BLIND SAMPLING RATE OFFSET ESTIMATION BASED ON COHERENCE DRIFT IN WIRELESS ACOUSTIC SENSOR NETWORKS

Mohamad Hasan Bahari, Alexander Bertrand, Marc Moonen

KU Leuven, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics

ABSTRACT

In this paper, a new approach for sampling rate offset (SRO) estimation between nodes of a wireless acoustic sensor network (WASN) is proposed using the phase drift of the coherence function between the signals. This method, referred to as least squares coherence drift (LCD) estimation, assumes that the SRO induces a linearly increasing phase-shift in the short-time Fourier transform (STFT) domain. This phase-shift, observed as a drift in the phase of the signal coherence, is applied in a least-squares estimation framework to estimate the SRO. Simulation results in different scenarios show that the LCD estimation approach can estimate the SRO with a mean absolute error of around 1%. We finally demonstrate that the use of the LCD estimation within a compensation approach eliminates the performance-loss due to SRO in a multichannel Wiener filter (MWF)-based speech enhancement task.

Index Terms— Wireless Acoustic Sensor Networks, Signal Enhancement, Sampling Rate Offset, Coherence Drift

1. INTRODUCTION

In a wireless acoustic sensor network (WASN), sampling rate offsets (SRO), caused by clock imperfections in different nodes, introduce many challenges for the fusion of signals captured by different microphones [1, 2]. For example, an SRO severely degrades the performance of different speech processing algorithms. Elson and Kay [3] addressed the time synchronization problem in wireless sensor networks by using a reference-broadcast synchronization algorithm to synchronize the clocks. The SRO estimation problem for acoustic beamforming in particular was tackled by Wehr et al. [4], using a modulated RF reference signal that is broadcast to each device. Pawig et al. [5] applied a reference signal for estimating the SRO between input and output channels in an echo cancellation system. It is noted that in the aforementioned works, the synchronization schemes rely on broadcasts of reference signals, which requires dedicated hardware, protocols, and/or communication channels. Reference-free ('blind') SRO estimation techniques have also been developed, which directly estimate the SRO from the recorded audio signals themselves without using a reference signal. For example, Miyabe et al. [2] developed a blind SRO estimation method based on a maximum likelihood estimation of the sampling frequency mismatch in the STFT domain. In [6] a link between SRO and the Doppler effect was considered and a wideband correlation processor for blind SRO estimation was applied. In [1] the phase-drift between the coherence of the speech-absent segments of signals was used, assuming the availability of a coherent noise source and a voice activity detector. However, these blind SRO estimation methods have a limited accuracy and suffer from robustness issues. In this paper, we aim to tackle both of these issues. For illustration purposes, we will use our SRO estimates in a multi-channel Wiener filter (MWF)-based speech enhancement task [7]. This also requires a compensation for the SRO.

In [1, 6], after the estimation of the SRO, the reference signal is re-sampled in the time domain using the Lagrange polynomials interpolation to attenuate the effect of SRO. While effective, this method is computationally expensive. In this paper, we aim to avoid such an explicit time-domain re-sampling, and instead we immediately compensate for the SRO within the MWF computations in the frequency domain.

The rest of this paper is organized as follows. In Section 2 the problem of SRO estimation is explained. In Section 3, the proposed SRO estimation approach is described. Section 4 briefly describes the applied SRO compensation approach. Our experimental setup and evaluation results are presented in Section 5. The paper ends with conclusions in Section 6.

2. 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 pair. The sound pressure of the i^{th} microphone and its corresponding discrete-

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), 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, The scientific responsibility is assumed by its authors.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

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. Since each node uses a local clock, relative SROs between them are inevitable and are mainly due to the variability of the oscillator in each clock. Therefore, the sampling frequency of the i^{th} microphone is equal to

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

where $|\epsilon_i| \ll 1$ is the relative SRO with respect to the reference sampling rate f_s^{ref} at an arbitrarily chosen reference node. Without loss of generality it is assumed that the sampling rate of the first microphone is the reference sampling rate, 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 audio 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 within the computation of the MWF. The MWF is typically conducted in the short-time Fourier transform (STFT) domain to reduce the computational load, hence we aim for SRO compensation in the STFT domain. The ι^{th} frame $X_i^\iota[k]$ 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-frame length, P is the STFT-frame shift, $w[l]$ is a user-defined window function, and k is the discrete frequency index ranging from 0 to $K-1$.

3. LEAST-SQUARES COHERENCE DRIFT ESTIMATION

In this section, a new SRO estimation approach, referred to as least-squares coherence drift (LCD) estimation, is described.

3.1. Coherence

The coherence of $x_1[n]$ and $x_i[n]$ within frame[1] m of length $\Gamma > K$ is obtained as

$$h_{1,i}^m[k] = \frac{q_{1,i}^m[k]}{\sqrt{q_{1,1}^m[k] q_{i,i}^m[k]}}, \tag{3}$$

where $q_{1,i}^m$ is the cross-spectrum between the signals in node 1 and i and $q_{i,i}^m$ denotes the auto-spectrum of the signals in node i. We define m as the sample index of the mid-frame sample of the frame that is used to compute $h_{1,i}^m[k]$.

All $q_{p,q}^m$ can be estimated using the Welch method [8], which is a common method to estimate periodograms. The Welch method chunks the m-th time frame of length Γ into several overlapping segments of length $K < \Gamma$, and then takes the average of the cross-correlation of the STFT of the segments.

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

Based on the shift-theorem, it is shown that a fixed delay of ϱ_i samples in $x_i[n]$ causes a shift of $\frac{2\pi k \varrho}{K}$ in the coherence phase, i.e.

$$h_{1,i}^m[k;\varrho_i] = h_{1,i}^m[k] \exp\left(\frac{2\pi k \varrho_i}{K} j\right), \tag{4}$$

where $h_{1,i}^m[k;\varrho_i]$ is the coherence between $x_1[n]$ and $x_i[n]$ after the latter is delayed by ϱ samples. Such a fixed delay can incorporate, e.g., an additional acoustic propagation delay when the microphones are not equidistant from the sound source or a fixed time-offset in the clocks of the nodes. However, note that these fixed delays are assumed to be unknown and are in principle absorbed within $h_{1,i}^m[k]$. Equation (4) is merely introduced to explain the shift theorem and for notation purposes in the sequel.

An SRO between $x_1[n]$ and $x_i[n]$ causes a linearly increasing delay in the time-domain, and hence a linearly increasing phase-shift in the coherence. The sample delay of $x_i[n]$ in the mid-frame sample (m) caused by the SRO (ϵ_i) w.r.t. $x_i[n]$ is denoted as ρ_i^m, and can be computed as

$$\rho_i^m = f_s^{\text{ref}} \left[\frac{m}{f_s^{\text{ref}}} - \frac{m}{(1+\epsilon_i) f_s^{\text{ref}}}\right] \approx m\epsilon_i. \tag{5}$$

The SRO induced delay is equal to (5) 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 of the reference signal and the signal with SRO ($h_{1,i}^m[k;\varrho_i;\rho_i^m]$) is difficult. However, assuming the maximum drift caused by the SRO inside a single coherence-frame is much smaller than 1 sample, i.e. $|\Gamma\epsilon_i| \ll 1$, the coherence $h_{1,i}^m[k;\varrho_i;\rho_i^m]$ can be approximated as

$$\begin{aligned} h_{1,i}^m[k;\varrho_i;\rho_i^m] &\approx h_{1,i}^m[k;\varrho_i+\rho_i^m] \\ &= h_{1,i}^m[k] \exp\left[\frac{2\pi k(\varrho_i+\rho_i^m)}{K} j\right]. \end{aligned} \tag{6}$$

3.2. LCD estimation

To estimate the SRO, we exploit the phase-drift of the coherence over different frames. To remove the effect of any arbitrary fixed delay ϱ_i, we use the phase difference between the coherence of two consecutive frames, such that, relying on (6),

$$\angle \frac{h_{1,i}^m[k;\varrho_i;\rho_i^m]}{h_{1,i}^{m-\Lambda}[k;\varrho_i;\rho_i^{m-\Lambda}]} \approx \frac{2\pi k(\rho_i^m - \rho_i^{m-\Lambda})}{K} = \frac{2\pi k \Lambda}{K}\epsilon_i, \tag{7}$$

where the $\angle$ denotes the phase of the phasor and Λ is the frame-shift (the last step follows from (5)). Therefore, the phase difference between the coherence of two different frames with frame shift Λ increases linear by the SRO.

To improve the estimation accuracy, we repeat this procedure for J consecutive frames and formulate the results in the following matrix form:

$$\mathbf{A} = \mathbf{B}\epsilon_i \tag{8}$$

where $\mathbf{A}$ is a matrix of size $K \times J$ and its elements $a_{k,j}$ are obtained as follows

$$a_{k,j} = \angle \frac{h_{1,i}^{m-j\Lambda}[k; \varrho_i; \rho_i^{m-j\Lambda}]}{h_{1,i}^{m-(j-1)\Lambda}[k; \varrho_i; \rho_i^{m-(j-1)\Lambda}]} \tag{9}$$

and $\mathbf{B}$ is a matrix of dimension $K \times J$ and its rows $\mathbf{b}_k$ are obtained as follows

$$\mathbf{b}_k = \frac{2\pi k\Lambda}{K}\mathbf{1}^T, \tag{10}$$

where $\mathbf{1}^T$ is a row vector of dimension J with all elements equal to 1.

A least-squares (LS) estimation[2] of ϵ_i can be obtained by solving

$$\hat{\epsilon}_i^{\text{LS}} = \arg\min_{\epsilon_i} \|\mathbf{A} - \mathbf{B}\epsilon_i\|_F^2, \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm (i.e., the squareroot of the sum of the squared entries). However, based on experiments, it is found that the data of several frequency bins, i.e. the data in some rows of $\mathbf{A}$, is not reliable. Furthermore, (8) compares phases, which are defined over a circular topology, i.e., a phase of 2π is the same as a zero-phase. However, for phases that are close to this phase ambiguity boundary, small errors due to noise may result in large absolute differences in (11), and hence minimization of (11) may result in an inaccurate estimation of the SRO.

Therefore, we apply a two-step procedure, which first performs an outlier detection and removal procedure, followed by the LS estimation (11). In the first step, we make a rough estimation of ϵ_i through the following least-absolute (LA) minimization:

$$\hat{\epsilon}_i^{\text{LA}} = \arg\min_{\epsilon_i} \|\mathbf{A} - \mathbf{B}\epsilon_i\|_1, \tag{12}$$

where $\|\cdot\|_1$ denotes the L_1-norm (i.e., the sum of the absolute value of the entries). The LA estimation is known to be more robust against outliers, and will give small weights to errors caused by the outliers. In the second step, the outliers are detected. The rows corresponding to any element of the data matrix $\mathbf{A}$ satisfying the following condition are considered as an outlier:

$$|a_{k,j} - b_{k,j}\hat{\epsilon}_i^{\text{LA}}| > 1.5\sigma_j, \tag{13}$$

where $|\cdot|$ denotes absolute value and σ_j is the standard deviation of elements in the j^{th} column of the residual matrix $\mathbf{R} = \mathbf{A} - \mathbf{B}\hat{\epsilon}_i^{\text{LA}}$.

After detection and removal of the outliers, a more accurate SRO estimation can be found this time using a LS estimation for computational convenience:

$$\hat{\epsilon}_i^{\text{LS}} = \arg\min_{\epsilon_i} \|\mathcal{A} - \mathcal{B}\epsilon_i\|_F^2, \tag{14}$$

[2]Other distance measures can also be applied for this problem. The procedure of solving a similar estimation problem using KullbackLeibler divergence is explained in [9]

where matrices $\mathcal{A}$ and $\mathcal{B}$ are equivalents of $\mathbf{A}$ and $\mathbf{B}$ after removing their outlier rows. Finally, the optimal solution of (14) can be obtained as follows:

$$\hat{\epsilon}_i^{\text{LS}} = \frac{(\vec{\mathcal{B}})^T \vec{\mathcal{A}}}{(\vec{\mathcal{B}})^T \vec{\mathcal{B}}}, \tag{15}$$

where T denotes transpose and $\vec{\cdot}$ denotes vectorization, where columns of a matrix are stacked on top of each other.

In the sequel, we refer to this approach as the least coherence drift (LCD) estimation method.

Remark: In LCD estimation, we assume that the phase of the coherence function is fixed between consecutive frames, when there is no SRO. In a single-source scenario, this implies spatial stationarity, but does not require spectral stationarity. In a multi-source scenario, we assume that each frequency bin is dominated by a single coherent source, such that the single-source assumption holds for each individual frequency bin[3]. Frequency bins that do not satisfy this assumption are effectively removed by the outlier removal operation.

4. SRO COMPENSATION

For the SRO compensation, two complementary operations are performed: skipping critical samples in the time-domain and phase compensation in the frequency domain based on the shift-theorem. We will explain why both have to be applied in a hybrid compensation framework. In this section, an estimation of the SRO is assumed to be available from the LCD method described in Section 3.

4.1. Time-domain operation

Assume w.l.o.g. that the i^{th} node has a positive relative SRO ϵ_i with respect to the reference node. 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:

$$f_s^{\text{ref}} \left[\frac{n_\tau}{f_s^{\text{ref}}} - \frac{n_\tau}{(1+\epsilon_i) f_s^{\text{ref}}} \right] > 1. \tag{16}$$

Therefore, this event will take place at approximately $n_\tau = \epsilon_i^{-1}$ (using the same approximation as in (5)). 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.

4.2. Frequency-domain operation

The SRO compensation in the frequency domain is performed based on the shift-theorem of the STFT. The shift-theorem

[3]This is a common assumption in speech processing and often holds when the number of sources is limited. In Section 5, we show that our method indeed works in a multi-source scenario.

states that a fixed delay of ϱ_i samples in $x_i^\iota[n]$ causes a phase rotation of $\frac{2\pi k\varrho}{K}$ 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 between the signals, not a constant one and we approximate a linearly increasing phase-shift as a fixed one assuming the drift caused by the SRO within a single STFT frame is much smaller than 1 sample, i.e. $|L\epsilon_i| \ll 1$. Therefore, the approximation is more accurate for a small frame-size and a small SRO. For each frame we calculate the SRO delay at the mid-frame sample based on the estimated SRO and obtain the corresponding phase-rotation $\frac{2\pi km\hat{\epsilon}_i^{\mathrm{LS}}}{K}$. After the frame is transformed to the frequency domain, the k^{th} frequency bin is multiplied by $\exp\left(-j\frac{2\pi km\hat{\epsilon}_i^{\mathrm{LS}}}{K}\right)$ 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.

4.3. Hybrid operation

If the frequency-domain operation would be applied alone, the signals at two different nodes drift more and more away from each other as the time increases, until there is hardly any overlap between the signal content of the current STFT frame in both nodes. Of course a phase rotation in the STFT domain cannot compensate for this. Therefore, we should also apply the time-domain operation to realign the frames.

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 due to the time-domain compensation, there is still 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 in subsections 4.1 and 4.2 are essential to compensate for the SRO effects in, e.g., a speech enhancement algorithm. The hybrid compensation is in fact split up in realigning the frames (coarse-scale compensation) and the compensation of small phase-shifts (fine-scale compensation).

The implementation in MWF is completely similar to what was discussed for each operation independently, the only difference is that an additional frequency domain compensation is applied each time a sample is skipped (compensating for a 1-sample delay corresponding to a phase shift of $\frac{2\pi k}{K}$).

5. VALIDATION

In this section, the accuracy of the proposed LCD estimation and SRO compensation is investigated.

5.1. Simulation setup

A $5m\times 5m\times 3m$ room with a reflection coefficient of 0.3 at all the walls is simulated. 25 speech signals are used comprised of read sentences from the Hearing in Noise Test (HINT) database [10]. Two microphones at positions [4.5 1 0.5] and [0.5 1 0.5] are considered, where the sampling rate of the reference microphone is set to 8kHz and the sampling rate of the second microphone is subject to offsets of 1, 10 and 100 parts per million (ppm) of the sampling rate of the first microphone, where ppm $= 10^{-6}$. In this simulation, re-sampling is performed using Sound eXchange (SOX) software[1]. The input SNR is 9.85 dB at the reference microphone and 10.45 dB at the second microphone. For each combination of SRO and speech signal, 4 Monte-Carlo experiments are conducted, where the locations of the speech and noise sources are randomly selected yielding a total of 100 Monte-Carlo runs per SRO.

5.2. SRO estimation

For the SRO estimation, we use coherence frames of length $\Gamma = 8192$ with 50% overlap. Coherence is calculated using the Welch method [8] with segment size of $K = 4096$, using a Hamming window, and with 75% overlap, i.e., P is equal to 4096/4=1024 samples.. The number of applied consecutive frames is 7, hence the number of columns of the data matrix $\mathbf{A}$ is $J = 6$.

The LCD estimation approach, is compared with a benchmark algorithm [1], which is also based on coherence of signals, here referred to as average coherence drift (ACD) estimation, since it is based on averaging rather than a robustified least-squares estimation.

Table 1 lists the mean absolute error E_{MA} of the SRO estimation. To study the effect of the outlier removal (OR) procedure, we report the results of LCD and ACD estimation with and without OR. Note that the original ACD does not use an OR method, and it assumes a coherent noise source as well as the availability of a VAD to estimate the SRO on speech-absent segments. In our experiments we extend it with OR, and apply it on all segments to obtain a fair comparison.

Table 1 demonstrates the accuracy of the LCD estimation compared to ACD. It is also observed that the proposed OR improves the results of both the LCD estimation and the benchmark ACD estimation.

Figure 1 illustrates the accuracy of the SRO estimation for different observation lengths (note that this impacts J, i.e., the number of data points that are used in the least-squares estimation). This figure shows that the accuracy of the LCD increases by increasing the observation length. Of course, this comes at the cost of decreased tracking capabilities and increased algorithmic delays.

5.3. SRO compensation for noise reduction

For noise reduction, the speech-distortion weighted MWF (SDW-MWF) [7] with square-root Hann window of size 1024, 50% window overlap and a forgetting factor of 0.997 is applied. Table 2 shows the output SNR of SDW-MWF

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

Table 1: *The E_{MA} SRO estimation using the proposed method and baseline system (ppm).*

System Configuration	LCD		ACD	
	with OR	without OR	with OR	without OR
100	0.64	15.95	9.22	19.26
10	0.04	1.37	8.15	14.03
1	0.04	0.50	8.30	19.57

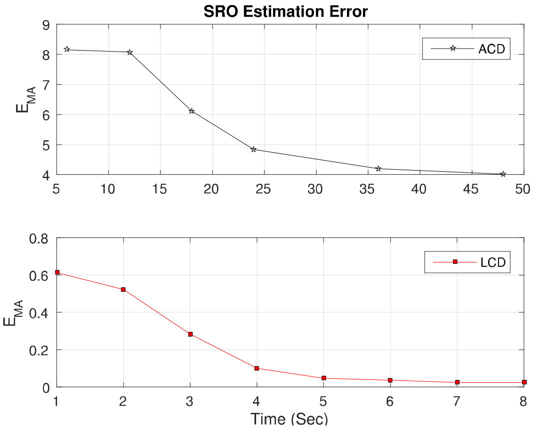

Fig. 1: The E_{MA} of SRO estimation versus observation length.

Table 2: *The SNR with and without compensation (dB).*

PPM	Regular	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

without SRO compensation (in the column 'Regular') and after SRO compensation using the LCD approach (in the column 'Compensated'). To investigate the effect of errors in the estimation of the SRO on the proposed compensation approach, 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 other rows shows that the SRO degrades the performance of the MWF, and this degradation increases with the SRO[4]. Therefore, a compensation method to avoid this performance-loss is necessary. The result of compensation with the true SRO is very close to the perfect case (no SRO) indicating the efficiency of the applied compensation scheme. When the SRO is estimated by means of the LCD estimation approach a similarly high SNR is obtained. Therefore it is confirmed that the presented approach can effectively recover the degradation of MWF performance caused by SRO.

[4]It is noted that the experiments are performed with an adaptive MWF, 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.

6. CONCLUSION

A new approach for blind sampling rate offset (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 rising phase-shift in the short-time Fourier transform (STFT) domain. After outlier removal, the obtained coherence drift, which has a linear relation with the SRO, is attained from a least-squares set of equations. Simulation results show the effectiveness of the proposed least-squares coherence drift (LCD) approach for SRO estimation. We have finally demonstrated that the LCD estimation scheme along with a hybrid compensation approach can eliminate the SRO-induced performance-loss of the multichannel Wiener filter (MWF) in a speech enhancement task.

REFERENCES

[1] Shmulik Markovich-Golan, Sharon Gannot, and Israel 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.

[2] Shigeki Miyabe, Nobutaka Ono, and Shoji Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in stft domain," in *Acoustics, Speech and Signal Processing, IEEE Int. Conf.*, 2013, pp. 674–678.

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

[4] Stefan Wehr, Igor Kozintsev, Rainer Lienhart, and Walter Kellermann, "Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation," in *Multimedia Software Engineering, IEEE 6th Int. Symp.*, 2004, pp. 18–25.

[5] Matthias Pawig, Gerald Enzner, and Peter 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.

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

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

[8] Peter 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.

[9] M.H. 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.

[10] Michael Nilsson, Sigfrid D Soli, and Jean 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.