BLIND SAMPLING RATE OFFSET ESTIMATION AND COMPENSATION IN WIRELESS ACOUSTIC SENSOR NETWORKS WITH APPLICATION TO BEAMFORMING

Shmulik Markovich-Golan[1]*, Sharon Gannot*[1] *and Israel Cohen*[2]

[1] Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 52900, Israel

[2] Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel

shmulik.markovich@gmail.com; sharon.gannot@biu.ac.il
icohen@ee.technion.ac.il

ABSTRACT

Beamforming methods for speech enhancement in wireless acoustic sensor networks (WASNs) have recently attracted the attention of the research community. One of the major obstacles in implementing speech processing algorithms in WASN is the sampling rate offsets between the nodes. As nodes utilize individual clock sources, sampling rate offsets are inevitable and may cause severe performance degradation. In this paper, a blind procedure for estimating the sampling rate offsets is derived. The procedure is applicable to speech-absent time segments with slow time-varying interference statistics. The proposed procedure is based on the phase drift of the coherence between two signals sampled at different sampling rates. Resampling the signals with Lagrange polynomials interpolation method compensates for the sampling rate offsets. An extensive experimental study, utilizing the transfer function generalized sidelobe canceller (TFGSC), exemplifies the problem and its solution.

Index Terms— WASN, beamforming, synchronization

1. INTRODUCTION

The topic of WASN has drawn attention of an increasing number of researchers in recent years [1]. Thanks to technological advances in nano-technology and communications, construction of WASNs comprising small-scale and energy efficient nodes has become feasible. Each node comprises sensors, a processor, actuators and a wireless communication module. The nodes sense a physical phenomenon propagating in space, and by collaborating, they aim to achieve a mutual goal of estimating a common parameter or of enhancing a desired signal.

As each node relies on its own clock source, synchronization offsets are inevitable. Elson and Kay [2] consider the time synchronization problem in wireless sensor networks (WSNs). They define several aspects of the problem, namely phase and frequency synchronization problems of the nodes' clock sources. They propose an algorithm, denoted reference-broadcast synchronization (RBS), for synchronizing the clocks in the WSN. Wehr et al. [3] consider the synchronization problem in distributed beamforming for blind source separation (BSS). They propose an algorithm for estimating the sampling rate offsets based on a modulated reference signal which is broadcast in the WASN. The estimated sampling rate offsets are then compensated for by resampling at the nodes. Pawig et al. [4] consider the problem of a sampling rate offset between the analog to digital converter (ADC) and the digital to analog converter (DAC) in a single channel echo cancelation system. They utilize the reference data to estimate the sampling rate offset, and propose a combined time-recursive algorithm for tracking both the room impulse response (RIR) and the sampling rate offset. Subsequently, the reference signal is resampled using the Lagrange polynomials interpolation method [5].

In this paper, we address the problem of blind sampling rate offset estimation and compensation in beamforming applications. As we are interested in blind scenarios, no reference signals (neither speech from a far-end nor a common modulated signal) are assumed to be available. To simplify the exposition, a WASN, for which all microphone signals are available at the fusion center, is considered. The same sampling rate compensation method is applicable in distributed constellations as well. The proposed method utilizes speech-absent time segments, where the interference statistics is assumed slowly time-varying, and the sampling rate offsets are assumed fixed [2]. An estimation procedure for the sampling rate offsets is proposed based on the coherence between the received signals. Following [4], we propose to resample the microphone signals with the Lagrange polynomials interpolation method. We incorporate the proposed sampling rate offset estimation and compensation scheme in the TFGSC, introduced by Gannot et al. [6].

The structure of the paper is as follows. In Sec. 2, the problem is defined. An algorithm for estimating the sampling rate offsets is derived in Sec. 3. The Lagrange polynomials interpolation method is described in Sec. 4. In Sec. 5, the sampling rate offset estimation and compensation is incorpo-

rated in the TFGSC-beamformer (BF), and its performance is evaluated in an extensive experimental study. Conclusions are drawn in Sec. 6.

2. PROBLEM FORMULATION

Consider a WASN comprising N nodes in a reverberant and noisy enclosure. The nth node comprises M_n microphones, and the total number of microphones is $M = \sum_{n=1}^{N} M_n$. Denote the $M_n \times 1$ microphone signals vector in continuous time at the nth node by:

$$\mathbf{z}_n(t) = \mathbf{x}_n(t) + \mathbf{v}_n(t). \tag{1}$$

Let $z_{n,r}(t)$ be the continuous-time received signal at microphone r in node n, and $x_{n,r}(t)$, $v_{n,r}(t)$ the respective speech and noise components. It is assumed that the received noise contains a spatially coherent component, and that a perfect voice activity detector (VAD) is available. Hence, noise-only segments can be determined, and hereafter only these segments will be considered.

Since nodes utilize individual clock sources, sampling rate differences are inevitable. Denote the sample rate at the nth node by $f_{s,n}$. Without loss of generality, the sampling rate offsets are defined with respect to the first node by:

$$f_{s,n} = (1 + \epsilon_n) f_s \tag{2}$$

where ϵ_n is the relative sampling rate offset and $f_s = f_{s,1}$ is the sampling rate at the first node.

The sampled microphone signals at the nth node are denoted by $\mathbf{v}_n[p] = \mathbf{v}_n(pT_{s,n})$ where $T_{s,n} = \frac{1}{f_{s,n}}$ is the sampling period at the nth node and p is the sample index. The notations $(\bullet)$ and $[\bullet]$ are used for denoting continuous-time and discrete-time functions, respectively. We assume that the WASN is fully connected and that all sampled microphone signals are transmitted to the fusion center, selected, without loss of generality, as the first node.

Let $V_{n,m}^{\ell}[k]$ be the discrete short time Fourier transform (STFT) of the rth microphone signal of the nth node at the ℓth sample, using the analysis window $c[p]$:

$$V_{n,r}^{\ell}[k] = \sum_{p=-\infty}^{\infty} v_{n,r}[p]\, c[p-\ell] \exp\left(-\mathrm{j}\frac{2\pi k p}{K}\right) \tag{3}$$

where $k = 0, 1, \ldots, K-1$. Throughout the paper, lowercase and uppercase letters denote functions in time and frequency domains, respectively.

Since $\mathbf{v}_n[p];\ n = 1, \ldots, N$ are sampled with different sampling rates, straightforward application of a beamforming algorithm may result in a degraded performance. In Sec. 5 we will demonstrate this degradation. In the next section, a sampling rate offset estimation procedure is derived.

3. SAMPLING RATE OFFSET ESTIMATION

We turn now to the derivation of a procedure for estimating ϵ_n, the sampling rate offset of the nth node. In Sec. 3.1 some notations are defined, and in Sec. 3.2 the estimation procedure is derived.

3.1. Notation

Consider microphones s and r at the first and the nth nodes, respectively. Let $R_{s,r}(\tau)$ and $\theta_{s,r}(\zeta)$ be their cross-correlation and cross-spectrum:

$$R_{s,r}(\tau) = \mathrm{E}\{v_{1,s}(t)\, v_{n,r}(t-\tau)\} \tag{4a}$$

$$\theta_{s,r}(\zeta) = \int_{-\infty}^{\infty} R_{s,r}(\tau) \exp(-\mathrm{j}\zeta\tau)\, \mathrm{d}\tau \tag{4b}$$

where we assume that $v_{1,s}(t)$ and $v_{n,r}(t)$ are jointly wide-sense stationary (WSS) processes. Similarly to (4a), $R_{s,r}(\tau+\Delta) = \mathrm{E}\{v_{1,s}(t)\, v_{n,r}(t-\tau-\Delta)\}$, and the corresponding cross-spectrum, denoted $\theta_{s,r}^{\Delta}(\zeta)$, is obtained by applying Fourier transform properties:

$$\theta_{s,r}^{\Delta}(\zeta) = \exp(\mathrm{j}\zeta\Delta)\, \theta_{s,r}(\zeta). \tag{5}$$

Due to sampling rate offset, the time difference between the ℓth sample at microphone s and microphone r is approximately $\ell T_s - \ell T_{s,n} \approx \ell T_s \epsilon_n$, where we replaced $T_{s,n} = \frac{T_s}{1+\epsilon_n}$ with its first-order Taylor series approximation $T_s(1-\epsilon_n)$. Let $\theta_{s,r}^{\ell}[k]$ be the discrete cross-spectrum of the sampled microphones s and r at the ℓth sample. Assuming that the continuous cross-spectrum is band-limited by $\frac{f_s}{2}$ and that the support of the window $c[n]$ is long enough, the following equivalence between the discrete and the continuous spectra holds:

$$\theta_{s,r}^{\ell}[k] = \theta_{s,r}^{\ell T_s \epsilon_n}\left(\frac{2\pi k f_s}{K}\right). \tag{6}$$

Now, applying the relation in (5) to (6) yields:

$$\theta_{s,r}^{\ell}[k] = \exp\left(\mathrm{j}\frac{2\pi k \ell \epsilon_n}{K}\right)\theta_{s,r}[k] \tag{7}$$

where

$$\theta_{s,r}[k] = \theta_{s,r}\left(\frac{2\pi k f_s}{K}\right). \tag{8}$$

Let $\theta_{s,s}[k]$ and $\theta_{r,r}[k]$ be the auto-spectra of microphones s and r, respectively. Denote the coherence between microphones s and r at the ℓth sample by:

$$\gamma_{s,r}^{\ell}[k] = \frac{\theta_{s,r}^{\ell}[k]}{\sqrt{\theta_{s,s}[k]\theta_{r,r}[k]}} \tag{9}$$

for $k = 0, 1, \ldots, K-1$. Substituting (7) in (9), the coherence is given by:

$$\gamma_{s,r}^{\ell}[k] = \alpha_n^{\ell} \gamma_{s,r}[k] \tag{10}$$

where

$$\gamma_{s,r}[k] = \frac{\theta_{s,r}[k]}{\sqrt{\theta_{s,s}[k]\theta_{r,r}[k]}} \tag{11a}$$

$$\alpha_n = \exp\left(\mathrm{j}\frac{2\pi k \epsilon_n}{K}\right). \tag{11b}$$

3.2. Estimation

We propose the following procedure for estimating $\epsilon_n;\ n = 2, \ldots, N$, given P_s samples of the microphone signals. For each $n = 2, \ldots, N$ the coherence between the microphones at the first and the nth nodes at samples $\ell = i \times P;\ i = 0, 1, \ldots, I-1$ is estimated by using the Welch method with K samples discrete Fourier transform (DFT), where $I = \lfloor \frac{P_s}{P} \rfloor$. We assume that the frequency offsets between the transformed microphone signals are negligible. The estimated $M_1 \times M_n$ cross-coherence matrix at sample iP is denoted by $\hat{\mathbf{\Gamma}}_{1,n}^{iP}[k]$. Assuming that the estimation error is low $\hat{\gamma}_{s,r}^{iP}[k] \approx \gamma_{s,r}^{iP}[k]$, where $\hat{\gamma}_{s,r}^{iP}[k]$ is the element in the sth row and the rth column of the matrix $\hat{\mathbf{\Gamma}}_{1,n}^{iP}[k]$. Considering (10), we note that $\alpha_n^P = \frac{\gamma_{s,r}^{iP}[k]}{\gamma_{s,r}^{(i-1)P}[k]}$.

Assume that the sampling rate offset is bounded to $|\epsilon_n| < \epsilon_{\max}$, and define

$$k_{\max} = \frac{K}{2P\epsilon_{\max}}. \tag{12}$$

Note that for $1 \leq k \leq k_{\max}$ it is guaranteed that the phase difference between $\gamma_{s,r}^{iP}[k]$ and $\gamma_{s,r}^{(i-1)P}[k]$ is bounded in the range $[-\pi, \pi]$. Therefore, we propose to estimate ϵ_n by averaging the results obtained from all available microphone couples:

$$\hat{\epsilon}_n = \frac{1}{M_1 M_n} \sum_{s=1}^{M_1} \sum_{r=1}^{M_n} \hat{\epsilon}_{n,s,r} \tag{13}$$

where $\hat{\epsilon}_{n,s,r}$ is the estimate of the sampling rate offset derived from microphones s and r. It is obtained by averaging the phase differences of consecutive estimates of $\hat{\gamma}_{s,r}^{iP}[k]$ and $\hat{\gamma}_{s,r}^{(i-1)P}[k]$ for all k in the allowable range with a proper frequency dependent normalization factor:

$$\hat{\epsilon}_{n,s,r} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} \frac{K}{2\pi P k} \angle \left\{ \frac{1}{I-1} \left(\sum_{i=1}^{I-1} \frac{\hat{\gamma}_{s,r}^{iP}[k]}{\gamma_{s,r}^{(i-1)P}[k]} \right) \right\}. \tag{14}$$

Note that the averaging is applied in both time and frequency.

In the following section, given estimates of $\epsilon_n;\ n = 2, \ldots, N$, we describe a procedure, applied by the fusion center, that compensates for sampling rate offsets by resampling the microphone signals.

4. RESAMPLING WITH LAGRANGE POLYNOMIALS INTERPOLATION

Consider the rth microphone signal of the nth node, i.e. $z_{n,r}[p]$. Given an estimate of the sampling rate offset at the nth node, $\hat{\epsilon}_n$, the signal is resampled to the sampling rate of the fusion center, f_s, by a fourth order Lagrange polynomials interpolation [5]. First, $z_{n,r}[p]$ is interpolated by a factor of 4, and the signal $\tilde{z}_{n,r}[\tilde{p}]$ is obtained. Denote $\dot{p} = \lfloor \frac{4pT_s}{T_{s,n}} \rfloor \approx \lfloor 4p(1+\hat{\epsilon}_n) \rfloor$, the closest interpolated sample index from the left to time pT_s. Then, the resampled value of $z_{n,r}(pT_s)$, denoted $\hat{z}_{n,r}[p]$, is calculated by proper weighting its four neighboring interpolated samples:

$$\begin{aligned}\hat{z}_{n,r}[p] = &\beta_{-1}^p \tilde{z}_{n,r}[\dot{p}-1] + \beta_0^p \tilde{z}_{n,r}[\dot{p}] \\ &+ \beta_1^p \tilde{z}_{n,r}[\dot{p}+1] + \beta_2^p \tilde{z}_{n,r}[\dot{p}+2]\end{aligned} \tag{15}$$

where

$$\eta = 4p(1+\hat{\epsilon}_n) - \dot{p} \tag{16a}$$

$$\beta_{-1}^p = -\frac{\eta(\eta-1)(\eta-2)}{6} \tag{16b}$$

$$\beta_0^p = \frac{(\eta+1)(\eta-1)(\eta-2)}{2} \tag{16c}$$

$$\beta_1^p = -\frac{(\eta+1)\eta(\eta-2)}{2} \tag{16d}$$

$$\beta_2^p = \frac{(\eta+1)\eta(\eta-1)}{6}. \tag{16e}$$

5. EXPERIMENTAL STUDY

Consider the following scenario. A $4\text{m} \times 3\text{m} \times 3\text{m}$ room, with a reverberation time of 300ms is simulated. A desired speaker and Q point source stationary interfering sources are picked up by the microphones, for $Q = 1, \ldots, 4$. Utterances of 75sec with 20% voice activity and a 6dB signal to interference ratio (SIR) are used. Two microphone arrays, each comprises 6 microphones with 5cm spacing, are located close to two perpendicular walls. The sampling rate of the first array is set to 8KHz, whereas the sampling rate of the second array is subject to offsets in the range of $\{-300, -250, \ldots, 300\}$ ppm of the sampling rate of the first array, where ppm $= 10^{-6}$. The sampling rate offsets are simulated using the Lagrange polynomials interpolation method, discussed above. For each combination of sampling rate offset and number of interferences, 5 Monte-Carlo experiments are conducted, where the locations of the sources are randomly selected. The proposed sampling rate estimation and compensation scheme incorporated in the TFGSC [6] is denoted the synchronized TFGSC.

The performances of the regular TFGSC and the synchronized TFGSC are compared in the various scenarios. The relative transfer function (RTF) is estimated once, using the subspace method [7], and is used to construct the fixed BF (FBF) and the blocking matrix (BM), which remain fixed during the entire utterance. The noise canceler (NC) is adapted using the normalized least mean squares (NLMS). The performance criteria are the excess distortion and the excess noise levels with respect to the corresponding TFGSC without a sampling rate offset. The following parameters are used in the proposed sampling rate estimation. The Welch method with a DFT size of 4096, 75% overlap and a Hamming window is applied to 32s speech-absent segments for estimating the auto and cross-covariances $\theta_{s,s}[k]$, $\theta_{r,r}[k]$ and $\theta^{\ell}_{s,r}[k]$. Coherence estimates, $\hat{\mathbf{\Gamma}}^{iP}_{1,n}[k]$, of $I = 6$ time segments with 50% overlap ($P = 128 \times 10^3$), are used for estimating the sampling rate offset, assuming that it is bounded by $|\epsilon_n| \leq 400$ppm. The TFGSC-BF uses a 4096 points STFT with 75% overlap, and the NLMS step is set to $\mu = 0.15$.

The results of the experimental study are summarized next. The standard deviation of the estimated sampling rate offset in the synchronized TFGSC is lower than 3.2ppm in all scenarios. The average signal to distortion ratio (SDR) of the regular TFGSC without a sampling rate offset is 14.8dB. Its corresponding SIRs levels are $34.1, 27.4, 24.4, 23.5$dB for 1-4 interferences, respectively. The excess distortion and excess noise levels in the regular TFGSC are 10.6dB in all scenarios and $9.2, 5.2, 3.6, 3.0$dB for 1-4 interferences, respectively. However, their counterparts in the synchronized TFGSC are significantly lower, 0.3dB excess distortion level and 0.1dB excess noise level in all scenarios. Clearly, the performance of the proposed synchronized TFGSC is equivalent to the regular TFGSC without sampling rate offsets and is highly superior to the regular TFGSC (with sampling rate offsets). The excess noise level in the regular TFGSC with respect to its counterpart without a sampling rate offset is depicted in Fig. 1.

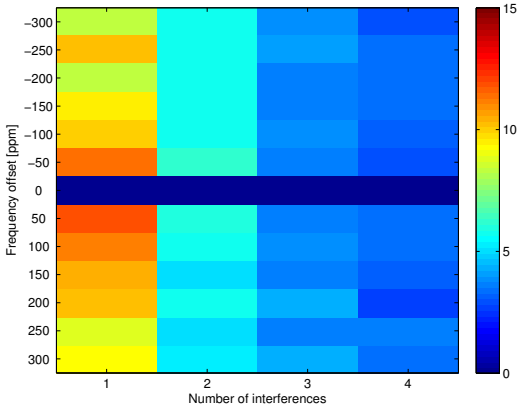

Fig. 1. The excess noise level in the regular TFGSC with respect to its counterpart without a sampling rate offset.

6. CONCLUSIONS

The problem of blind sampling rate offset estimation and compensation in WASN was considered. An interference with slowly time-varying statistics, and speech-absent segments obtained by a perfect VAD were assumed available. A procedure for estimating the sampling rate offsets was derived. It was based on the phase drift in the coherence between two microphone signals, sampled at different sampling rates. The estimated sampling rate offsets were compensated for by resampling the signals with the Lagrange polynomials interpolation method. The estimation and compensation scheme was incorporated in the TFGSC, denoted the synchronized TFGSC. It was shown that, under sampling rate offsets, the synchronized TFGSC significantly outperforms the regular TFGSC. Moreover, the proposed method achieves the performance of the regular TFGSC as if there were no sampling rate offsets.

7. REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communication Mag.*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] J. Elson and R. Kay, "Wireless sensor networks: A new regime for time synchronization," *SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 1, pp. 149–154, 2003.

[3] S. Wehr, I. Kozintsev, R. Lienhart, and W. Kellermann, "Synchronization of acoustic sensors for distributed adhoc audio networks and its use for blind source separation," in *Proc. IEEE 6th Int. Symp. on Multimedia Software Eng.*, Dec. 2004, pp. 18 – 25.

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

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

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[7] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.