$$

The SRO induced delay is equal to (14) for the mid-frame sample and equal to $(m-1)\epsilon_i$ and $(m+1)\epsilon_i$ for the sample before and after, etc. Since this delay increases for each consecutive sample in a frame, calculating the coherence $\Phi^m_{1,i}[k;\rho_i^m]$ of the reference signal and the signal with SRO is difficult. However, assuming the maximum drift caused by the SRO inside a single frame is much smaller than 1 sample, i.e. $|\Gamma\epsilon_i| \ll 1$, the coherence $\Phi^m_{1,i}[k;\rho_i^m]$ can be approximated as

$$\Phi^m_{1,i}[k;\rho_i^m] = \Phi^m_{1,i}[k] \exp\left(\frac{2\pi k\,(\rho_i^m)}{K} j\right).$$

To remove the phase-shift due to acoustic propagation, we use the phase difference between the coherence of two consecutive frames with frame-shift equal to Λ samples such that, relying on (11),

$$\begin{aligned} \angle \frac{\Phi^m_{1,i}[k;\rho_i^m]}{\Phi^{m-\Lambda}_{1,i}[k;\rho_i^{m-\Lambda}]} \;\; &\approx \tfrac{2\pi k\left(\rho_i^m - \rho_i^{m-\Lambda}\right)}{K} \\ &= \tfrac{2\pi k \Lambda}{K}\epsilon_i, \end{aligned} \tag{15}$$

where the $\angle$ denotes the phase (the last step follows from (14)). From (15), we observe that the phase difference between the coherence of two different frames with frame shift Λ increases linearly with the SRO.

Remark 1: The source signal $|S_z^{\iota}[m;k]|$ is cancelled out from the numerator and denominator in (10) and (11). Therefore, there is no stationarity assumption required on the source signal for (15) to hold.

To improve the estimation accuracy, we repeat the above procedure for $Q+1$ consecutive frames and collect the results in matrix form, i.e.

$$\mathbf{A} = \mathbf{B}\epsilon_i \tag{16}$$

where $\mathbf{A}$ is a matrix of size[2] $\lfloor \frac{K}{2} \rfloor \times Q$ with elements $a_{k,q}$

$$a_{k,q} = \angle \frac{\Phi_{1,i}^{m-q\Lambda}[k;\rho_i^{m-q\Lambda}]}{\Phi_{1,i}^{m-(q-1)\Lambda}[k;\rho_i^{m-(q-1)\Lambda}]} \tag{17}$$

and $\mathbf{B}$ is a matrix of dimension $\lfloor \frac{K}{2} \rfloor \times Q$ with elements $b_{k,q}$

$$b_{k,q} = \frac{2\pi k\Lambda}{K}, \tag{18}$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

A LS[3] estimation of ϵ_i can be obtained by solving

$$\hat{\epsilon}_i^{\mathrm{LS}} = \arg\min_{\epsilon_i} \|\vec{\mathbf{A}} - \vec{\mathbf{B}}\epsilon_i\|_2, \tag{19}$$

where $\|\cdot\|_2$ denotes the L_2-norm and $\vec{\cdot}$ denotes vectorization, where columns of a matrix are stacked on top of each other. The optimal solution of (19) can be obtained as follows:

$$\hat{\epsilon}_i^{\mathrm{LS}} = \frac{\vec{\mathbf{B}}^T\vec{\mathbf{A}}}{\vec{\mathbf{B}}^T\vec{\mathbf{B}}}, \tag{20}$$

where T denotes the transpose operator.

2) Multiple sources scenario: For the multiple sources scenario, we modify (19) by developing a WLS framework.

Although relation (15) is invalid for the multiple sources scenario (3), it still holds for frequency bins where at least one of the following conditions is met.

1- One of the sources is predominant for two consecutive frames, i.e.

$$\exists\, z \in \{1,\dots,Z\}:\ \Psi_{1,i}^{m}[k] \approx \Psi_{1,i,z}^{m}[k] \ \text{ and } \ \Psi_{1,i}^{m-\Lambda}[k] \approx \Psi_{1,i,z}^{m-\Lambda}[k] \tag{21}$$

[2]Since the second half of the STFT bins is just a mirror image of the first half, we use the first half without losing performance.

[3]Other distance measures can also be applied for this problem. The procedure of solving a similar estimation problem using KullbackLeibler divergence is explained in [24]

This condition is typically satisfied in the case of speech sources, due to their sparse nature in the time-frequency-domain. This is often exploited in speech processing, and has been empirically validated in several studies [25]–[27]. Meeting the predominant source condition turns the multiple sources scenario into a single source scenario in the majority of the time-frequency bins. Therefore, its validity is shown in relations (8)-(15).

2- All active sources are stationary for two consecutive frames, i.e.

$$\forall z \in \{1, \cdots, Z\}: \quad \bar{S}_z^m[k] = \bar{S}_z^{m-\Lambda}[k], \tag{22}$$

where

$$\bar{S}_z^m[k] = \sum_\iota \left|S_z^\iota[m;k]\right|^2. \tag{23}$$

This condition is commonly met in noise-only frames in scenarios with localized (i.e., coherent) stationary noise sources.

To show the importance of the second condition, let us start by expanding (8) using (7)

$$\Phi_{1,i}^m[k] = \frac{\sum_z H_{1,z}[k](H_{i,z}[k])^* \sum_\iota \left|S_z^\iota[m;k]\right|^2}{\sqrt{\sum_z \left|H_{1,z}[k]\right|^2 \sum_\iota \left|S_z^\iota[m;k]\right|^2 + \sigma_1^2}\sqrt{\sum_z \left|H_{i,z}[k]\right|^2 \sum_\iota \left|S_z^\iota[m;k]\right|^2 + \sigma_i^2}}, \tag{24}$$

or

$$\Phi_{1,i}^m[k] = \frac{\sum_z H_{1,z}[k](H_{i,z}[k])^* \bar{S}_z^m[k]}{\sqrt{\sum_z \left|H_{1,z}[k]\right|^2 \bar{S}_z^m[k] + \sigma_1^2}\sqrt{\sum_z \left|H_{i,z}[k]\right|^2 \bar{S}_z^m[k] + \sigma_i^2}}. \tag{25}$$

Coherence $\Phi_{1,i}^{m-\Lambda}[k]$ can be calculated as

$$\Phi_{1,i}^{m-\Lambda}[k] = \frac{\sum_z H_{1,z}[k](H_{i,z}[k])^* \bar{S}_z^{m-\Lambda}[k] \exp\left(-\frac{2\pi k \Lambda \epsilon_i}{K} j\right)}{\sqrt{\sum_z \left|H_{1,z}[k]\right|^2 \bar{S}_z^{m-\Lambda}[k] + \sigma_1^2}\sqrt{\sum_z \left|H_{i,z}[k]\right|^2 \bar{S}_z^{m-\Lambda}[k] + \sigma_i^2}}. \tag{26}$$

Assuming the second condition (22) is met, we replace $\bar{S}_z^{m-\Lambda}[k]$ by $\bar{S}_z^m[k]$, i.e.

$$\Phi_{1,i}^{m-\Lambda}[k] = \frac{\exp\left(-\frac{2\pi k \Lambda \epsilon_i}{K} j\right) \sum_z H_{1,z}[k](H_{i,z}[k])^* \bar{S}_z^m[k]}{\sqrt{\sum_z \left|H_{1,z}[k]\right|^2 \bar{S}_z^m[k] + \sigma_1^2}\sqrt{\sum_z \left|H_{i,z}[k]\right|^2 \bar{S}_z^m[k] + \sigma_i^2}}. \tag{27}$$

Relation (15) is obtained by dividing (25) with (27).

By inserting the following WG scheme into the proposed LS estimation problem (19), we decrease the deteriorating effect of these data points that do not meet any of the conditions mentioned above.

$$\hat{\epsilon}_i^{\mathrm{WLS}} = \arg\min_{\epsilon_i} \|\vec{\mathbf{A}}_{\mathrm{v}} - \vec{\mathbf{B}}_{\mathrm{v}}\epsilon_i\|_2 \tag{28}$$

$$\vec{\mathbf{A}}_{\mathrm{v}} = \vec{\mathbf{V}} \circ \vec{\mathbf{A}} \tag{29}$$

$$\vec{\mathbf{B}}_{\mathrm{v}} = \vec{\mathbf{V}} \circ \vec{\mathbf{B}}, \tag{30}$$

where $\circ$ denotes Hadamard product and $\mathbf{V}$ is a weighting matrix of dimension $\lfloor \frac{K}{2} \rfloor \times Q$ with rows $v_{k,q}$

$$v_{k,q} = \frac{\sqrt{\left(\left|\Phi_{1,i}^{m-q\Lambda}[k]\right| \left|\Phi_{1,i}^{m-(q-1)\Lambda}[k]\right|\right)^{\beta}}}{\exp\left(\left||\Phi_{1,i}^{m-q\Lambda}[k]| - |\Phi_{1,i}^{m-(q-1)\Lambda}[k]|\right|^2\right)}, \tag{31}$$

where β is a hyperparameter and can be tuned for different applications. The weight $v_{k,q}$ attains its global maximum ($v_{k,q} = 1$) if condition 1 (21) is satisfied as shown in Appendix A and any typical discrepancies from this condition decrease the weight.

For the frequency bins where the second condition (22) is satisfied, the denominator of (31) is also minimized (which result in a larger weight) as shown in Appendix B. To understand the motivation behind the numerator in this case, note that the denominator alone would result in a large weight even if there is a low coherence between the signals, i.e. both $\left|\Phi_{1,i}^{m}[k]\right|$ and $\left|\Phi_{1,i}^{m-\Lambda}[k]\right|$ are close to 0. To avoid this problem, the numerator is used to down-weight low-coherence frequency bins.

Remark 2: Please note that the WG scheme (31) assumes unchanged transfer functions between the sources and microphones during each consecutive frames. It is also noted that this WG scheme depends on the coherence amplitude only and completely ignores the coherence phase. This may cause undesired large weight in some cases such as a turn-taking scenario where the fixed transfer function assumption is violated. In this case, the coherence amplitudes can be similar (or even not affected) and instead the coherence phase abruptly changes. To avoid this problem an outlier removal (OR) procedure is proposed in the subsequent section which gives binary weights of zeros and ones to each frequency bin depending on the coherence phase.

3) Outlier Removal: Although the WG scheme (28)-(31) significantly improves the performance of the LS estimate, there are still many outlier frequency bins which have a non-negligible effect on the LS estimator. For example, (15) compares phases which are defined over a circular topology, i.e., a phase of π is the same as a phase $-\pi$. However, for phases that are close to this phase ambiguity boundary, small errors due to noise may result in large absolute differences, and then (28) may result in an inaccurate estimation of the SRO. Figure 1 shows an example of the coherence phase drift in different frequency bins. The SRO estimation here involves fitting a line to this observation. Note that the slope of the line has a linear relation with the value of the SRO. As it can be seen in this figure, there are some outliers in different frequency bins causing a harmful effect on the accuracy of data fitting[4].

[4]Note that in practice, we can never determine the exact cause of existing outliers, however in this specific case, the number of outliers is more in higher frequencies, which can indicate that the outliers occur due to phase wrapping. By the way the proposed outlier removal procedure remove all outliers without considering their cause.

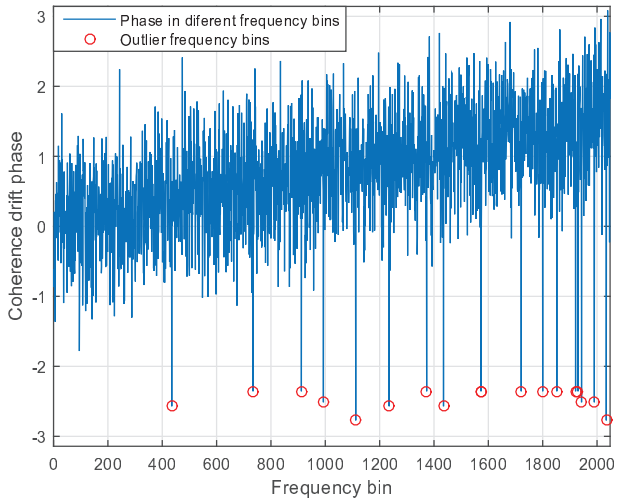

Fig. 1: Outliers in the coherence drift phase.

Furthermore, (31) only focuses on the coherence amplitude and completely ignores the coherence phase, while in some scenarios many outlier frequency bins can occur due abrupt changes in the coherence phase. For example, Figure 2 shows the coherence drift[5] in a turn-taking scenario over a 6 frames segment, where in frames 3 and 4 the first source stops speaking the second source starts speaking. As can be seen in this example this turn-taking scenario abruptly changes the phase drift in frames 3 and 4 and causes many outlier frequency bins in this two frames.

Therefore, we adopt a two-step outlier removal (OR) procedure focusing on the coherence phase and yielding a binary (0 and 1) weigh to each frequency bin. In the first step, we make a rough estimation of ϵ_i through the following least absolute value (LA) minimization:

$$\hat{\epsilon}_i^{\mathrm{LA}} = \arg\min_{\epsilon_i} \|\vec{\mathbf{A}}_{\mathrm{v}} - \vec{\mathbf{B}}_{\mathrm{v}}\epsilon_i\|_1 \tag{32}$$

where $\|\cdot\|_1$ denotes the L_1-norm. The LA estimation is known to be more robust against outliers compared to the ordinary LS estimation. Furthermore, solving (32) also allows to detect the outliers, e.g., using thresholding of the absolute error. LA minimization of (32) does not have an analytical solution and usually an iterative approach is applied such as, e.g., a simplex-based approach [28], iteratively re-weighted least-squares [29], Wesolowsky's direct descent approach [30] and Li-Arce's maximum

[5]In Figure 2, the coherence drift of six frames is concatenated one after the other one to show the effect of turn-taking sources in frames 3 and 4.

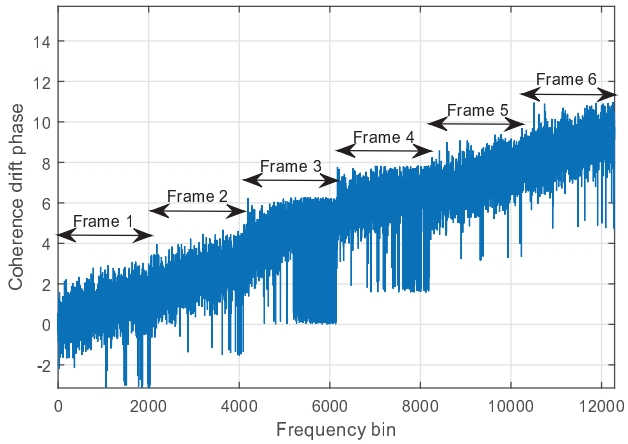

Fig. 2: Outliers in the coherence drift phase of six frames, where two sources take a turn in frames 3 and 4.

likelihood approach [31].

In the second step, the outliers are detected and removed, after which a more accurate SRO estimate can be computed, this time using a LS minimization for computational convenience. The frequency bins (rows of $\mathbf{A}$ and $\mathbf{B}$) satisfying the following condition are considered as outlier frequency bins:

$$\exists q \in \{1, \ldots, Q\}: \quad |a_{k,q} - b_{k,q}\hat{\epsilon}_i^{\mathrm{LA}}| > \alpha\sigma_q, \tag{33}$$

where σ_q is the standard deviation of the elements in the q^{th} column of the residual matrix $\mathbf{R} = \mathbf{A} - \mathbf{B}\hat{\epsilon}_i^{\mathrm{LA}}$ and α is a tuning parameter, which is usually around 1.

After detection and removal of the outlier frequency bins, we proceed with the WLS minimization.

$$\hat{\epsilon}_i^{\mathrm{WLS}} = \arg\min_{\epsilon_i} \|\vec{\mathcal{A}}_{\mathrm{v}} - \vec{\mathcal{B}}_{\mathrm{v}}\epsilon_i\|_2, \tag{34}$$

where matrices $\vec{\mathcal{A}}_{\mathrm{v}}$ and $\vec{\mathcal{B}}_{\mathrm{v}}$ are equivalents of $\vec{\mathbf{A}}_{\mathrm{v}}$ and $\vec{\mathbf{B}}_{\mathrm{v}}$ after removal of the outlier rows in $\mathbf{A}$ and $\mathbf{B}$. Finally, the optimal solution of (34) can be obtained as

$$\hat{\epsilon}_i^{\mathrm{WLS}} = \frac{(\vec{\mathcal{B}}_{\mathrm{v}})^T \vec{\mathcal{A}}_{\mathrm{v}}}{(\vec{\mathcal{B}}_{\mathrm{v}})^T \vec{\mathcal{B}}_{\mathrm{v}}}. \tag{35}$$

IV. SRO Compensation

speech enhancement is required in many applications such as speech recognition, hearing aids and speaker characterisation and verification [32]–[34]. In this paper, we focuse on multi-channel speech enhancement using MWF and try to perform SRO compensation over MWF. For the SRO compensation,

two complementary operations are performed: skipping critical samples in the time-domain and phase compensation in the frequency-domain. We will explain why both have to be applied in a hybrid compensation framework. In this approach, an estimate of the SRO is assumed to be available from the LCD described in Section III.

A. Time-domain compensation

Assume w.l.o.g. that the i^{th} microphone signal has a positive relative SRO ϵ_i with respect to the reference signal. The SRO then causes a linearly increasing delay between the two signals. Therefore, after a certain time τ, the signals are drifted more than 1 sample apart from each other. The corresponding sample n_τ is found as the first sample for which the following inequality is satisfied:

$$f_s^{\text{ref}} \left[\frac{n}{f_s^{\text{ref}}} - \frac{n}{(1+\epsilon_i) f_s^{\text{ref}}} \right] > 1, \tag{36}$$

i.e. $n_\tau = \epsilon_i^{-1}$ (using the same approximation as in (14)). By skipping one sample after n_τ samples, the signals will be re-aligned again. This procedure can be repeated after each n_τ samples indefinitely and will ensure that the two of signals will never drift further apart than 1 sample.

B. Frequency-domain compensation

The SRO compensation in the frequency-domain is performed based on the fact that a fixed delay of ϱ_i samples in $x_i[n]$ causes a phase rotation of $\frac{2\pi k \varrho_i}{K}$ in frequency bin k. In other words, two signals shifted relative to each other in the time-domain can be re-aligned by a simple phase-shift in the frequency-domain. However, an SRO causes a linearly increasing delay instead of a fixed delay. Still we compensate for a linearly increasing delay with a fixed phase-shift assuming the drift caused by the SRO within a single STFT segment is much smaller than 1 sample, i.e. $|K\epsilon_i| \ll 1$. Therefore, the compensation is more accurate for a small segment-size and a small SRO. For each segment we calculate the SRO induced delay at the mid-segment sample based on the estimated SRO and obtain the corresponding phase rotation $\frac{2\pi k m \hat{\epsilon}_i^{\text{WLS}}}{K}$. For the STFT segment, the k^{th} frequency bin is then multiplied by $\exp\left(-j\frac{2\pi k m \hat{\epsilon}_i^{\text{WLS}}}{K}\right)$ to compensate for the phase rotation caused by the SRO.

Since the MWF is typically applied in the STFT domain, this frequency-domain compensation is computationally very cheap.

C. Hybrid compensation

If the frequency-domain compensation would be applied alone, the signals at two different nodes drift more and more away from each other as the time increases, until their STFT segments no longer relate

to the same source signal STFT segments. A phase rotation in the STFT domain can obviously no longer compensate for this. Therefore, the frequency-domain compensation cannot be applied without the time-domain compensation.

Applying the time-domain compensation without the frequency-domain compensation is also not sufficient. Even though the signals will then never drift further apart than one sample, there will be a significant performance drop due to short-term time-varying coherence phases in the second-order signal statistics used in, e.g., the MWF.

Therefore, both compensation schemes are essential and have to be combined into a hybrid scheme to compensate for the SRO effects in, e.g., a speech enhancement algorithm. The hybrid compensation is in fact split up in realigning the segments (coarse-scale compensation) and the compensation of small phase-shifts (fine-scale compensation).

The hybrid compensation is straightforwardly integrated into the MWF. To implement the hybrid compensation in MWF, we basically apply time-domain compensation and then a frequency-domain compensation is applied each time a sample is skipped (compensating for a 1-sample delay corresponding to a phase-shift of $\frac{2\pi k}{K}$).

V. Validation

In this section, we briefly describe two benchmark methods with which we will compare our LCD. Then we present our evaluation setup and investigate the accuracy of the proposed methods for SRO estimation and compensation.

A. Benchmark methods

We will compare the LCD with two benchmark methods, which we refer to as averaged coherence drift (ACD) SRO estimation [9] and maximum likelihood (ML) SRO estimation [12].

1) Averaged coherence drift (ACD): In [9], a method for SRO estimation has been proposed, based on the phase-drift of the coherence of noise-only segments of the signals. The SRO is estimated as follows:

$$\hat{\epsilon}_i^{\mathrm{ACD}} = \frac{1}{K_{max}} \sum_{k=1}^{K_{max}} \frac{K}{2\pi\Lambda k Q} \sum_{q=1}^{Q} a_{k,q} \tag{37}$$

$$= \frac{1}{K_{max}} \sum_{k=1}^{K_{max}} \hat{\epsilon}_{i,k}^{\mathrm{ACD}}, \tag{38}$$

where $a_{k,q}$ is defined in (17), $\hat{\epsilon}_{i,k}^{\mathrm{ACD}}$ is the ACD estimated SRO in k-th frequency bin and $K_{max} < K$ is the maximum number of considered frequency bins and is determined such that $a_{k,q}$ is bounded in the range $[-\pi, \pi]$ to avoid phase ambiguity.

Eq. (38) implies that the ACD computes an SRO for each frequency bin and then averages over the estimated SROs to obtain an overall SRO between the two signals. This method also assumes the availability of a coherent noise source and a VAD to detect noise-only segments.

The LCD enjoys four distinct advantages compared to the ACD. First, in the ACD, the final SRO is computed by averaging over the estimated SROs in each frequency bin. Instead, we use a least squares estimation framework, which minimizes the sum of the squared residuals (errors). In Appendix C, we prove that, for the case of Gaussian noise and for the same data points in both methods, the mean and variance of the SRO estimation error in ACD is always larger compared to LCD (even if no OR or WG schemes are used).

Second, the ACD looses available information in speech frames by applying a VAD. This problem significantly deteriorates the accuracy of the ACD when there are few noise-only frames. The LCD solves this problem by developing an OR and WG procedures in frequency domain, which results in exploiting available information in both speech and noise frames. In this method, improper frequency bins are ignored or down-weighted and the rest are incorporated in the estimation.

Third, to avoid the phase ambiguity, the ACD completely neglects frequency bins larger than K_{max} to avoid the phase wrapping point, whereas many of these bins contain useful information about the SRO and only a few of them are affected by phase wrapping. The applied OR procedure of LCD implicitly removes the frequency bins with phase ambiguity, and hence exploits a lot more informative frequency bins compared to the ACD. The effect of the OR procedure on both LCD and ACD is studied in our experiments.

Finally, the ACD suggests no method to deal with multiple source scenarios, while the WG technique improves the results of the LCD in such a case.

2) Maximum likelihood (ML): A blind SRO estimation method based on a maximum likelihood estimation of the sampling frequency mismatch in the STFT domain has been proposed by [10], [12]. The SRO is again translated into a phase-rotation in the STFT domain (**??**) and then estimated by solving the following likelihood maximization problem

$$\hat{\epsilon}_i^{\mathrm{ML}} = \arg\max_{\epsilon_i} \Omega\left(\epsilon_i\right) \tag{39}$$

$$\Omega\left(\epsilon_i\right) = -\sum_{k=1}^{K} \log\left(1 - \left|\Phi_{1,i}[k;-\epsilon_i]\right|^2\right) \tag{40}$$

$$\Phi_{1,i}[k;-\epsilon_i] = \frac{\sum_\iota X_1^\iota[k] X_i^\iota[k]^* \exp\left(\frac{2\pi k(\iota\Lambda+1)\epsilon_i}{K} j\right)}{\sqrt{\sum_\iota |X_1^\iota[k]|^2}\sqrt{\sum_\iota |X_i^\iota[k]|^2}}. \tag{41}$$

This optimization problem (39) does not have an analytical solution and a numerical approach, namely a golden section search, is applied.

The ML assumes a fixed acoustic transfer functions between the sources and the microphones during the full batch size –over which the SRO is estimated– to yield an accurate SRO estimation. Therefore, its accuracy severely degrades in the case of turn-taking sources, e.g., during a conversation. In a turn-taking speakers scenario the coherence phase changes drastically at certain time instances as shown in Figure 2. The LCD is less affected since in such a scenario, the time instances in which a drastic change occurs will be considered as outliers and removed effectively by the proposed OR procedure. However, the ML suggests no method to deal with such abrupt changes in the coherence as it considers the overall change in coherence over the full batch size, and hence yields less accurate results in such a case.

B. Experimental setup

For the SRO estimation, the LCD uses frames of length $\Gamma = 4096$ with 50% overlap. Coherence is calculated using the Welch method [23] with segment size $K = 2048$, using a Hamming window, and with 75% overlap, i.e., m is incremented with 4096/4=1024 samples between consecutive segments (note that a frame of length $\Gamma = 4096$ is then chunked into 5 smaller segments of length $K = 2048$ with %75 overlap). We assume a nominal sampling rate of 8kHz in all experiments.

The accuracy of the SRO estimation is measured using the mean absolute error E_{MA} and median absolute error E_{MdA} calculated as

$$E_{\mathrm{MA}} = \frac{1}{L}\sum_{l=1}^{L} \left|\epsilon_l - \hat{\epsilon}_l\right|, \tag{42}$$

$$E_{\mathrm{MdA}} = \mathbf{Median}\left(\left|\epsilon_1 - \hat{\epsilon_1}\right|, \cdots, \left|\epsilon_l - \hat{\epsilon_l}\right|, \cdots, \left|\epsilon_L - \hat{\epsilon_L}\right|\right), \tag{43}$$

where ϵ_l and $\hat{\epsilon}_l$ are the true and estimated SRO respectively and L is the total number of experiments.

C. SRO Estimation on Real Audio Recordings

We validate the proposed method in two different real-world scenarios. In the first experiment, we recorded a static speech source scenario when multiple noise sources were also present. In the second experiment, we recorded a real conversation between two persons when multiple noise sources were also present.

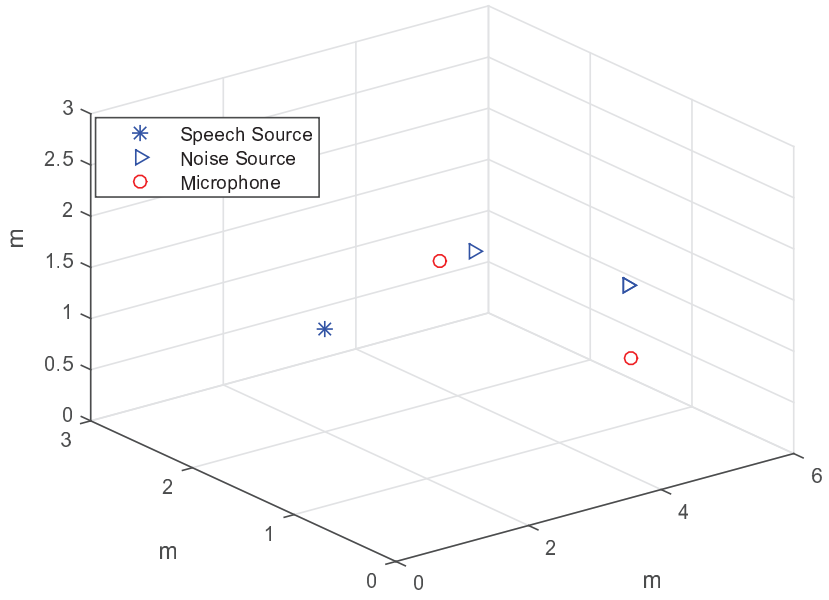

Fig. 3: The location of sources and microphones with actual SRO in the room.

1) Static speech sources scenario: In this experiment, we perform SRO estimation over real speech data recorded in an office environment. We repeated this experiment for 10 times using different speech signals from the Hearing in Noise Test (HINT) database [35]. The room contains 3 sound sources, which produce either speech or background noise and 2 recording devices, i.e. the microphones of laptops from two different brands, namely an Apple MacBook Pro and a Sony VAIO. The sampling frequency of each module was set to 8 kHz and the recording was performed with a single channel and 16 Bits-Per-Sample. The location of the sources and microphones is depicted in Figure 3.

There is no ground-truth about the actual SRO between the two signals to measure the accuracy of the applied SRO estimation methods directly. However, by compensating for different values of SRO in an MWF-based noise reduction we can determine which SRO yields a better enhancement and use it as a ground-truth. Applying the SDW-MWF with the specifications mentioned in Section V-E to the received signals after compensation for different values of SRO shows that the maximum SNR is obtained at 20.60 ppm, hence we use this value as a ground-truth.

Table I lists the mean absolute error (E_{MA}) and the median absolute error (E_{MdA}) of the SRO estimates, where a signal of 6 seconds is available for the SRO estimation in all methods. We repeated this experiment for 10 times using different speech signals. To study the effect of the WG scheme described

TABLE I: *The E_{MA} and E_{MdA} of the SRO estimation by LCD, ACD and ML in the static source scenario using 6 seconds of signal, averaged over 10 different experiments (all units are in ppm).*

System Configuration		E_{MA}	E_{MdA}
ACD		9.55	8.94
ML		1.98	1.97
LCD	without WG, without OR	5.09	4.93
	without WG, with OR	1.77	1.80
	with WG, without OR	2.56	2.61
	with WG, with OR	**0.59**	**0.41**

in Section (III-B2) and the OR procedure explained in Section (III-B3), we reported the results of the LCD with and without OR and WG. It is observed that the applied WG technique is effective and the proposed OR substantially improves the performance of the LCD. Furthermore, the combination of WG and OR is more powerful than any of them separately and remarkably improve the estimation results, which suggests that they have a complementary effect on the performance.

Table I also demonstrates that the accuracy of the LCD with OR and WG is considerably more than both ML and ACD in SRO estimation. Applying the SDW-MWF to the received signals after compensation for estimated SRO using using LCD, ML and ACD yield SNR improvement of 6.89%, 3.26% and 1.33% respectively.

It is also shown that the ACD, which uses only the noise-only segments detected through a perfect VAD yields less accurate results compared to ML and LCD in this scenario, which can be due to the fact that ACD does not apply useful information in speech segments and requires on consecutive noise-only segments.

To further investigate the performance of ACD compared to ML and LCD, the accuracy of ACD, ML and LCD for different signal segment lengths is depicted in Figure 4. This figure illustrates that ACD requires a long signal segment to yield reliable results, while ML and LCD can estimate the SRO by processing much shorter signal segments (note that the horizontal and vertical axes are scaled differently in both figures). This figure also shows that the accuracy of ACD, ML and LCD improves by increasing the signal segment length. Of course, this comes at the cost of decreased tracking capabilities and increased algorithmic delays. This figure demonstrates that for a short batch-size of only 1 second LCD is considerably more accurate than ML, which suggests that the tracking capabilities of LCD is superior compared to ML.

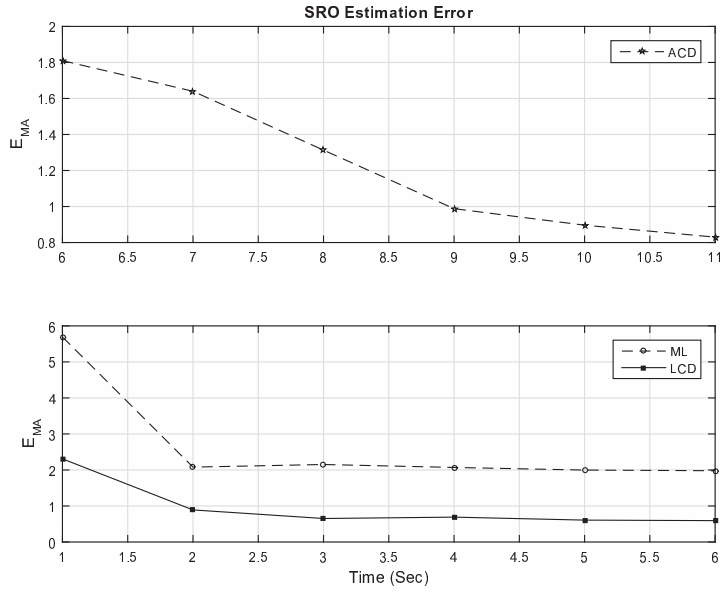

Fig. 4: The E_{MA} of the SRO estimation versus signal segment length (batch size) in a static source scenario, averaged over 10 experiments. Note that the horizontal and vertical axes are scaled differently in both figures.

2) Turn-taking speech sources scenario: In this experiment, we perform SRO estimation over a real conversation of two persons recorded in an office environment.

The room contains 4 sound sources, which produce either speech or background noise. The same recording devices mentioned in Section V-C1 are used.

We have recorded the 5 different conversations while the speaker were located at different positions of the room and the microphones locations were the same as those of the last experiment depicted in Figure 3.

Table II lists the mean absolute error (E_{MA}) and the median absolute error (E_{MdA}) of the SRO estimates, where a signal of 6 seconds is available for the SRO estimation in all methods. Table **??** demonstrates that the results of LCD, ML and ACD in SRO estimation. As expected, the accuracy of ACD and ML is considerably lower than that of the LCD with OR and WG due to the violation of the fixed transfer function assumption. In this case, since each speaker generates a different transfer function at the microphones, the coherence abruptly changes when one speaker becomes active and the other becomes silent. This abrupt change in coherence will only affect one or two measurement frames in the LCD method, which will be detected and removed via the applied OR procedure, whereas ML and ACD

TABLE II: *The E_{MA} and E_{MdA} of the SRO estimation by LCD, ACD and ML in the turn-taking source scenario using 6 seconds of signal, averaged over 5 different experiments (all units are in ppm).*

System Configuration		E_{MA}	E_{MdA}
ACD		13.05	13.87
ML		3.22	2.47
LCD	without WG, without OR	16.69	16.87
	without WG, with OR	9.22	6.19
	with WG, without OR	12.27	12.10
	with WG, with OR	**0.80**	**0.79**

estimate the SRO from the coherence over the full batch size.

D. SRO Estimation on Simulated Data

It is noted that our experiments on real recorded data is very limited and we could not study the statistical significance of the obtained results. Therefore, to perform a Monte-Carlo experiment over different controlled experimental settings, we have simulated a $5m \times 5m \times 3m$ reverberant room with a T60 reverberation time of 0.3 seconds using the image method [36], [37]. We used 25 speech signals from the Hearing in Noise Test (HINT) database [35]. Signal re-sampling is performed using Sound eXchange (SOX) software[6]. An uncorrelated (diffuse) additive white Gaussian noise is present in each microphone with power equal to 20% of the speech signal power. We considered three cases a static source case, a turn-taking source case and a multiple source case.

1) Static speech source scenario: In this scenario, the microphones are located at positions [4.5 1 0.5] and [0.5 1 0.5]. The sampling rate of the reference microphone is set to 8kHz and the sampling rate of the second microphone is subject to an offset of 1, 10, 40 and 80 parts per million (ppm) of the sampling rate of the first microphone. In this experiment, a speech source and a localized white noise source are used. The ratio of the power of the speech signal and the power of the localized noise signal is around 9 dB. For every SRO value, we conducted 100 Monte-Carlo experiments (4 experiments for each speech signal), where the location of the speech and noise sources are randomly selected.

Table III lists the mean absolute error (E_{MA}) of the SRO estimates in static, turn-taking and multiple scenarios, where a signal of 6 seconds is available for the SRO estimation. This table demonstrates that

[6]http://sox.sourceforge.net/

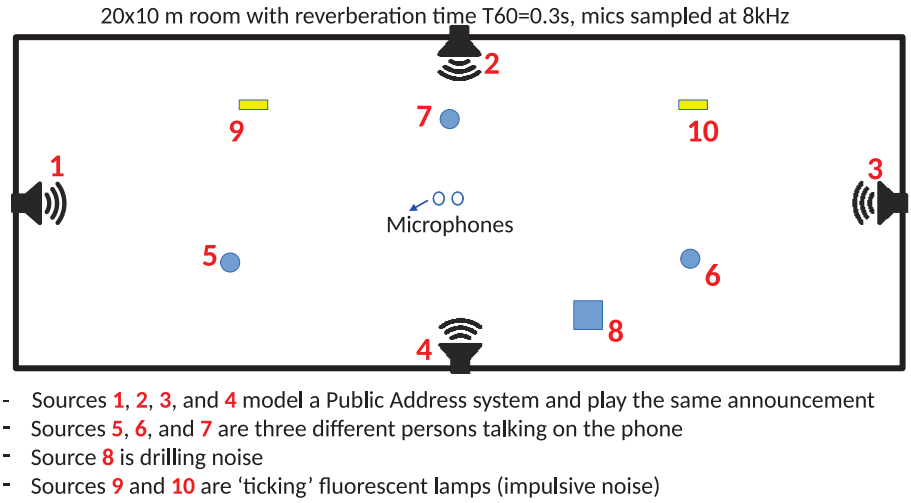

Fig. 5: Description of case 2 in multiple speech sources scenario.

LCD outperforms ACD and ML in static source scenario. This results concur with those of real-world recorded data.

2) Turn-taking speech source scenario: In the first case, we simulate a conversation between three speakers. Consisting of three time intervals of 2 seconds each. In each time interval, only one of the speakers is active. The sampling rate of the reference microphone is set to 8kHz and the sampling rate of the second microphone is subject to an offset of 10, 20, 40 and 80 ppm. Table III shows that performance of LCD, ACD and ML in this scenario. Similar to the results of real-world recorded data, the LCD yields more accurate results compared to ACD and ML in the turn-taking scenario.

3) Multiple speech source scenario: In this case, a 20mx10m reverberant room is simulated with a T60 reverberation time of 0.3 seconds. The room contains 10 sound sources, which produce either continuous speech or background noise and 2 recording devices with nominal sampling rate of 8kHz. The location of the sources and microphones is depicted in Figure 5.

Table III lists the E_{MA} of the SRO estimates in these cases, where the available data of SRO estimation for both methods is 6 seconds. This table shows that the LCD still estimates the SRO without a severe performance degradation, which shows the effectiveness of the proposed OR and WG which up-scale the contribution of good frequency bins.

E. SRO compensation for noise reduction

For noise reduction, the speech-distortion weighted MWF (SDW-MWF) [38] with square-root Hann window of size 1024, 50% window overlap and a forgetting factor of 0.997 is applied to the static speech

TABLE III: *The E_{MA} of the SRO estimation by the LCD and ML in the static, turn-taking and multiple speech sources scenarios (all units are in ppm).*

PPM	Static			Turn-taking			Multiple		
	ACD	LCD	ML	ACD	ML	LCD	ACD	ML	LCD
80	10.25	0.18	0.13	20.48	18.62	14.07	5.23	0.42	0.24
40	8.67	0.16	0.13	24.54	17.69	9.61	6.44	0.48	0.27
20	8.34	0.14	0.12	26.44	14.82	8.12	7.54	0.50	0.50
10	8.15	0.14	0.12	29.82	11.39	6.67	9.01	0.50	0.65

TABLE IV: *The SNR of signals after using the proposed hybrid compensation approach (dB).*

PPM	Uncompensated	Compensated	
		with true SRO	with estimated SRO
100	17.43	20.52	20.55
10	19.94	20.55	20.54
1	20.50	20.50	20.50
0	20.55	20.55	20.55

source scenario described in Section V-D1. Table IV shows the output SNR of the SDW-MWF without SRO compensation (in the column 'Uncompensated') and after SRO compensation using the LCD (in the column 'Compensated'). To investigate the effect of errors in the estimation of the SRO on the proposed compensation, we list both the results with perfectly known SRO and with the estimated SRO.

Comparison of the last row of this table -where there is no SRO between the nodes- with the rest of the rows shows that the SRO substantially degrades the performance of the MWF. This degradation increases with the SRO[7]. Therefore, a compensation method to avoid this performance-loss is necessary. The result after compensation with the true SRO is very near to the perfect case (no SRO) indicating the efficiency of the applied compensation scheme. The compensation, in which the SRO is estimated, produces similar results. Therefore it is confirmed that the applied method can effectively recover the degradation of the MWF performance caused by SRO.

[7]It is noted that the experiments are performed with an adaptive MWF [39], which implicitly already performs some SRO compensation due to the continuous updating of the second-order statistics of the signals within the MWF. However, this implicit SRO compensation is not sufficient, since the SRO causes variations in these statistics which are usually too fast to be tracked with an adaptive MWF.

VI. Conclusions

A new approach to blind SRO estimation in an asynchronous wireless acoustic sensor network has been proposed in this paper. This method assumes that the SRO causes a linearly increasing time-delay between two signals, hence induces a linearly increasing phase-shift in the STFT domain. For the SRO estimation, the phase of the coherence function between two microphone signals is monitored, where the SRO induces a phase-drift over time. After outlier frequency bin removal, the obtained coherence phase-drift, which has a linear relation with the SRO of the signals, is used in a weighted LS framework to estimate the SRO. Experimental results in different scenarios with static and turn-taking sources show the effectiveness of the LCD. Finally it has been demonstrated that the proposed SRO estimation along with a hybrid SRO compensation can eliminate the performance loss due to SRO in an MWF-based signal enhancement.

Appendix A

The WG scheme (31) implies that the global maximum of $v_{k,q}$ is 1, which is attained if and only if

$$\left|\Phi_{1,i}^{m}[k;\rho_i^m]\right| = 1 \tag{44}$$

$$\left|\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}]\right| = 1. \tag{45}$$

We prove that if (21) holds, then $v_{k,q} = 1$. W.l.o.g. let's assume $q = 1$. Assuming the first condition (21) is satisfied, the coherence (8) can be simplified to (11) as shown in equations (8)-(11), and hence it follows that

$$\left|\Phi_{1,i}^{m}[k;\rho_i^m]\right| = \left|\frac{H_{1,z}[k](H_{i,z}[k])^*}{|H_{1,z}[k]|\,|H_{i,z}[k]|}\right| = 1. \tag{46}$$

In equations (8)-(15), which hold under assumption (21), we show that $\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}] = \Phi_{1,i}^{m}[k;\rho_i^m]\exp\left(\frac{2\pi k\Lambda}{K}\epsilon_i\right)$, hence

$$\left|\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}]\right| = 1. \tag{47}$$

Therefore, $v_{k,q}$ is maximized if the condition (21) is satisfied.

This proof shows that (21) is a sufficient condition to maximize the weight $v_{k,q}$, although it may not be necessary condition. Nevertheless, significant discrepancies from condition (21) will typically decrease the weight unless in contrive cases.

APPENDIX B

In this appendix, we show that (22) yields a sufficient condition to minimize the denominator of (31). W.l.o.g. let's assume $q = 1$. Since $|\cdot|^2$ is a convex function, the global minimum of the denominator of (31) is 1, which is attained if

$$\left|\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}]\right| = \left|\Phi_{1,i}^{m}[k;\rho_i^{m}]\right|. \tag{48}$$

In equations (24)-(27), we calculated $\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}]$ and $\Phi_{1,i}^{m}[k;\rho_i^{m}]$ assuming the second condition (22) is satisfied. Comparing (25) and (27) shows that the magnitudes of $\Phi_{1,i}^{m-\Lambda}[k;\rho_i^{m-\Lambda}]$ and $\Phi_{1,i}^{m}[k;\rho_i^{m}]$ are the same. Therefore, the denominator of (31) is minimized if the condition (22) is satisfied.

APPENDIX C

Assume both LCD and ACD methods use the same measurements $a_{k,q}$ in all frequency bins and consider

$$a_{k,q} = a_{k,q}^{\diamond} + e_{k,q}, \tag{49}$$

where $a_{k,q}^{\diamond}$ is the actual phase drift between the two signals due to SRO and $e_{k,q}$ is its corresponding error. To obtain a fair comparison between ACD and LCD, we do not include WG and OR in LCD. To make the analysis mathematically tractable we assume that both methods use the same number of bins and we ignore phase wrapping, i.e., we assume that $K = K_{max}$ and the error $e_{k,q}$ are small enough such that no phase wrapping occurs. The SRO estimation[8] of the ACD can be obtained by inserting (49) into (37), i.e.,

$$\hat{\epsilon}_i^{\text{ACD}} = \frac{1}{K}\sum_{k=1}^{K}\frac{K}{2\pi\Lambda k}\left(a_{k,1}^{\diamond} + e_{k,1}\right), \tag{50}$$

or

$$\hat{\epsilon}_i^{\text{ACD}} = \epsilon_i + \frac{1}{K}\sum_{k=1}^{K}\frac{K}{2\pi\Lambda k}e_{k,1}. \tag{51}$$

Assuming that the measurement errors are independent and identically distributed (iid) with a zero-mean Gaussian distribution with variance σ_e^2, the variance of the ACD SRO estimation error, denoted as $(\sigma_i^{\text{ACD}})^2$, is equal to

$$(\sigma_i^{\text{ACD}})^2 = E\{(\hat{\epsilon}_i^{\text{ACD}} - \epsilon_i)^2\} \tag{52}$$

$$= \frac{\sigma_e^2}{K^2}\sum_{k=1}^{K}\left(\frac{K}{2\pi\Lambda k}\right)^2. \tag{53}$$

[8]For easy exposition and w.l.o.g we assume that $Q = 1$ in this proof.

The SRO estimation of LCD can be obtained by inserting[9] (49) into (20), i.e.,

$$\hat{\epsilon}_i^{\mathrm{LS}} = \frac{\vec{\mathbf{B}}^T \left(\vec{\mathbf{A}}^{\diamond} + \vec{\mathbf{e}} \right)}{\vec{\mathbf{B}}^T \vec{\mathbf{B}}}, \tag{54}$$

where $\mathbf{A}^{\diamond}$ and $\mathbf{e}$ are matrices of appropriate size with elements $a_{k,q}^{\diamond}$ and $e_{k,q}$ respectively. Expanding (54) results in

$$\hat{\epsilon}_i^{\mathrm{LS}} = \epsilon_i + \frac{\sum_{k=1}^{K} \frac{2\pi\Lambda k}{K} e_{k,1}}{\sum_{k=1}^{K} \left(\frac{2\pi\Lambda k}{K}\right)^2}. \tag{55}$$

The variance of LCD SRO estimation error, denoted as $(\sigma_i^{\mathrm{LCD}})^2$, equal to

$$(\sigma_i^{\mathrm{LCD}})^2 = E\{(\hat{\epsilon}_i^{\mathrm{LCD}} - \epsilon_i)^2\} \tag{56}$$

$$= \frac{\sigma_e^2}{\sum_{k=1}^{K} \left(\frac{2\pi\Lambda k}{K}\right)^2}. \tag{57}$$

Consider the SRO estimation variance of LCD and ACD obtained in (57) and (53) respectively. The Cauchy-Schwarz inequality implies that

$$|\bar{t}_1\bar{u}_1 + \cdots + \bar{t}_K\bar{u}_K|^2 \leq$$
$$(|\bar{t}_1|^2 + \cdots + |\bar{t}_K|^2)(|\bar{u}_1|^2 + \cdots + |\bar{u}_K|^2). \tag{58}$$

Replacing $\bar{t}_k$ by$\frac{1}{\bar{u}_k}$ and $\bar{u}_k$ by $\frac{2\pi\Lambda k}{K\sigma_e}$, results in $\sigma_i^{\mathrm{LCD}} \leq \sigma_i^{\mathrm{ACD}}$, which implies that LCD yields a more accurate estimation compared to ACD.

In the case of dynamic source scenario, the measurement error mean μ_e can be non-zero. In this case, the mean of the ACD and LCD estimations are obtained as

$$\mu_i^{\mathrm{ACD}} = E\{\hat{\epsilon}_i^{\mathrm{ACD}}\} = \epsilon_i + \frac{\mu_e}{K} \sum_{k=1}^{K} \frac{K}{2\pi\Lambda k}, \tag{59}$$

and

$$\mu_i^{\mathrm{LCD}} = E\{\hat{\epsilon}_i^{\mathrm{LCD}}\} = \epsilon_i + \frac{\sum_{k=1}^{K} \frac{2\pi\Lambda k}{K}}{\sum_{k=1}^{K} \left(\frac{2\pi\Lambda k}{K}\right)^2} \mu_e, \tag{60}$$

respectively.

By expanding the series in numerator and denominator of (60), the mean of LCD estimation error can be obtained as

$$\mu_i^{\mathrm{LCD}} = \epsilon_i + \frac{3K}{2\pi\Lambda(2K+1)} \mu_e \tag{61}$$

[9]To have a fair comparison between the ACD and LCD, we ignore using OR and WG in this proof.

By comparing (61) and (59), it is trivial to validate that $|\mu^{\mathrm{LCD}} - \epsilon_i| \leq |\mu^{\mathrm{ACD}} - \epsilon_i|$, where the equality holds for $K = 1$ only. This implies that the expected value of the SRO estimation error using LCD is almost always lower compared to ACD.

REFERENCES

[1] M. H. 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)*, Nice, France, Sep. 2015, pp. 2326–2330.

[2] A. Bertrand and M. Moonen, "Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology," *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2196–2210, 2011.

[3] L. Lanbo, Z. Shengli, and C. Jun-Hong, "Prospects and problems of wireless communication for underwater sensor networks," *Wireless Communications and Mobile Computing*, vol. 8, no. 8, pp. 977–994, 2008.

[4] M. F. F. B. Ismail and L. W. Yie, "Acoustic monitoring system using wireless sensor networks," *Procedia Engineering*, vol. 41, pp. 68–74, 2012.

[5] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

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

[7] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP Journal on Applied Signal Processing*, p. 12, 2009.

[8] M. H. Bahari, J. Plata-Chaves, A. Bertrand, and M. Moonen, "Distributed labelling of audio sources in wireless acoustic sensor networks using consensus and matching," in *Proc. European signal processing conf. (EUSIPCO)*, Hungary, 2016, pp. 2345–2349.

[9] 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 *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*. VDE, 2012, pp. 1–4.

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

[11] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, April 2003, pp. IV–840–3 vol.4.

[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 Processing*, vol. 107, pp. 185 – 196, 2015.

[13] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," *International Workshop on Acoustic Signal Enhancement 2014*, 2014.

[14] 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," in *Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on*. IEEE, 2004, pp. 18–25.

[15] J. Elson and K. Römer, "Wireless sensor networks: A new regime for time synchronization," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 1, pp. 149–154, 2003.

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

[17] F. Hoflinger, R. Zhang, J. Hoppe, A. Bannoura, L. Reindl, J. Wendeberg, M. Buhrer, and C. Schindelhauer, "Acoustic self-calibrating system for indoor smartphone tracking (assist)," in *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*, Nov 2012, pp. 1–9.

[18] T. Janson, C. Schindelhauer, and J. Wendeberg, "Self-localization application for iphone using only ambient sound signals," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, Sept 2010, pp. 1–10.

[19] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware–software approach for acoustic sensor network synchronization," *Signal Processing*, vol. 107, pp. 171–184, 2015.

[20] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," in *Proc. IWAENC*, 2008.

[21] L. Erup, F. M. Gardner, and R. A. Harris, "Interpolation in digital modems. ii. implementation and performance," *Communications, IEEE Transactions on*, vol. 41, no. 6, pp. 998–1008, 1993.

[22] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230 – 2244, Sep. 2002.

[23] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[24] M. Bahari, N. Dehak, H. Van hamme, L. Burget, A. Ali, and J. Glass, "Non-negative factor analysis of gaussian mixture model weight adaptation for language and dialect recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1117–1129, 2014.

[25] Z. Yermeche, N. Grbic, and I. Claesson, "Blind subband beamforming with time-delay constraints for moving source speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2360–2372, 2007.

[26] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1503–1512, 2012.

[27] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[28] I. Barrodale and F. D. Roberts, "An improved algorithm for discrete l_1 linear approximation," *SIAM Journal on Numerical Analysis*, vol. 10, no. 5, pp. 839–848, 1973.

[29] E. Schlossmacher, "An iterative technique for absolute deviations curve fitting," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 857–859, 1973.

[30] G. Wesolowsky, "A new descent algorithm for the least absolute value regression problem: A new descent algorithm for the least absolute value," *Communications in Statistics-Simulation and Computation*, vol. 10, no. 5, pp. 479–491, 1981.

[31] Y. Li and G. R. Arce, "A maximum likelihood approach to least absolute deviation regression," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1762–1769, 2004.

[32] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE, 2014, pp. 7–12.

[33] N. Dehak, O. Plchot, and M. H. Bahari, "Gmm weights adaptation based on subspace approaches for speaker verification," in *Proceedings Odyssey 2014*, 2014, pp. 48–53.

[34] M. Bahari, M. McLaren, h. Van hamme, and D. van Leeuwen, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[35] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[36] E. A. P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep 2.2.4*, 2006.

[37] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

[38] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[39] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.