Source Enumeration and Robust Voice Activity Detection in Wireless Acoustic Sensor Networks

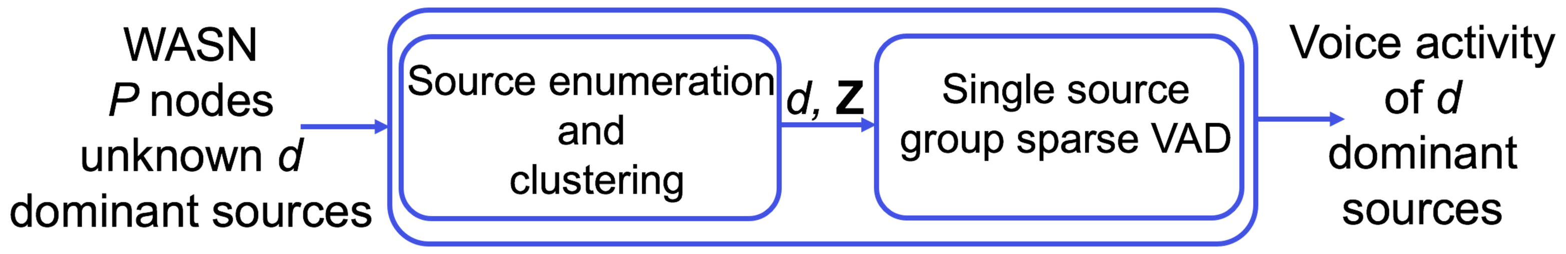
Tanuj Hasija[1], Martin Gölz[2], Michael Muma[2], Peter J. Schreier[1] and Abdelhak M. Zoubir[2]
[1]Signal & System Theory Group, Universität Paderborn, Germany
[2]Signal Processing Group, Technische Universität Darmstadt, Germany

Overview

- Multispeaker voice activity detection (VAD) prerequisite for speech enhancement and noise cancellation in WASN
- Traditional techniques: Based on energy signatures, e.g., Multiplicative non-negative ICA (M-NICA), degrade in performance
  i. as number of speakers increase
  ii. in presence of impulsive noise
- Proposed technique:
  i. clusters nodes around each speaker
  ii. single-source M-NICA with block-sparse penalization
  iii. no knowledge of source/node positions or number of speakers required

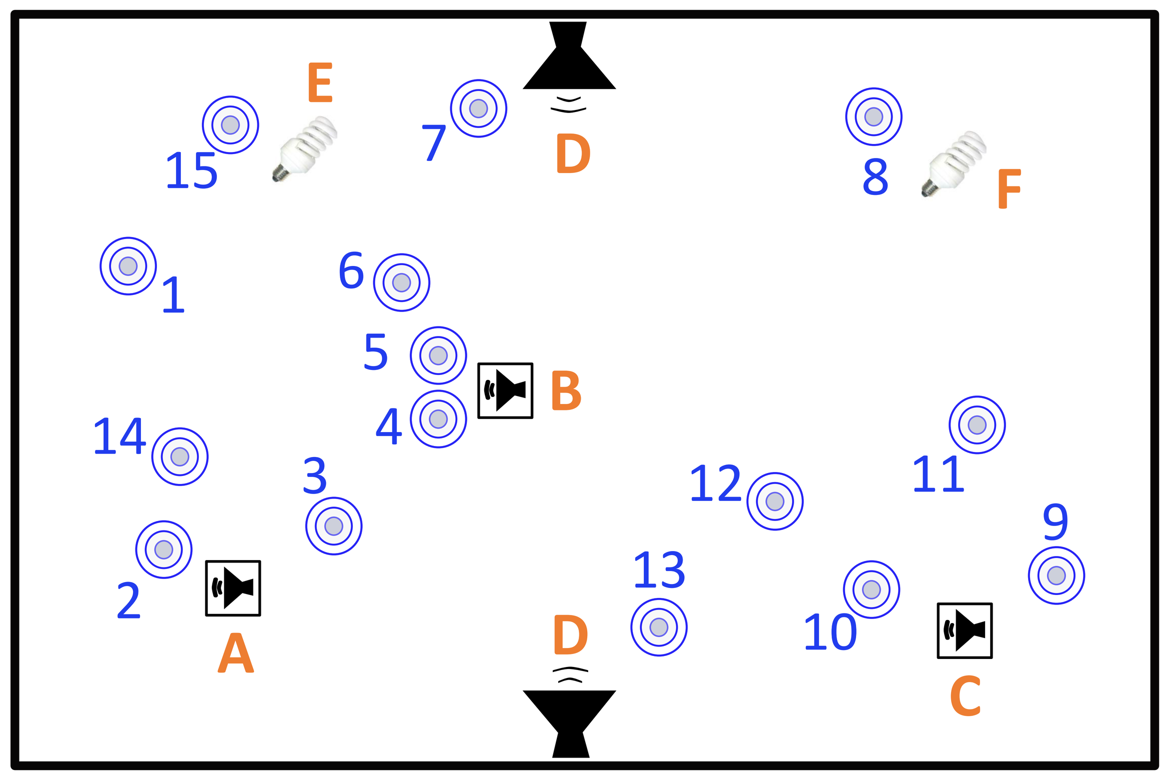
Problem Formulation

- P sensor nodes each with n microphones
- Unknown number of d dominant sources among all q sources
- Goal: Robustly estimate voice activity of d dominant sources



- Cluster information matrix $\mathbf{Z} \in \mathbb{B}^{d \times P}$ whose $\{ij\}$th entry is 1 if ith dominant source is observed by jth node and 0 otherwise

WASN Setup



- 20m × 10m room with $P = 15$ nodes (in blue), $n = 3$ microphones
- $d = 4$ dominant sources (A-D), bulb-flickering noise by sources E and F

Source Enumeration and Node Clustering

- P-channel model: Observed data at fth frequency index, $\mathbf{x}_p(f) \in \mathbb{C}^n$

$$\mathbf{x}_p(f) = \mathbf{A}_p(f)\mathbf{s}_p(f), \qquad p = 1, \dots, P,$$

  $\mathbf{A}_p(f) \in \mathbb{C}^{n \times m_p}$ is acoustic transfer function, $\mathbf{s}_p(f) \in \mathbb{C}^{m_p}$ contains m_p uncorrelated sources
- Strong correlation among nodes observing same dominant source
- Let $\mathbf{R}_{pq} = E[\mathbf{x}_p \mathbf{x}_q^H]$ and $\mathbf{C}_{pq} = \mathbf{R}_{pp}^{-\frac{1}{2}} \mathbf{R}_{pq} \mathbf{R}_{qq}^{-\frac{H}{2}}$
- Composite coherence matrix $\mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1P} \\ \mathbf{C}_{21} & \mathbf{I} & \cdots & \mathbf{C}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{P1} & \mathbf{C}_{P2} & \cdots & \mathbf{I} \end{bmatrix}$
- Main results (adapted from [1]):
  i. $\mathbf{C}$ has exactly d eigenvalues greater than one
  ii. Let $\mathbf{u}^{(i)} = [\mathbf{u}_1^{(i)T}, \mathbf{u}_2^{(i)T}, \dots, \mathbf{u}_P^{(i)T}]^T$, be the ith eigenvector associated with the above eigenvalue of $\mathbf{C}$. The ith source is observed by the pth node iff $\mathbf{u}_p^{(i)} \neq \mathbf{0}$
- Bootstrap-based hypothesis tests for estimating d and $\mathbf{Z}$

No heuristic thresholds required

Group Sparse Voice Activity Detection

- Received energies $\mathbf{y}$ composed of N time samples summarized in matrix $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{a}\mathbf{s} + \mathbf{W}$$

$$\mathbf{Y} \in \mathbb{R}^{n^{(i)} \times N} \qquad \mathbf{a} \in \mathbb{R}^{n^{(i)} \times 1} \qquad \mathbf{s} \in \mathbb{R}^{1 \times N} \qquad \mathbf{W} \in \mathbb{R}^{n^{(i)} \times N}$$

- Singular value decomposition $\mathrm{SVD}(\mathbf{Y}) = \sigma \mathbf{u}\mathbf{v}^\top$
- Information on shape of dominant source energy signature contained in $\mathbf{v}$ [2]
- Divide received energy matrix $\mathbf{Y}$ and $\mathbf{v}$ into L groups of length $N_\mathbf{g}$

$$\mathbf{Y} = \left[\mathbf{Y}_{\mathbf{g},1}, \dots, \mathbf{Y}_{\mathbf{g},L}\right], \qquad \mathbf{Y}_{\mathbf{g},l} \in \mathbb{R}^{n^{(i)} \times N_\mathbf{g}}$$

- Penalize SVD optimization problem to enforce group sparsity

$$\underset{\mathbf{v}}{\operatorname{argmin}} \left\| \sum_{l=1}^{L} \mathbf{Y}_{\mathbf{g},l} - \sigma \mathbf{u}\mathbf{v}_{\mathbf{g},l}^\top \right\| + \lambda_\mathbf{v} \underbrace{\sum_{l=1}^{L} \sqrt{\sum_{k}^{N_\mathbf{g}} |v_{\mathbf{g},l}[k]|^2}}_{\text{mixed } \ell 1/\ell 2 \text{ norm}}$$

- Tuning parameter $\lambda_\mathbf{v}$ determines degree of sparsity in $\mathbf{v}$

Intrinsic Voice Activity Detection without heuristic threshold

- Without sparsity enforcing penalty term, all entries of $\mathbf{v}$ are generally non-zero
- Group sparse structure of $\mathbf{v}$ forces entire groups $\mathbf{v}_{\mathbf{b},l}$ to be zero

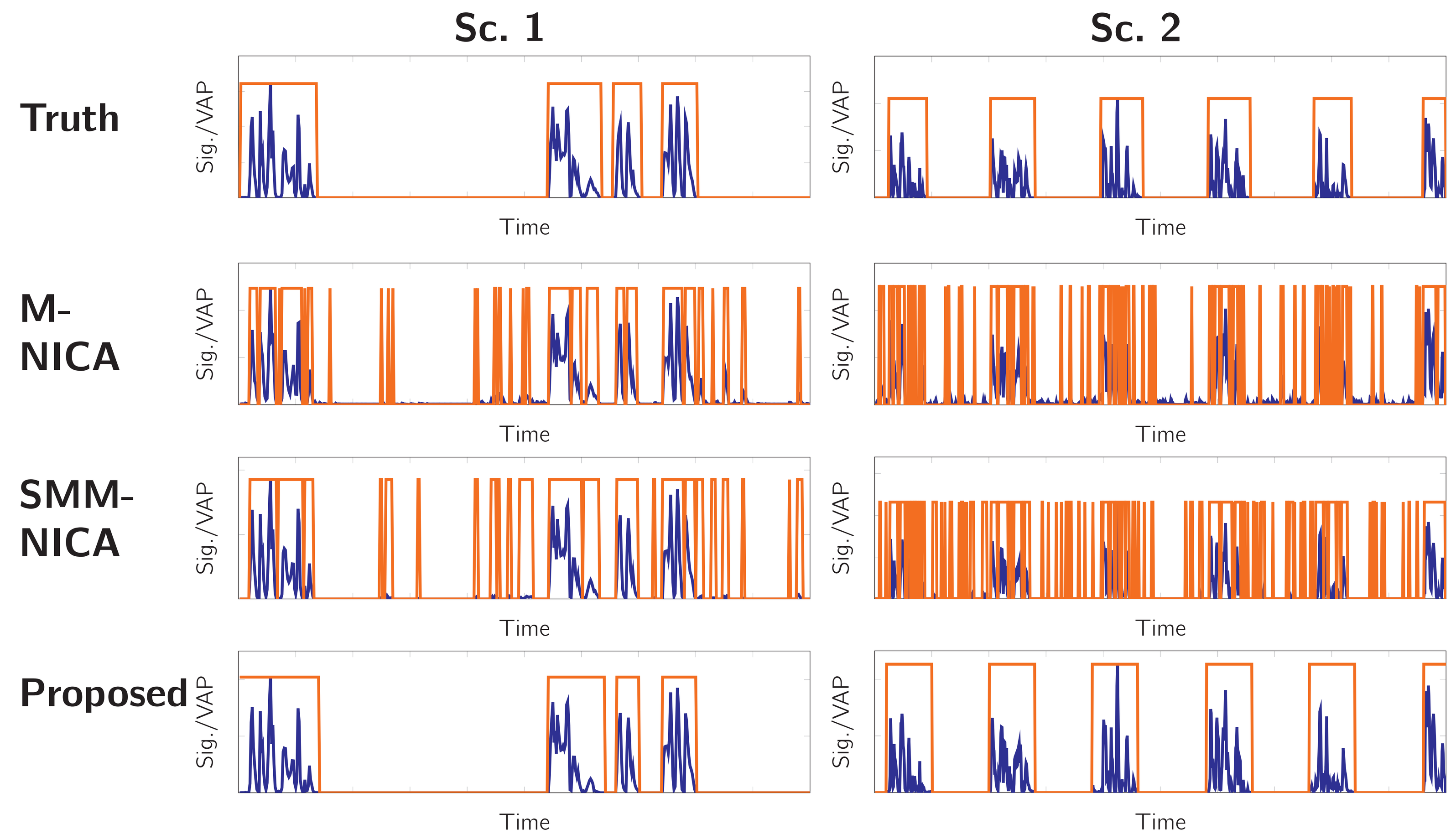
Results

- Each node corrupted by AWGN, sampled at 16kHz
- Hamming window STFT
- Sc. 1: $\hat{d}=3$, 15s
- Sc. 2: $\hat{d}=4$, 30s

Clustering results

Source	Cluster Nodes (Sc. 1)	Cluster Nodes (Sc. 2)
A	2 and 3	2, 3 and 14
B	4, 5 and 6	4, 5 and 6
C	9 and 10	9, 10 and 11
D	Not active	7, 12 and 13

VAD Results



Advantages of Proposed Group-Sparse Method

+ Outperforms standard methods and clustered sparse approach
+ Is scalable to large WASNs with many speakers
+ Handles impulsive noise
+ Does not require heuristic thresholds for node clustering and VAD

Acknowledgements

This research was supported by the German Research Foundation (DFG) under grants SCHR 1384/3-2 and ZO 215/17-2. The WASN speech dataset has been generated within the EU FET-Open Project HANDiCAMS (GA no. 323944).

Bibliography

[1] T. Hasija, C. Lameiro, T. Marrinan & P.J. Schreier,"Determining the Dimension and Structure of the Subspace Correlated Across Multiple Data Sets," *arXiv*, 2019

[2] A. Bertrand & M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," *ICASSP*, 2010

[3] L. K. Hamaidi, M. Muma & A.M. Zoubir, "Multi-Speaker Voice Activity Detection by an Improved Multiplicative Non-Negative Independent Component Analysis with Sparseness Constraints", *ICASSP*, 2017