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 and System Theory Group, Universität Paderborn, Germany
[2]Signal Processing Group, Technische Universität Darmstadt, Germany
Email: {tanuj.hasija, peter.schreier}@sst.upb.de, {goelz, muma, zoubir}@spg.tu-darmstadt.de

***Abstract*— We propose a robust technique for multi-speaker voice activity detection and source enumeration in wireless acoustic sensor networks (WASN). The proposed technique first clusters the nodes that observe a single speaker as dominant source, and then estimates the voice activity of each speaker by introducing a block-sparsity penalizing term in the unmixing problem. The method is scalable in terms of the number of simultaneously active speakers, does not require setting empirical thresholds, and is robust to impulsive noise sources. The results are validated using a WASN with four human speakers and two impulsive noise sources observed by 15 nodes.**

***Index terms*— Group-sparse penalization, multiplicative nonnegative ICA, node clustering, source enumeration, voice activity detection, wireless acoustic sensor network**.

I. Introduction

Wireless acoustic sensor networks (WASNs) provide a next generation system with great potential for new services, e.g., in ambient assisted living, habitat monitoring, smart cities [1], [2]. WASNs combine many comparatively low-resource, distributed nodes with sensing, computing and communication capabilities. Compared to traditional microphone arrays, the spacial field is sampled in a larger area, which leads to a significant performance increase. WASNs, however, also provide new signal processing challenges. This work is concerned with multi-speaker voice activity detection (VAD) for WASNs, which is a prerequisite for many speech enhancement algorithms [3]. Multi-speaker VAD is challenging because hidden voice activity patterns must be recovered for all sources, given observed mixtures only.

Existing VAD approaches, e.g., [4] rely on energy separation by means of multiplicative nonnegative independent component analysis (MNICA). However, they suffer from a significant performance loss, as the number of sources increases, making the MNICA problem more and more difficult. Further, non-active speech may yield a small, but non-zero energy value, making the detection task non-trivial. Finally, existing methods are non-robust against impulsive noise.

We propose a multi-speaker VAD approach that addresses the above three challenges. To provide scalability in terms of the active number of sources, we propose a new method to identify the so-called dominant source model that has been introduced in [5]. Measurements from nodes that observe the same dominant source are highly correlated. The number of dominant sources and their associated node clusters are, therefore, estimated based on correlation information. We formulate the task as model-selection problem in multiple data sets which has been addressed in [6], [7]. The results of [7] are extended because it uses joint information from all nodes and has shown to have superior performance in various scenarios compared to [6]. Thus, by determining node clusters $i = 1, \ldots, d$ that observe the same source as dominant source, the energy separation problem is divided into d simpler energy separation tasks. The second challenge of thresholding voice activity has been recently addressed by incorporating sparseness constraints on the energy signatures [8], [9]. Shrinking small energy values to zero relieves the practitioner from the necessity of defining a heuristic voice activity threshold. However, such approaches are not robust against impulsive noise. Therefore, we introduce a group sparseness constraint that matches the characteristics of human speech and suppresses impulsive noise in the energy unmixing step. The proposed method outperforms existing approaches for a WASN in a 20 $\times$ 10 m room with four simultaneously active human speakers and two impulsive noise sources observed by 15 nodes.

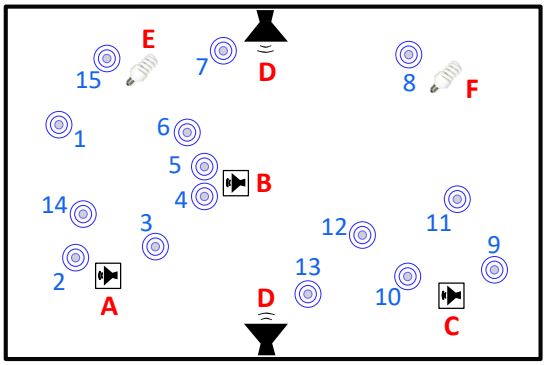

Fig. 1: An example of a WASN with 6 sources (A-F) and 15 sensor nodes in a 20 $\times$ 10 m room.

II. Problem Formulation

A WASN of P sensor nodes, each equipped with n microphones, is considered. Let q acoustic sources be active in the network. We assume there are $d(\leq q)$ dominant sources, i.e., sources that are observed dominantly in more than one sensor node. Fig. 1 shows an example of a WASN with 6 sources (A-F, shown in red) and 15 sensor nodes (1-15, shown in blue). In this case, nodes 2,3 and 14 mainly observe speaker A. We will therefore, call speaker A as the dominant source for nodes 2,3 and 14. Similarly, speaker B is the dominant source for nodes 4,5 and 6, and so on. There are $q - d$ sources that are observed mainly by one sensor node. For example, in Fig. 1,

sources E and F (generating bulb-flickering noise) are mainly observed by sensors 15 and 8, respectively. Our aim is to robustly estimate the voice activity of the unknown number of d dominant sources in WASN, given a received mixture at each sensor.

The proposed method consists of two processing steps: first, d along with the associated node clustering information, $\mathbf{Z}$, is computed. The binary matrix $\mathbf{Z}$ is of size $d \times P$ whose $\{ij\}$th entry is 1 if the ith dominant source is observed by the jth node, and 0 otherwise. Based on the estimates $\hat{d}$ and $\hat{\mathbf{Z}}$, in the second step, the voice activity pattern (VAP) is determined separately for the ith dominant source by only using the nodes marked as 1 in the ith row of $\hat{\mathbf{Z}}$. The two steps are explained in detail in Sections III and IV, respectively.

III. Source Enumeration and Node Clustering

Frequency-domain signal model: Let the observed vector $\mathbf{x}_p(f) \in \mathbb{C}^n$ at node p and frequency index f be modeled as

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

where $\mathbf{A}_p(f) \in \mathbb{C}^{n \times m_p}$ is the full column rank mixing matrix (acoustic transfer function), and $\mathbf{s}_p(f) \in \mathbb{C}^{m_p}$ refers to the source vector. We assume an unknown $m_p (\leq n)$ number of uncorrelated and (locally) stationary sources in $\mathbf{s}_p(f)$ that, without loss of generality, are assumed to be zero-mean and unit variance. The kth signal component of the pth data set is denoted by $s_p^{(k)} = u_p^{(k)} + jv_p^{(k)}$, where $u_p^{(k)}$ and $v_p^{(k)}$ are the real and imaginary parts of $s_p^{(k)}$ [1]. Between any two nodes p and q, sources may be correlated only pairwise, i.e., the source $s_p^{(k)}$ may only correlate with source $s_q^{(k)}$ for $1 \leq k \leq m_{pq} (= \min(m_p, m_q))$ with an unknown (possibly zero) correlation coefficient

$$\rho_{pq}^{(k)} = E[u_p^{(k)}u_q^{(k)}] + E[v_p^{(k)}v_q^{(k)}] + j(E[v_p^{(k)}u_q^{(k)}] - E[u_p^{(k)}v_q^{(k)}]). \tag{2}$$

Let there be d number of dominant sources such that for $i = 1, \ldots, d$, there exist $P^{(i)}$ nodes whose kth sources are correlated with each other.

Let $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T$ be the composite data vector of P nodes and $\mathbf{R} = E[\mathbf{x}\mathbf{x}^H]$. Let $\mathbf{R}_D = \text{blkdiag}(\mathbf{R}_{11}, \ldots, \mathbf{R}_{PP})$ be a block diagonal matrix with $\mathbf{R}_{pp} = E[\mathbf{x}_p\mathbf{x}_p^H]$. The composite coherence matrix $\mathbf{C} \in \mathbb{C}^{nP \times nP}$ is defined as

$$\mathbf{C} = \mathbf{R}_D^{-\frac{1}{2}}\mathbf{R}\mathbf{R}_D^{-\frac{1}{2}}, \tag{3}$$

where the exponent $^{-\frac{1}{2}}$ denotes the square-root matrix inverse (or square-root pseudo-inverse of a rank-deficient matrix). It is shown in [7] that the eigenvalue decomposition of $\mathbf{C}$ can completely characterize the correlation structure among multiple sets of data. In this context, d is equal to the number of correlated sources among all the nodes, and determining the correlation structure is equivalent to finding the cluster of nodes for the d sources. However, the results in [7] are derived for a real-valued $\mathbf{C}$.

[1]From now on, we will drop the frequency index (f) from the sources, observed vectors and their covariance matrices for conciseness.

For this work, we assume that covariances of real and imaginary parts of correlated sources are equal, i.e., $E[v_p^{(k)}u_q^{(k)}] = E[u_p^{(k)}v_q^{(k)}]$. Thus, using (2), $\rho_{pq}^{(k)}$ is real-valued. This assumption is reasonable since the correlated sources in different nodes are generated from a common underlying speaker. This means, the signal cross-covariance matrix between nodes p and q $(p \neq q)$

$$\mathbf{R}_{s_p s_q} = E[\mathbf{s}_p\mathbf{s}_q^H] = \text{diag}(\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(m_{pq})}) \tag{4}$$

is a diagonal and real-valued matrix. Under this assumption, Theorem I and II derived in [7] for a real-valued C can be straightforwardly extended for a complex-valued C. More specifically,

i) $\mathbf{C}$ has exactly d eigenvalues greater than one if and only if there exists d dominant sources in the WASN, and
ii) Let $\mathbf{u}^{(i)}$ be the eigenvector associated with the ith largest eigenvalue of $\mathbf{C}$. $\mathbf{u}^{(i)}$ can be partitioned into P subvectors, $\mathbf{u}^{(i)} = [\mathbf{u}_1^{(i)T}, \mathbf{u}_2^{(i)T}, ..., \mathbf{u}_P^{(i)T}]^T$, where $\mathbf{u}_p^{(i)} \in \mathbb{C}^n$ contains the elements of $\mathbf{u}^{(i)}$ associated with the pth node. The ith source is observed by the pth node if and only if $\mathbf{u}_p^{(i)} \neq \mathbf{0}$.

Bootstrap-based hypothesis testing: In practice, $\mathbf{C}$ is unknown and is estimated from observations. Let M samples of each node form the columns of the sample matrices $\mathbf{X}_1, \ldots, \mathbf{X}_P$. The sample coherence matrix, $\hat{\mathbf{C}}$ can be computed from the sample estimates of covariance matrices, $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}_D$, using (3). In this case, the number of eigenvalues of $\hat{\mathbf{C}}$ that are greater than one will often not equal to d. A sequence of binary hypothesis tests can be used to determine d. Starting with $s = 0$, each test compares the null hypothesis $H_0 : d = s$ with the alternative hypothesis $H_1 : d > s$. If H_0 is rejected, s is incremented and the test is repeated until H_0 is not rejected or s reaches its maximum possible value. We propose a statistic based on the assumption that there is at least one independent source (which can also correspond to noise) among all nodes. This means that at least one eigenvalue of $\mathbf{C}$ is equal to one. The null hypothesis for each test is

$$H_0 : \lambda^{(s+1)} = 1 \tag{5}$$

and the proposed statistic is

$$T(s) = (\lambda^{(s+1)} - 1)^2. \tag{6}$$

We use bootstrap to estimate the unknown distribution of $T(s)$ under H_0 since it has been shown to work well with non-Gaussian data and limited number of observations, both being relevant to this application [10].

Moreover, the subvectors $\mathbf{u}_p^{(i)}$ will not be zero when computed using $\hat{\mathbf{C}}$. Thus, for $i = 1 \ldots d$ and $p = 1 \ldots P$ we test the hypotheses

$$\begin{aligned} H_0 &: \|\mathbf{u}_p^{(i)}\| \leq u_0, \\ H_1 &: \|\mathbf{u}_p^{(i)}\| > u_0, \end{aligned} \tag{7}$$

for the threshold u_0. The distribution of the proposed statistic, $T_p^{(i)} = \|\mathbf{u}_p^{(i)}\|$ under H_0 is estimated using the bootstrap [10]. If H_0 is not rejected, $\mathbf{Z}\{ip\} = 0$, otherwise $\mathbf{Z}\{ip\} = 1$.

The threshold u_0 controls the selection of nodes observing the dominant source. If the ith source is correlated across all nodes with equal pairwise correlation coefficients, $\|\mathbf{u}_p^{(i)}\| = \frac{1}{\sqrt{P}}$, $\forall p = 1, \ldots, P$. However, often in a WASN, a source is not observed by all nodes. Thus, some of the subvectors in $\mathbf{u}^{(i)}$ will be close to zero. Due to the constraint that $\|\mathbf{u}^{(i)}\| = 1$, this will push the subvectors corresponding to the nodes observing the ith dominant source to be significantly higher than $\frac{1}{\sqrt{P}}$. For this reason, we chose $u_0 = \frac{1}{\sqrt{P}}$.

IV. Group Sparse Voice Activity Detection

Because of the on/off behavior of human speech, the individual signal energies are block sparse. Thus, for each of the $\hat{d}$ clusters of nodes obtained, we employ a singular value decomposition (SVD) on the received mixed energies and impose a block-sparsity constraint on the right rotation matrix. We improve upon recent work [8] that assumed only sparse energy sources. In contrast to non-sparse methods as M-NICA [4], sparse median-based M-NICA (SMM-NICA) [8] and also our group-sparse median-based M-NICA (GSMM-NICA) perform VAD intrinsically as all non-zero entries in the reconstructed energy signature are automatically labeled as active speech. The entries of reconstructed energy signatures for non-sparse methods are, in general, all non-zero and an activity threshold τ has to be defined, which heavily depends on the deployed application scenario.

Time-domain energy model: Let the received energy vector $\mathbf{y}_p(k) \in \mathbb{R}^n$ for all n microphones at node p and frame index k be

$$\mathbf{y}_p(k) = \mathbf{A}_p \mathbf{e}(k) + \mathbf{w}_p(k), \tag{8}$$

where $\mathbf{A}_p \in \mathbb{R}^{n \times q}$ is the mixing matrix for node p, $\mathbf{e}(k) \in \mathbb{R}^q$ contains the signal energy from all q speakers in the WASN, i.e., the sum of squared received signal values during time frame k per speaker, and $\mathbf{w}_p(k) \in \mathbb{R}^n$ is additive noise. The individual node observations are summarized in a network received energy vector $\mathbf{y}(k) = \left[\mathbf{y}_1^T(k), \ldots \mathbf{y}_P^T(k)\right]^T$. Finally, the observations at all time instants $k = 1, \ldots, N$ and nodes are expressed as

$$\mathbf{Y} = \mathbf{A}\mathbf{E} + \mathbf{W}, \tag{9}$$

where $\mathbf{Y} \in \mathbb{R}^{nP \times N}$, the WASN mixing matrix $\mathbf{A} \in \mathbb{R}^{nP \times q}$, energy matrix $\mathbf{E} \in \mathbb{R}^{q \times N}$ and noise matrix $\mathbf{W} \in \mathbb{R}^{nP \times N}$.

Since the initial source enumeration and node clustering algorithm divides the P nodes into $\hat{d}$ clusters, we define cluster-wise received energy matrices

$$\mathbf{Y}^{(i)} = \mathbf{a}^{(i)}\mathbf{e}^{(i)} + \mathbf{W}^{(i)}, \tag{10}$$

where $\mathbf{Y}^{(i)} \in \mathbb{R}^{n^{(i)} \times N}$, $\mathbf{a}^{(i)} \in \mathbb{R}^{n^{(i)} \times 1}$, $\mathbf{e}^{(i)} \in \mathbb{R}^{1 \times N}$ and $\mathbf{W}^{(i)} \in \mathbb{R}^{n^{(i)} \times N}$ with the number of nodes $P^{(i)}$ and the number of microphones $n^{(i)} = n \cdot P^{(i)}$ per cluster $i = 1, \ldots, \hat{d}$.

Singular value decomposition: M-NICA separates sources by applying an SVD to the energy matrix $\mathbf{Y}$, i.e., $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. $\mathbf{\Sigma}$ contains the singular values of $\mathbf{Y}$ on its diagonal, $\mathbf{U}$ and $\mathbf{V}$ are composed of the left and right singular vectors of $\mathbf{Y}$. As the principle components of $\mathbf{Y}$ are represented by $\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{\Sigma}$ acts as a scaling factor without influencing the signature shape, $\mathbf{V}$ contains the required information on the time-domain shape of the underlying energies [4]. Due to the previous decomposition of the WASN into clusters of microphones using the dominant source approach, we aim to extract exactly one source, the dominant source, from each of the $\hat{d}$ clusters. Hence, the SVD provides us with one singular value $\mathbf{\Sigma}^{(i)} = \sigma^{(i)}$ and one left/right singular vector $\mathbf{U}^{(i)} = \mathbf{u}^{(i)}$ and $\mathbf{V}^{(i)} = \mathbf{v}^{(i)}$, respectively. Since we only work with cluster-wise quantities in the following, we drop superscript (i) for readability.

To enforce group sparsity in the right singular vector $\mathbf{v}$, we first decompose $\mathbf{Y}$ by a standard SVD to obtain estimates for $\mathbf{u}, \sigma$ and $\mathbf{v}$. Then, we divide the cluster-wise received energy matrix $\mathbf{Y} \in \mathbb{R}^{n^{(i)} \times N}$ into L groups of time samples, each of length N_g and extract the cluster-wise group received energy matrices $\mathbf{Y}_{\mathrm{g},l} \in \mathbb{R}^{n^{(i)} \times N_\mathrm{g}}$ such that $\mathbf{Y} = [\mathbf{Y}_{\mathrm{g},1}, \ldots, \mathbf{Y}_{\mathrm{g},L}]$. Equivalently, we define $\mathbf{v}_{\mathrm{g},l} = [v_{\mathrm{g},l}[1], \ldots, v_{\mathrm{g},l}[N_\mathrm{g}]] \in \mathbb{R}^{N_\mathrm{g}}$. We reformulate the SVD optimization problem w.r.t. the right singular vector $\mathbf{v}$ as

$$\underset{\mathbf{v}}{\operatorname{argmin}} \left|\left| \sum_{l=1}^{L} \mathbf{Y}_{\mathrm{g},l} - \sigma \mathbf{u} \mathbf{v}_{\mathrm{g},l}^T \right|\right| + \lambda_\mathbf{v} \Omega(\mathbf{v}), \tag{11}$$

where $\Omega(\mathbf{v})$ is a sparsity inducing penalty-term and $\lambda_\mathbf{v}$ is a tuning parameter that determines the degree of sparsity of $\mathbf{v}$. In contrast to [8], we work with grouped quantities and define $\Omega(\mathbf{v}) = \sum_{l=1}^{L} \sqrt{\sum_{k}^{N_\mathrm{g}} |v_{\mathrm{g},l}[k]|^2}$ as a mixed $\ell 1/\ell 2$ norm. The optimization problem (11) is a model-adjusted variant of a problem in [11], whose solution is commonly referred to as the group least absolute shrinkage and selection operator (group LASSO). Since $\sigma\mathbf{u}$ is orthonormal as $\mathbf{u}$ is a singular vector, the group LASSO [11] for our data model results in

$$\mathbf{v}_{\mathrm{g},l}^T = \begin{cases} \left(1 - \frac{\lambda_\mathbf{v}\sqrt{N_\mathrm{g}}}{||\mathbf{u}^T\mathbf{Y}_{\mathrm{g},l}||^2}\right) \mathbf{u}^T\mathbf{Y} & x \geq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where we iterate until convergence of $\mathbf{v}$ over all L groups.

The tuning parameter $\lambda_\mathbf{v}$ is selected as $\lambda_\mathbf{v} = \operatorname{argmin} C_\mathrm{N}$, where Mallows' C_N [12] is equivalent to the Akaike information criterion for our time-domain signal energy model, thus,

$$\begin{aligned} C_N = &\frac{\left|\left|\mathbf{Y} - \mathbf{u}\mathbf{v}^T(\lambda_\mathbf{v})\right|\right|^2}{\sigma_\mathbf{Y}^2} - n^{(i)} \\ &+ 2\sum_{l=1}^{L} (||\mathbf{v}_{\mathrm{g},l}|| > 0) + 2\sum_{l=1}^{L} \frac{||\mathbf{v}_{\mathrm{g},l}||}{\left|\left|\mathbf{v}_{\mathrm{g},l}^{\mathrm{LS}}\right|\right|}(N_\mathrm{g} - 1), \end{aligned} \tag{13}$$

where $\mathbf{v}^{\mathrm{LS}} = \left[\mathbf{v}_{\mathrm{g},1}^{\mathrm{LS}\,T}, \ldots, \mathbf{v}_{\mathrm{g},L}^{\mathrm{LS}\,T}\right]^T = ((\mathbf{u}^T\mathbf{u})\mathbf{u}^T\mathbf{Y})^T$ and $\sigma_\mathbf{Y}^2$ is the variance of $\mathbf{Y}$.

Proposed Algorithm: Let us briefly recapitulate our proposed GSMM-NICA. We decompose the cluster-wise received energies initially by a standard SVD. Then, we iteratively compute (12) for all L groups, which provides us with a group sparse energy signature shape estimate $\mathbf{v}$. $\lambda_\mathbf{v}^* = \operatorname{argmin}_{\lambda_\mathbf{v}} C_N$ is selected in each iteration individually. We stop iterating when the update on solution $\mathbf{v}$ falls below a convergence threshold. We then continue with de-correlation of $\mathbf{Y}$ using σ, $\mathbf{u}$ and (group-sparse) $\mathbf{v}$ as SMM-NICA [8].

V. Results

We validate the proposed technique on a WASN generated from real speakers in a 20×10m room with a reverberation time of 0.3s [2]. Each node is equipped with a uniform linear array of $n = 3$ microphones sampled at 16kHz. We set the duration of one energy time frame to 30 ms, which fits well to speech characteristics. Each microphone is corrupted by an additive white Gaussian noise with variance of 0.05. For source enumeration and node clustering, short-term Fourier transform (STFT) with a Hamming window and frame length of 1024 with 50% overlap is applied to the data obtained from each microphone. The number of frequency bins for the STFT is chosen as 32. The proposed technique provides an estimate of the number of dominant sources and the corresponding clusters for each frequency bin. The estimated number of sources are averaged over all frequency bins and rounded off to the nearest integer to obtain the final estimate of d. The top d majority voted clusters from all frequency bins are selected as the final clusters associated with the d dominant sources. We provide a brief summary of the selected competitors from the literature.

Competitors: We compare our proposed method with two algorithms from the literature, namely, the standard M-NICA [4] algorithm and SMM-NICA [8]. The later one is deployed on a cluster-wise level, meaning that the presented results for SMM-NICA were obtained using the dominant source model presented in Section III, to demonstrate the effect of the group sparsity constraint in GSMM-NICA. Also, for M-NICA, one has to select a voice activity threshold as all entries of the resulting energy signature are generally non-zero. To compare our method to the best possible results for M-NICA, we decided to perform VAD for a grid of possible energy threshold τ and select the one τ_{opt} that provides us with the largest number of correctly classified time samples. In a real-world scenario, the underlying ground truth VAP is unknown and determining τ_{opt} is impossible. Therefore, we refer to this competitor as oracle M-NICA.

The results for two different scenarios are presented.

Scenario 1: A total of 10 nodes observe three spatially well-separated speakers for a duration of 15s as shown in Fig. 2. The clustering result is shown in Table I for $\hat{d} = 3$ estimated speakers. The nodes 1, 7 and 8, which are comparatively far away from all speakers are not selected in any of the clusters. Based on the clustering results in Table I, we run the group sparse VAD as presented in Section IV. As an exemplary result, we plot the obtained energy signatures and VAPs for speaker B in Fig. 3. In general, we observed in simulations that a group length N_{g} between 8 and 12 samples at 16 kHz seems to best fit the human speech characteristics. Thus, we choose $N_{\text{g}} = 10$ samples.

The proposed clustered group sparse method clearly outperforms the competitors. The results for speakers A and C are similar to the ones displayed in the figure.

[2]WASN data and MATLAB code for our and the competing techniques are available at https://github.com/SSTGroup/Voice-Activity-Detection-in-WASN/

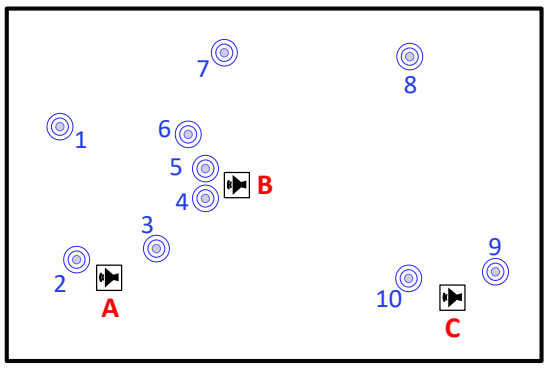

Fig. 2: An example of WASN with 3 sources (A-C) and 10 sensor nodes in a 20×10 m room.

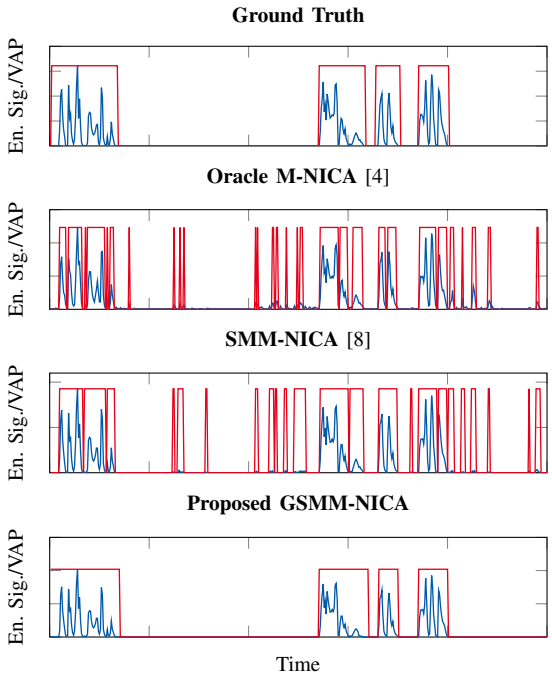

Fig. 3: Extracted energy signatures (blue) and voice activity patterns (red) for speaker B in Scenario 1. GSMM-NICA outperforms the competitors and the extracted VAP is closest to the ground truth VAP.

Scenario 2: In a more challenging scenario, a public announcement loudspeaker is introduced at two opposite sides of the room denoted as speaker D in Fig. 1 along with two sources E and F which generate uncorrelated bulb-flickering impulsive noise. The six sources are observed by 15 nodes for 30s duration. The clustering result is listed in Table I. The proposed technique estimates $\hat{d} = 4$ speakers correctly along with their node clusters. For loudspeaker D, nodes 7, 12 and 13 form one cluster even though they are are spatially far from each other. Nodes 1, 8 and 15 which either observe impulsive noise or are far from all speakers do not belong to any cluster.

Again, we run the group sparse VAD on the cluster observation matrices. In this case, we suggest using a larger group length, which suppresses the flickering noise most effectively. The energy signatures and VAPs for speaker C in Fig. 4 were obtained for $N_{\text{g}} = 15$ samples. The proposed method again clearly outperforms the competitors, which are not able to treat

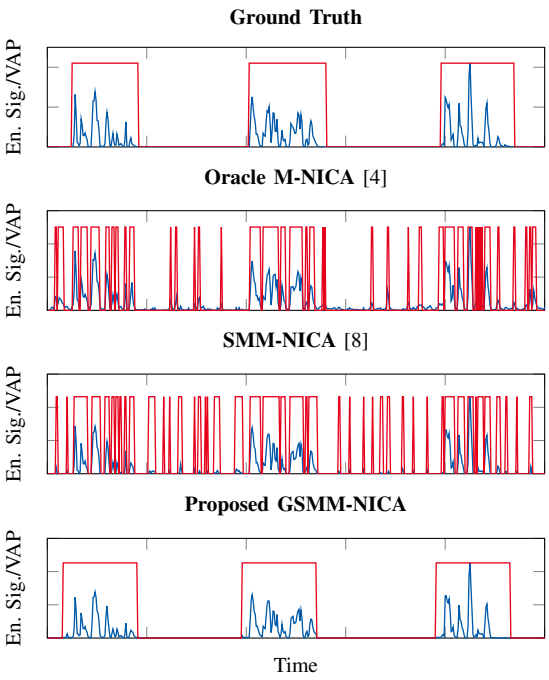

Fig. 4: Extracted energy signatures (blue) and voice activity patterns (red) for speaker C in Scenario 2. GSMM-NICA outperforms the competitors and the extracted VAP is closest to the ground truth VAP.

the noise as efficiently. The results for speakers A, B and D are similar to the ones displayed in the figure.

In Table II, we provide the percentage of correctly labeled (speech/pause) frames for the two scenarios and all sources. In scenario 1, the proposed technique with group length $N_{\mathrm{g}} = 10$ samples outperforms the competitors for speakers B and C and shows similar performance for speaker A. For the second scenario, we decided for $N_{\mathrm{g}} = 15$ samples to completely suppress the flickering noise. However, the resulting VAPs are a bit longer for speakers which are not strongly disturbed by the flickering. Therefore, we observe a slight degradation in the performance for speakers A and B. A more sophisticated approach to determine the ideal group length for each speaker would allow for a general improvement over M-NICA and SMM-NICA.

Dominant Source	Cluster of Nodes (Scenario 1)	Cluster of Nodes (Scenario 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

TABLE I: The clustering result of the proposed technique for scenarios 1 and 2.

VI. Conclusion

We have devised a new method to perform multi-speaker voice activity detection and source enumeration in WASNs. Due to the clustering of nodes according to the dominant source model, the approach outperforms existing standard procedures that process all nodes in the network jointly. The sparseness constraint relieves the practitioner from the necessity of defining a heuristic voice activity threshold. The group sparseness constraint suits better to the characteristics of human speech than a simple sparseness constraint.

	Dominant Source	Oracle M-NICA [4]	SMM-NICA [8]	Proposed GSMM-NICA
Scenario 1	A	81.4	86.8	84.2
	B	83.2	85.0	97.6
	C	84.2	86.0	89.8
Scenario 2	A	61.7	67.4	62.7
	B	83.3	81.6	81.5
	C	79.1	77.7	93.6
	D	75.1	71.8	78.7

TABLE II: The percentage of correctly labeled frames for all sources.

Acknowledgment

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). We thank L. K. Hamaidi for providing her codes that implement [8].

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*. IEEE, 2011, pp. 1–6.

[2] M. Cobos, F. Antonacci, A. Mouchtaris, and B. Lee, "Wireless acoustic sensor networks and applications," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[3] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, 2015.

[4] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 85–88.

[5] M. H. Bahari, L. K. Hamaidi, M. Muma, J. Plata-Chaves, M. Moonen, A. M. Zoubir, and A. Bertrand, "Distributed multi-speaker voice activity detection for wireless acoustic sensor networks," *arXiv preprint arXiv:1703.05782*, 2017.

[6] T. Marrinan, T. Hasija, C. Lameiro, and P. J. Schreier, "Complete model selection in multiset canonical correlation analysis," in *Proceedings of the European Signal Processing Conference*, 2018.

[7] T. Hasija, C. Lameiro, T. Marrinan, and P. J. Schreier, "Determining the dimension and structure of the subspace correlated across multiple data sets," *arXiv preprint arXiv:1901.11366*, 2019.

[8] L. K. Hamaidi, M. Muma, and A. M. Zoubir, "Multi-speaker voice activity detection by an improved multiplicative non-negative independent component analysis with sparseness constraints," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4611–4615.

[9] ——, "Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 161–165.

[10] A. M. Zoubir and D. R. Iskander, *Bootstrap techniques for signal processing*. Cambridge University Press, 2004.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc., Series B (Stat. Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x

[12] C. L. Mallows, "Some comments on c_p," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973. [Online]. Available: http://www.jstor.org/stable/1267380