PADERBORN UNIVERSITY

Computer Science, Electrical Engineering and Mathematics

NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

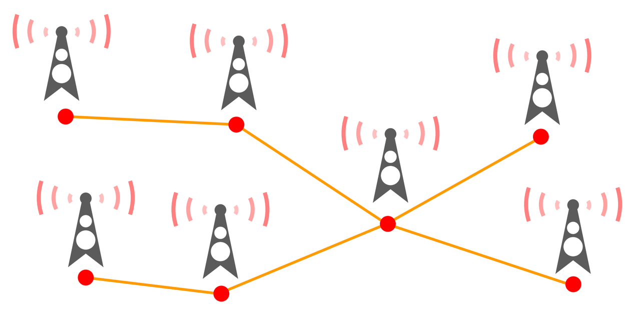
Multi-Stage Coherence Drift Based Sampling Rate Synchronization for Acoustic Beamforming

Joerg Schmalenstroeer, Jahn Heymann, Lukas Drude, Christoph Boeddecker and Reinhold Haeb-Umbach

Paderborn University, Germany
{schmalen, heymann, drude, haeb}@nt.uni-paderborn.de
http://nt.uni-paderborn.de

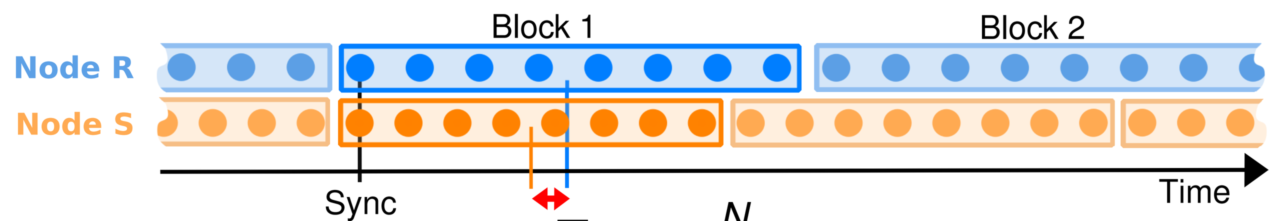
Introduction

- Wireless Acoustic Sensor Networks (WASN)
  - Opportunities
    - Always a microphone close to a source
    - Multi-channel signal recording & processing
  - Challenges & Tasks
    - Synchronization of sampled audio streams
    - Distribution of algorithms & data
    - Condition changes: Environment, network, ...
  - Applications: Telecommunication, environmental monitoring, ...
- Contributions of this paper:
  - Multi-stage synchronization of two audio signals
  - Beamformer normalization for enhanced speech recognition performance



Sampling Rate Offset

- Nodes in the network have individual oscillators
  - Frequencies differ from target frequency with ± 200 ppm
  - Microphone signals are sampled individually
  - Block-oriented processing of time-discrete values $x_i(n)$
- Sampling Rate Offset (SRO) ϵ_{RS} defined with $f_S = (1+\epsilon_{RS}) \cdot f_R$



$$\overline{\tau}_{RS} = \frac{N}{2}\epsilon_{RS}$$

Two node example showing increase of delay $\overline{\tau}_{RS}$ by non-zero SRO

- Rough synchronization (e.g. GCC-PHAT)
- Signal model: Single coherent source with additive noise

$$X_i(l,k) = H_i(k) \cdot S_i(l,k) + V_i(l,k)$$

- Short Time Fourier Transform (STFT) of data streams
  - Frequency bin k (FFT size N) & Block index l (Block size B)

$$X_i(l,k) = \sum_{n=0}^{N-1} w(n) \cdot x_i(n + l \cdot B) \cdot e^{-j\frac{2\pi}{N}kn}$$

- Approximation of inter-node STFT dependency on SRO:
  - Model different sampling starting points with delay τ_{RS}
  - Model SRO by block-wise increasing delay: $(\frac{N}{2} + lB)\epsilon_{RS}$
  - Constant delay within STFT block

$$S_R(l,k) \approx S_S(l,k) \cdot e^{-j\frac{2\pi}{N}[\tau_{RS} + (\frac{N}{2}+lB)\epsilon_{RS}]k}$$

Coherence Drift Estimate

- Complex coherence $\Gamma_{R,S}(l,k) = \frac{\Psi_{R,S}(l,k)}{\sqrt{\Psi_{R,R}(l,k)\cdot\Psi_{S,S}(l,k)}}$, where

$\Psi_{R,S}(l,k) = \frac{1}{N_W}\sum_{\kappa=0}^{N_W-1} X_R(l+\kappa,k) \cdot X_S(l+\kappa,k)^*$ evaluates to

$$\Gamma_{R,S}(l,k) = \frac{H_R(k)H_S^*(k)}{\sqrt{|H_R(k)|^2}\cdot\sqrt{|H_S(k)|^2}} \cdot \frac{\sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa,k)|^2 e^{j\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\overline{|X_{RS}(l,k)|^2}} e^{j\frac{2\pi}{N}[\tau_{RS}+(\frac{N}{2}+lB)\epsilon_{RS}]k}$$

with

$$\overline{|X_{RS}(l,k)|^2} = \sqrt{\overline{|X_R(l,k)|^2}\cdot\overline{|X_S(l,k)|^2}} \text{ and } \overline{|X_R(l,k)|^2} = \sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa,k)|^2 + \frac{|V_R(l+\kappa,k))|^2}{|H_R(k)|^2}$$

- Average Coherence Drift (ACD) [ACD]
  - Ratio of coherence functions $\rightarrow$ SNR information lost

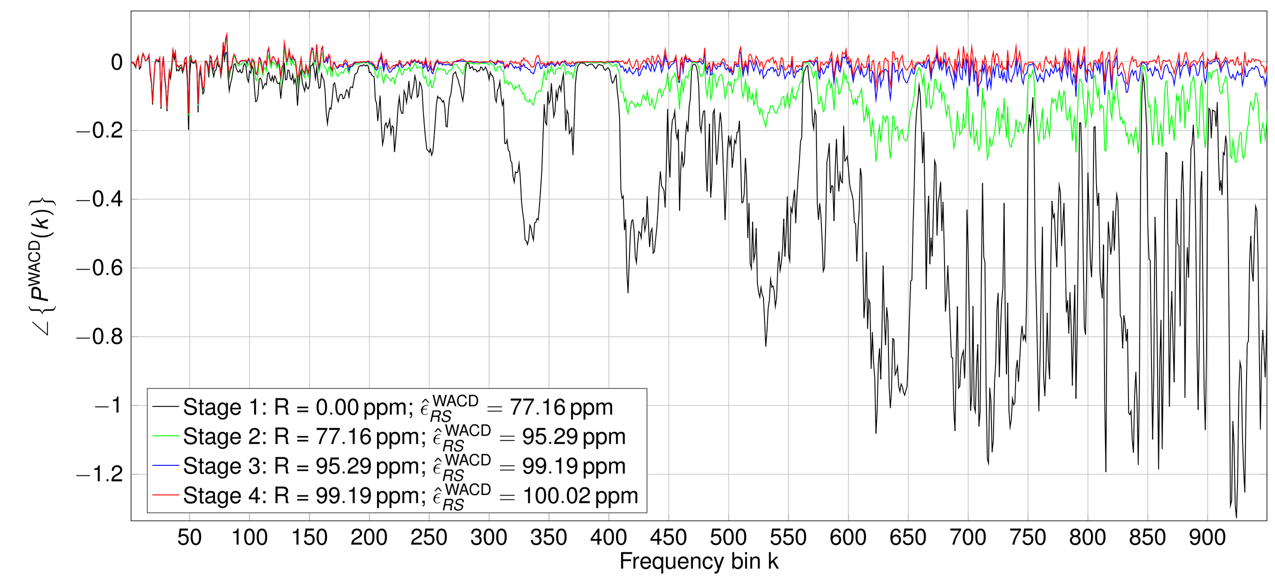
$$\frac{\Gamma_{R,S}(l+p,k)}{\Gamma_{R,S}(l,k)} = \frac{\overline{|X_{RS}(l,k)|^2}}{\overline{|X_{RS}(l+p,k)|^2}} \cdot \frac{\sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa,k)|^2 e^{j\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa+p,k)|^2 e^{j\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}} e^{j\frac{2\pi}{N}(pBk)\epsilon_{RS}}$$

- Weighted Average Coherence Drift (WACD)
  - Complex conj. product of coherence functions $\rightarrow$ Keeps SNR information

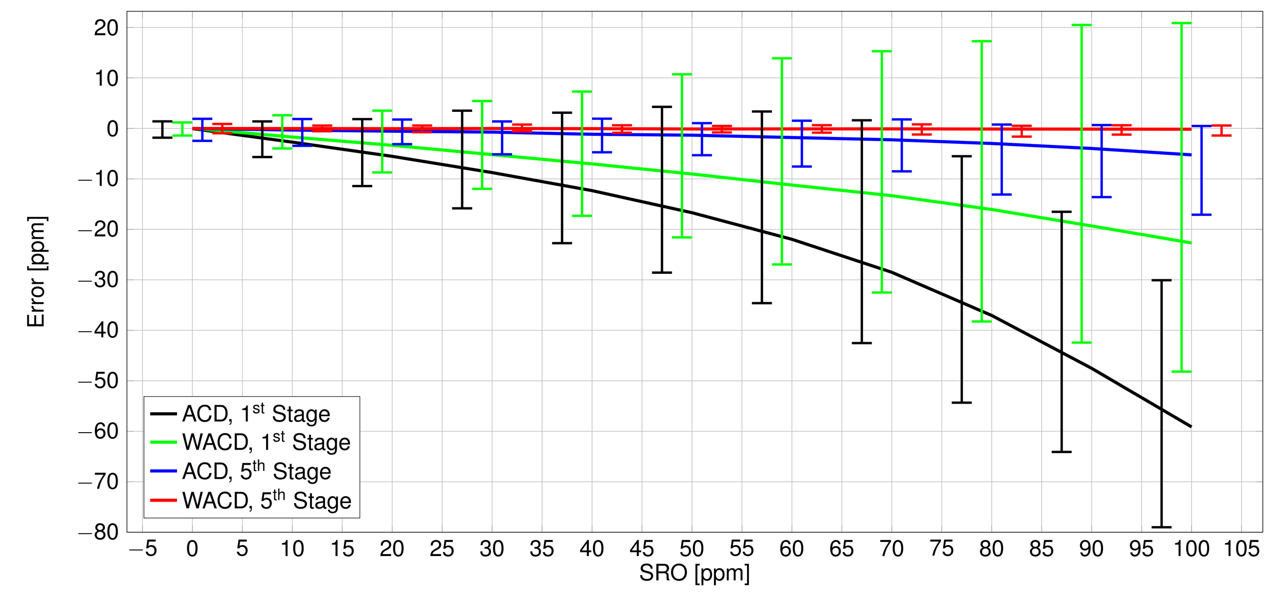
$$\Gamma_{R,S}(l+p,k)\cdot\Gamma^*_{R,S}(l,k) = \frac{\sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa+p,k)|^2 e^{+j\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\overline{|X_{RS}(l+p,k)|^2}} \cdot \frac{\sum_{\kappa=0}^{N_W-1}|S_R(l+\kappa,k)|^2 e^{-j\frac{2\pi}{N}(\kappa B k)\epsilon_{RS}}}{\overline{|X_{RS}(l,k)|^2}} e^{j\frac{2\pi}{N}(pBk)\epsilon_{RS}}$$

Multi-Stage SRO Estimation

- Database: Recordings of TIMIT data with hardware-defined SROs



- Observation: Estimation error bias increases with SRO
  - Bias error mitigation by multi-stage SRO estimation



Generalized Eigenvalue Beamformer

- GEV beamformer: $\mathbf{\Phi_{XX}}(k)\,\mathbf{F}(k) = \lambda\,\mathbf{\Phi_{NN}}(k)\,\mathbf{F}(k)$
  - Spatial correlation matrices $\mathbf{\Phi_{XX}}(k)$ and $\mathbf{\Phi_{NN}}(k)$ estimated using time-frequency masks generated by neural network
- Beamforming vector $\mathbf{F}_{\text{GEV}}(k)$ has arbitrary complex scale factor
  - $\Rightarrow$ Normalization required to prevent "arbitrary distortions"
- Normalization to reference microphone $\tilde{d}$

$$\mathbf{F}'_d(k) = \mathbf{F}_d(k) \cdot \exp(-\mathrm{j}\angle\{\mathbf{F}_{\tilde{d}}(k)\})$$

- Normalization by minimizing group delay

$$\mathbf{F}'(k) = \mathbf{F}(k) \cdot \exp\left(-j\angle\{\mathbf{F}'^{\mathrm{H}}(k-1)\mathbf{F}(k)\}\right)$$

Speech Recognition Results

- CHiME database: eval. test set, real data, 6 channels
  - Noisy environment (≈ 3 dB SNR), utterance lengths (1.2 s - 13 s)
- Resampling: Random SRO (± 50 ppm) for each channel & utterance
  - No information aggregation across consecutive files

Beamformer	GEV-BAN WER [%]			MVDR WER [%]			σ_{SRO}
Normalization	-	Grp.-Delay	Ref.-Mic	-	Grp.-Delay	Ref.-Mic	[ppm]
No Sync.	9.57	9.26	10.02	9.44	8.87	9.68	25.68
ACD,1st Stage	8.45	7.93	8.46	8.49	7.80	8.17	18.34
ACD,10th Stage	7.17	6.65	6.88	7.41	6.73	7.02	7.63
ACD,15th Stage	7.26	6.70	6.87	7.36	6.73	7.05	7.35
WACD,1st Stage	7.55	7.14	7.65	7.81	7.06	7.45	13.81
WACD,10th Stage	7.30	6.71	7.08	7.43	6.72	6.99	6.71
WACD,15th Stage	7.03	6.56	6.77	7.40	6.61	6.92	6.63
CORR	6.80	6.38	6.62	7.28	6.52	6.62	6.29
No Offset	6.92	6.38	6.77	7.24	6.45	6.84	0

[ACD] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), pp. 4-6, 2012.

[CORR] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," IEEE/ACM Transactions on Speech and Language Processing, vol. 24, no. 3, pp. 571-582, 2016.

Conclusions

- Coherence drift based SRO estimation for WASN scenarios
  - Proposed WACD approach: Matched-filter like technique
- New phase normalization technique for GEV beamformer
- SRO compensation and phase normalization improves ASR results

Workshop on Multimedia Signal Processing 2017, London

FOR 2457 Acoustic Sensor Networks

Work supported by DFG under contract no. <Schm 3301/1-1>
Research Unit FOR2457 "Acoustic Sensor Networks"