Robust Distributed Multi-Speaker Voice Activity Detection Using Stability Selection for Sparse Non-Negative Feature Extraction

L. Khadidja Hamaidi*†, Michael Muma* and Abdelhak M. Zoubir*†

* Signal Processing Group
Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt, Germany
{hamaidi, muma, zoubir}@spg.tu-darmstadt.de

† Graduate School CE
Technische Universität Darmstadt
Dolivostraße 15, 64293 Darmstadt, Germany
{hamaidi, zoubir}@gsc.tu-darmstadt.de

***Abstract*—In this paper, we propose a robust multi-speaker voice activity detection approach for wireless acoustic sensor networks (WASN). Each node of the WASN receives a mixture of sound sources. We propose a non-negative feature extraction using stability selection that exploits the sparsity of the speech energy signals. The strongest right singular vectors serve as source-specific features for the subsequent voice activity detection (VAD). To separate active speech frames from silent frames, we propose a robust Mahalanobis classifier that is based on an M-estimator of the covariance matrix. The proposed approach can also be applied to a distributed setting, where no fusion center is available. Highly accurate VAD results are obtained in a challenging WASN of 20 nodes observing 6 sources in a reverberant environment.**

I. Introduction

Voice activity detection (VAD), i.e., detecting the presence or absence of human speech, is crucial for several speech processing applications, such as noisy speech enhancement [1], speaker recognition [2], speech coding systems [3], echo cancellation and hands-free telephony [4]. VAD algorithms trade off noise sensitivity, precision, and computational complexity and consist of two consecutive phases, namely, speech-related feature extraction and a discriminating model. Classical single-speaker VAD methods use energy features [5], signal periodicity [6], and zero-crossing rates [7]. While other more sophisticated approaches rely on statistical model-based speech classification [8], [9]. Compared to single-speaker VAD, very few solutions exist for the multiple-concurrent-speaker case. For centralized WASN, [10], [11] propose a multi-speaker energy pattern extraction by designing an efficient energy unmixing algorithm. After energy separation, no VAD is performed in [10], [11]. In [12], independent component analysis (ICA) is used combined with beampattern analysis to identify the active speaker and perform VAD based on the precise knowledge of the direction of arrival of the speech signals. An integrated multi-source speaker localization and multi-channel VAD framework is introduced in [13]. The paper exploits the behavior of the spatial gradient steered response power function using the phase transform method. While in [14], identifying a single target speaker from multiple speakers is considered. Thus, an energy-based information from the interfering channels is included to adaptively adjust the decision threshold of the targeted channel. Recently, a VAD method [15] is developed that exploits processed information recorded from a camera-assisted microphone array. Moreover, a centralized sparse median-based multiplicative non-negative ICA (M-NICA), abbreviated by SMM-NICA, is proposed for energy source unmixing in [16]. The idea is to enhance the energy features with a penalized ℓ_1-norm model and apply a straightforward zero-threshold VAD, which detects speech activity. Concerning, distributed WASN, the literature is even scarcer. [17] proposes a distributed multi-speaker VAD (DMVAD) algorithm that first unmixes and then detects the activity of multiple interfering energy signals in a WASN.

Contributions: We improve upon [16] and [17] with a two-step robust solution to the multi-speaker VAD problem by exploiting sparse coding [18], [19]. The novelty of our approach lies in first using a sub-sampling stability approach that selects the degree of sparseness parameter in the penalized regression suitable for a time-domain sparse energy feature extraction. Additionally, a subsequent robust classification step that uses robust Mahalanobis distance based on M-estimation is performed. Hence, our suggested method addresses the multiple speech activity detection task and makes it unnecessary to use an energy unmixing method, as proposed in the SMM-NICA method in [16]. Both centralized and decentralized multi-speaker VAD is considered, and in both cases highly accurate results are obtained.

II. Signal model and Problem Formulation

We analyze an ad-hoc distributed wireless acoustic sensor network (WASN) accommodating N speakers and $k = 1, \ldots, K$ devices. Each device k comprises a uniform linear array (ULA) equipped with an identical number of microphone sensing elements J_k. The overall number of microphones throughout the network is $J = \sum_{k=1}^{K} J_k$. Fig. 1 sketches the studied audio scenario. A speaker n generates signals $\tilde{s}_n[\eta]$, $\eta = 1, \ldots T$, where η denotes the sample time index. The matrix $[\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_N]^\top \in \mathbb{R}^{N \times T}$ consists of speech source

The work of L. K. Hamaidi is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt (e-mail: hamaidi@gsc.tu-darmstadt.de).

ISBN 978-0-9928626-7-1 © EURASIP 2017

column vectors $\tilde{\mathbf{s}}_n \in \mathbb{R}^T$ that are mutually independent and uniquely labeled using the algorithm presented in [20]. We assume statistical second-order stationarity for blocks of length L and define the instantaneous power of a signal $\tilde{s}_n[\eta]$ at each block as

$$s_n[i] = \frac{1}{L}\sum_{l=0}^{L-1} \tilde{s}_n[iL+l]^2, \tag{1}$$

where $i = 1, \dots, I$ is the frame index. The $s_n[i]$ are stacked in an N dimensional vector $\mathbf{s}[i]$. The instantaneous noisy signal power at the jth microphone of the kth device is

$$y_{k,j}[i] = \frac{1}{L}\sum_{l=0}^{L-1} \tilde{y}_{k,j}[iL+l]^2, \quad j \in \{1, \dots, J_k\}, \tag{2}$$

where $\tilde{y}_{k,j}$ denotes the observed signal at the jth microphone of the kth device. Assuming a centralized network, the system-wide non-negative $y_{k,j}[i]$ of all devices k are stacked in a J-dimensional vector $\mathbf{y}[i]$. The mixture is modeled by

$$\mathbf{y}[i] \approx \mathbf{A}\mathbf{s}[i] + \boldsymbol{\omega}[i], \quad i = 1, \dots, I, \tag{3}$$

with $\mathbf{A} \in \mathbb{R}^{J\times N}$ being the mixing matrix that describes the power attenuation between speaker n and microphone j. The additive white noise term $\boldsymbol{\omega}[i]$ follows the same design introduced in Eqs. (1)-(2). In the centralized setup, as in [10], the instantaneous linear mixtures in Eq. (3) allow for the estimation of $\mathbf{s}[i]$. Our focus is to extract well-separated sparse features similar to $\mathbf{s}[i]$, which are the input of a subsequent classification-based multi-speaker VAD.

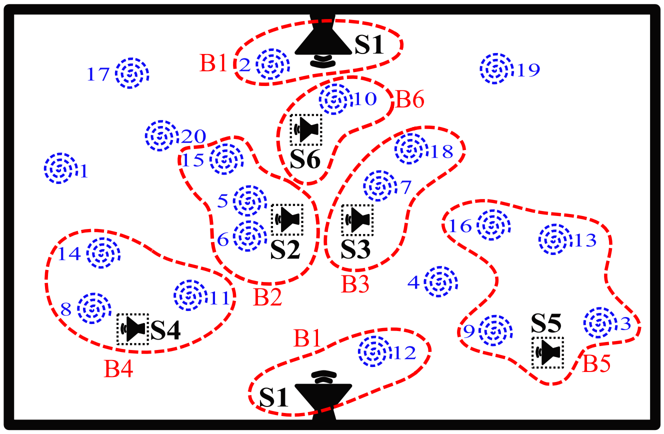


Fig. 1: A 20 × 10 meter room with a reverberation time of T60=0.3 seconds. The room describes a wireless acoustic sensor network (WASN) of $N = 6$ speech sources (black speakers) and $K = 20$ devices (blue dashed circles). Each device k is equipped with $J_k = 3$ microphones sampled at 16 kHz.

III. Robust and Sparse Energy Feature Extraction Based Stability Selection

Let $\mathbf{Y} \in \mathbb{R}_+^{J\times I}$ denote the matrix composed of entries $\mathbf{y}[i]$, $i = 1, \dots, I$. Singular value decomposition (SVD) projects $\mathbf{Y}$ onto

$$\text{SVD}(\mathbf{Y}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{J\times J}$ and $\mathbf{V}^\top \in \mathbb{R}^{I\times I}$ describe the left and right orthogonal rotations of singular vectors, respectively. $\boldsymbol{\Sigma} \in \mathbb{R}^{J\times I}$ contains the singular values on its diagonal. In essence, we target a robust derivation of sparse right-singular vectors. Thus, we suggest as in [16] to impose sparsity-inducing penalties solely on $\mathbf{V}$ within the iterative rank-one SVD layer extraction. Sparse right rotation components serve as features for the subsequent VAD phase. Accordingly, we consider an ℓ_1-regularized term that minimizes a penalized sum-of-squares criterion, such that

$$\underset{\sigma,\mathbf{u},\mathbf{v}}{\text{argmin}}\|\mathbf{Y} - \sigma\mathbf{u}\mathbf{v}^\top\|^2 + \lambda_{\mathbf{v}}\Phi(\sigma\mathbf{v}), \tag{5}$$

with $\mathbf{u}$ and $\mathbf{v}$ being unit vectors of length J and I, respectively. We interpret the right singular vectors $\mathbf{v}$ as regression coefficients of a linear penalized regression fit as to design their sparse map. $\lambda_{\mathbf{v}}$ describes the tuning parameter of the penalization and $\Phi(\sigma\mathbf{v})$ is the ℓ_1 regularization function

$$\Phi(\sigma\mathbf{v}) = \sigma\sum_{i=1}^{I}|v_i|. \tag{6}$$

Based on the Lasso penalized regression in Eq. (5), the selection of $\lambda_{\mathbf{v}}$ corresponds to selecting the degree of sparsity of $\mathbf{v}$, i.e., the number of non-zero components in $\mathbf{v}$. In [16], we use the BIC based penalty parameter selection proposed in [21]. However, the resulting sparse vectors $\mathbf{v}$ require a subsequent unmixing step, see [16]. In this work, we favor the use of stability selection [22], [23] to accurately deduce the sparseness level of the right singular vectors $\mathbf{v}$ and thus determine the minimal penalization value of the regularization parameter $\lambda_{\mathbf{v}}$. This approach is promising as it surmounts the imperative use of an unmixing procedure, such as M-NICA. Let $\mathcal{L}_{\mathbf{v}}$ be the set of possible $\lambda_{\mathbf{v}}$ parameters that we adapt to Eq. (5). Every $\lambda_{\mathbf{v}} \in \mathcal{L}_{\mathbf{v}}$ points to a distinct subspace of non-zero indicators $i \in I$ of $\mathbf{v}$ denoted $\hat{\mathcal{Z}}_{\mathbf{v}}^{\lambda_{\mathbf{v}}}(J)$. The probability of selecting a non-zero coefficient, i.e., $P(\cdot)$ in Eq. (7), is obtained via estimating the relative selection frequencies of i pertaining to all possible subsamples $J^\circ \subset J$ for an arbitrary threshold τ and a given value $\lambda_{\mathbf{v}}$. The subsamples J° are drawn from J without replacement.

$$\hat{\mathcal{Z}}_{\mathbf{v}} = \left\{ i : \max_{\lambda_{\mathbf{v}}\in\mathcal{L}_{\mathbf{v}}} P(i \in \hat{\mathcal{Z}}_{\mathbf{v}}^{\lambda_{\mathbf{v}}}(J^\circ)) \geq \tau \right\}. \tag{7}$$

Here, $\hat{\mathcal{Z}}_{\mathbf{v}}$ encloses the stable selection set of non-attenuated candidates from I. The value of τ is chosen in the range of $[0.6, 0.9]$ according to [22]. The minimal penalization value $\lambda_{\mathbf{v}}^{\text{min}}$ that verifies a maximum estimated probability $P(\cdot)$ in Eq. (7) is used to adjust the components of $\mathbf{v}_i$. A component-wise minimizer derived in [21] that incorporates the minimal regularization parameter $\lambda_{\mathbf{v}}^{\text{min}}$ is utilized to estimate the elements of $\mathbf{v}_i$, such that

$$\mathbf{v}_i = \frac{1}{\sigma}\left[\text{sgn}\left\{[\mathbf{Y}^\top\mathbf{u}]_i\right\}\left(|[\mathbf{Y}^\top\mathbf{u}]_i| - \frac{\lambda_{\mathbf{v}}^{\text{min}}}{2}\right)\right]. \tag{8}$$

IV. Robust Mahalanobis-Based VAD

A. *K-medians Based Speech/Silence Prior Partitioning*

Our focus is to first estimate a pair of centroids $\mathbf{c}_q, q = \{1, 2\}$ associated to two separate classes, namely the active and non-active speech data points $\mathbf{C}_q, q = \{1, 2\}$, respectively.

Algorithm 1 Centralized stability selection based sparse feature extraction and robust Mahalanobis classifier for VAD (SRM-VAD)

```
    Input: Form $\mathbf{Y} = (\mathbf{y}[1], \cdots, \mathbf{y}[I]) \in \mathbb{R}_{+}^{J \times I}$ using Eq. (3).
    VAD procedure
 1: for $n = 1, \ldots, N$ do
 2:     Minimize Eq. (5) subject to the $\ell_1$-norm constraints imposed
        on the right-singular vector $\mathbf{v}$.
 3:     Deduce $\lambda_{\mathbf{v}}^{\text{min}}$ through a stability approach that selects the best
        set of non-zero indicators $i$ guaranteeing sparsity in $\mathbf{v}$, based
        on Eq. (7).
 4:     Adjust $\mathbf{v}$ with its new elements using Eq. (8).
 5:     Update the singular value $\sigma = \mathbf{u}^{\top}\mathbf{Y}\mathbf{v}$.
 6:     Construct a sparse lower-rank matrix $\mathbf{Y}^{\star} = \sigma\mathbf{u}\mathbf{v}^{\top}$.
 7:     Collect the matrix of residue $\mathbf{Y} = \mathbf{Y} - \mathbf{Y}^{\star}$.
 8:     Based on $|\mathbf{v}|$, extract $\mathbf{f}_n^i = [f_{n,1}^i, f_{n,2}^i, f_{n,3}^i]^{\top}, \forall i \in I$,
        with $|\cdot|$ being the absolute value operator.
 9:     Initial speech/silence segregation $\mathbf{C}_q$ based on $\mathbf{c}_q^{\top}$,
        $q = \{1, 2\}$.
10:     Compute $\hat{\mathbf{R}}_{n,q}, \forall q$ using the $p$-variate $t_{\nu}$ M-estimator from
        Eq. (9).
11:     Evaluate robust Mahalanobis distance given in Eq. (10).
12:     Decide upon speech activity for source $n$ using Eq. (11).
13: end for
    Output: VAD patterns $\mathbf{d}_1^{\top}, \cdots, \mathbf{d}_N^{\top}$
```

For this, we collect three statistical short-term feature series $\mathbf{f}_n^i = [f_{n,1}^i, f_{n,2}^i, f_{n,3}^i]^{\top}$ analogous to [17] that well characterize the sparse vector $\mathbf{v}$ related to a given source n. These features capture information about the energy average, the standard deviation, and the energy difference. In this study, we use the K-medians partitional clustering technique as a robust variation of the K-means to determine conforming estimates of the active and non-active centroids, namely $\mathbf{c}_q, q = \{1, 2\}$, respectively, while utilizing the features $\mathbf{f}_n^i$. A centroid $\mathbf{c}_q^{\top}$ is defined as a 3-dimensional vector accommodating the individual centroids relating to the energy average feature, the standard deviation, and the energy difference features at the speech/non-speech clusters. Subsequently, we form two disjoint classes $\mathbf{C}_q, q = \{1, 2\}$, of speech/silence by assigning the realizations of $\mathbf{f}_n^i$ to the closest class $\mathbf{C}_q$ depending on their corresponding distances to the estimated centroids $\mathbf{c}_q$.

B. Robust Mahalanobis-Based Speech Detection

In this subsection, we design a Mahalanobis-based similarity measure using the robust p-variate t_{ν} M-estimator of ν degrees of freedom, see [24], for the estimation of the covariance matrix $\hat{\mathbf{R}}_{n,q}, q = \{1, 2\}$, of the speech/non-speech feature's distributions, respectively. The latter can be formulated as

$$\hat{\mathbf{R}}_{n,q} = \frac{1}{\#(\mathbf{C}_q)} \sum_{i=1}^{\#(\mathbf{C}_q)} u_{\nu}(\mathbf{C}_{q,i}^{\top} \hat{\mathbf{R}}_{n,q}^{-1} \mathbf{C}_{q,i}) \mathbf{C}_{q,i} \mathbf{C}_{q,i}^{\top}, \quad (9)$$

with $u_{\nu}(t) = \frac{p+\nu}{\nu+t}$ being the weight function, p the dimension of $\mathbf{f}_n^i$, $t = \mathbf{C}_{q,i}^{\top} \hat{\mathbf{R}}_{n,q}^{-1} \mathbf{C}_{q,i}$, and $\hat{\mathbf{R}}_{n,q}^{-1}$ corresponding to the inverse covariance matrix. The symbol $\#(\cdot)$ represents the cardinality operator. The robust Mahalanobis distance for the speech/silence classes then becomes

$$M_q(\mathbf{f}_n^i) = \sqrt{(\mathbf{f}_n^i - \hat{\mathbf{c}}_q)^{\top} \hat{\mathbf{R}}_{n,q}^{-1} (\mathbf{f}_n^i - \hat{\mathbf{c}}_q)}, \quad (10)$$

Next, speech activity is determined following the decision rule

$$d_n^i = \begin{cases} 1 & \text{if } M_1(\mathbf{f}_n^i) < M_2(\mathbf{f}_n^i) \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The proposed multi-speaker stability selection based sparseness combined with the robust Mahalanobis classifier for VAD (SRM-VAD) is summarized in Alg. 1.

V. Distributed stability based sparseness and robust Mahalanobis classifier for VAD

Assuming a distributed network of devices, our aim is to obtain speaker-specific VAD patterns using clusters of devices that share a common interest in the described multi-source scheme in Fig. 1. A preliminary divide-and-conquer-based approach is performed. To do this, we apply the LONAS algorithm, see [17], which partitions the network into N clusters by grouping devices around a unique dominant source based on adaptive distributed eigenvalue decomposition. Figure 1 illustrates the resulting device clusters (dashed red), each observing a specific source of interest n. We define $\mathcal{B}_n$ as the set of devices k that observe speaker n as a dominant source. Based on this distributed device structure, we construct the $(J_k\#(\mathcal{B}_n))$-dimensional vector $\mathbf{y}_{\mathcal{B}_n}[i]$ by stacking the non-negative $y_{k,j}[i]$ for every device k present in $\mathcal{B}_n$. $\#(\mathcal{B}_n)$ is the device cardinality for a given source n. Based on Eq. (3), the dominant source model becomes

$$\mathbf{y}_{\mathcal{B}_n}[i] \approx \mathbf{a}_{\mathcal{B}_n} \mathbf{s}[i] + \boldsymbol{\omega}_{\mathcal{B}_n}[i], \quad i = 1, \ldots, I. \quad (12)$$

Here, $\mathbf{a}_{\mathcal{B}_n}$, $\boldsymbol{\omega}_{\mathcal{B}_n}[i] \in \mathbb{R}^{J_k\#(\mathcal{B}_n) \times 1}$ reduce to the mixing vector and noise for the ensemble of devices in $\mathcal{B}_n$. In such a distributed setup, our goal is to provide a sparse estimate $\hat{\mathbf{v}}_{\mathcal{B}_n}[i]$ by observing only the linear mixture $\mathbf{y}_{\mathcal{B}_n}[i]$. The vectors $\hat{\mathbf{v}}_{\mathcal{B}_n}[i], \forall n \in N$, are features used to decide upon speaker-specific activity as outlined in Alg. 2.

Algorithm 2 Distributed stability selection based sparseness and robust Mahalanobis classifier for VAD (DSRM-VAD)

```
1: for $n = 1, \ldots, N$ do
2:     $\mathbf{Y}_{\mathcal{B}_n} = (\mathbf{y}_{\mathcal{B}_n}[1], \ldots, \mathbf{y}_{\mathcal{B}_n}[I]) \in \mathbb{R}_{+}^{(J_k\#(\mathcal{B}_n)) \times I}$ using
       Eq. (12).
3:     Perform Alg. 1. (2) until Alg. 1. (7) to extract a unique sparse
       layer of $\mathbf{Y}_{\mathcal{B}_n}$ and deduce a speaker-specific $\mathbf{v}$ for source $n$.
4:     Apply Alg. 1. (8) until Alg. 1. (12) based on the vector $|\mathbf{v}|$
       related to source $n$.
5:     Extract the speaker-specific VAD pattern $\mathbf{d}_n$ for the current
       observations in $\mathbf{Y}_{\mathcal{B}_n}$.
6: end for
```

VI. Results for VAD

1) Centralized Two-Source Scenario Use-Case: We assess the outcome of our proposed VAD approach on the basis of a centralized multi-speaker WASN presented in Fig. 1 with two simultaneously active speech sources S1 and S6 and an additive white Gaussian noise (AWGN) of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$. In this case, the speech mixture is recorded at every device as shown in Eq. (3). We apply the centralized SRM-VAD method summarized in Alg. 1 on the collected noisy speech mixture $\mathbf{Y}$. The degree of freedom for the robust Mahalanobis is empirically chosen as $\nu = 49$. Figure 2 shows the impact of choosing

ν on the correct detection (CD), misdetection (MD), and false alarm (FA) rates. From Tab. I, we see that the proposed SRM-VAD noticeably outperforms M-NICA in speech activity decision. More than 95% of CD is achieved as displayed in Tab. I. Additionally, we deliver the generated decisions when the proposed standalone sparseness based stability selection for VAD (S-VAD) and its improved version with Mahalanobis distance (SM-VAD) are considered. Comparable results are drawn from both SM-VAD and the fully robust version SRM-VAD. Both algorithms outperform S-VAD for S1. Meanwhile, marginally decreased performance is obtained for S6. This is explained by the concern of stability selection in reducing Type I error rates, while the proposed improved versions SM-VAD and SRM-VAD are biased towards misdetection reduction. Our justification is clearly supported by the measures given in columns MD and FA of Tab. I. Moreover, we assess the separation quality reached by M-NICA and the introduced sparse stability-selection-based methods for VAD. For this, we measure the signal-to-distortion ratios (SDR) as exhibits Tab. I. Distinctly superior separation quality is reached in the energy signatures used for our proposed VAD approaches. We also achieve less distorted signals compared to the SMM-NICA.

Centralized Use-Case						
Variance	**Source**	**Method**	CD (%)	MD (%)	FA (%)	SDR
$\sigma_\omega^2 = 0.01$	S1	M-NICA [10]	62.4	1.9	35.7	-3.23
		SMM-NICA [16]	87.2	5.8	7	7.63
		S-VAD	80.7	6.5	12.8	6.9
		SM-VAD	85.44	1.92	12.64	6.9
		SRM-VAD	85.03	1.52	13.45	6.9
	S6	M-NICA [10]	54.7	1.1	44.2	5.75
		SMM-NICA [16]	80.7	0.9	18.4	5.4
		S-VAD	96.1	2	1.9	5.91
		SM-VAD	95.3	1.3	3.4	5.91
		SRM-VAD	95.45	0.4	4.15	5.91

TABLE I: Comparative results of the original M-NICA [10], SMM-NICA [16], and the proposed S-VAD, SM-VAD, and the SRM-VAD (with $\nu = 49$), in a centralized scenario of two sources (S1 and S6) with AWGN of variance $\sigma_\omega^2 = 0.01$.

2) Distributed Multi-Source Scenario Use-Case: As a second experiment, we consider a WASN observing six speech sources, see Fig. 1, affected with AWGN of variance $\sigma_\omega^2 = 0.01$ variance. We deal with grouped devices following their unique dominant source [17]. Devices hearing a source with higher power are more likely to cluster together in order to cooperate for an accurate VAD. Eq. (12) accumulates mixtures from clustered devices per primary dominant source. For the scenario sketched in Fig. 1, we apply Alg. 2. The input is a sub-matrix $\mathbf{Y}_{\mathcal{B}_n}$ assembled from the $\#(\mathcal{B}_n)$ devices for source n. Table II outlines the higher decision results for the proposed distributed VAD algorithms compared to M-NICA and DMVAD. Figure 3 depicts the estimated VAD patterns with high precision layered on the energy ground truth in the distributed scenario for three different speech sources S3, S4, and S5.

Distributed Use-Case						
Variance	**Source**	**Method**	CD (%)	MD (%)	FA (%)	SDR
$\sigma_\omega^2 = 0.01$	S1	M-NICA [10]	60.8	6	33.2	-55.73
		DMVAD [17]	86.3	3.5	10.1	7.7
		S-VAD	79.6	10.4	10	7.4
		SM-VAD	85.44	5.7	8.9	7.4
		DSRM-VAD	85.04	2.33	12.63	7.4
	S2	M-NICA [10]	46.85	3	50.15	-9.5
		DMVAD [17]	96.3	0.8	2.9	6.73
		S-VAD	93.1	3	3.9	6.7
		SM-VAD	96	0.2	3.8	6.7
		DSRM-VAD	95.1	0	4.9	6.7
	S3	M-NICA [10]	56.96	3.90	39.14	-34.6
		DMVAD [17]	97	0.9	2.1	7
		S-VAD	89.4	10.5	0.1	6.6
		SM-VAD	96.6	0.3	3.1	6.6
		DSRM-VAD	96.2	0.3	3.5	6.6
	S4	M-NICA [10]	55.85	6.41	37.74	-14.52
		DMVAD [17]	93.6	6.4	0	8.6
		S-VAD	77.6	22.4	0	8.2
		SM-VAD	95.96	3.03	1.01	8.2
		DSRM-VAD	96.4	2.4	1.2	8.2
	S5	M-NICA [10]	45.75	5	49.25	-34.52
		DMVAD [17]	96.2	3.8	0	2.3
		S-VAD	94.5	4	1.5	2.3
		SM-VAD	98.2	1.7	0.1	2.3
		DSRM-VAD	98.9	0.8	0.3	2.3
	S6	M-NICA [10]	46.55	2.8	50.65	-20.3
		DMVAD [17]	94.8	2.2	2.9	5.9
		S-VAD	91.4	8.6	0	5.3
		SM-VAD	94.85	2.12	3.03	5.3
		DSRM-VAD	95.7	0.6	3.7	5.3

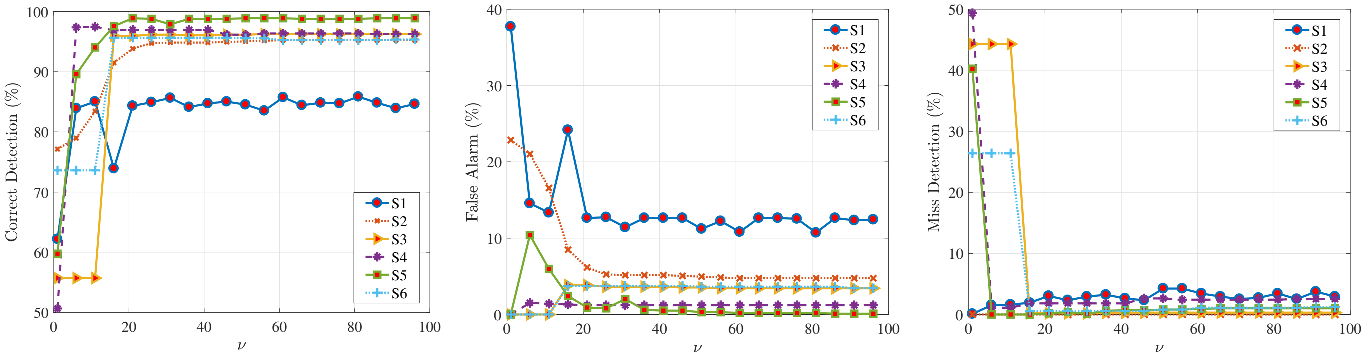
TABLE II: Detection comparision of the original M-NICA algorithm [10], the DMVAD approach [17], and the proposed methods: the S-VAD, the SM-VAD and the DSRM-VAD (with a degree of freedom robustness parameter $\nu = 49$), for the speech use-case scenario presented in Fig. 1, with AWGN of variance $\sigma_\omega^2 = 0.01$.

VII. Conclusion

We suggest a new method for solving the multi-speaker VAD problem for WASN in a distributed reverberant environment. Our proposed method relies on a stability selection assisted technique to promote a robust and sparse speaker-specific feature extraction from a noisy observed signal mixture. The extracted sparse components are sufficiently well-separated for VAD, so the use of M-NICA is no longer required. A robust Mahalanobis classifier is then designed to reveal speaker-specific activity patterns.

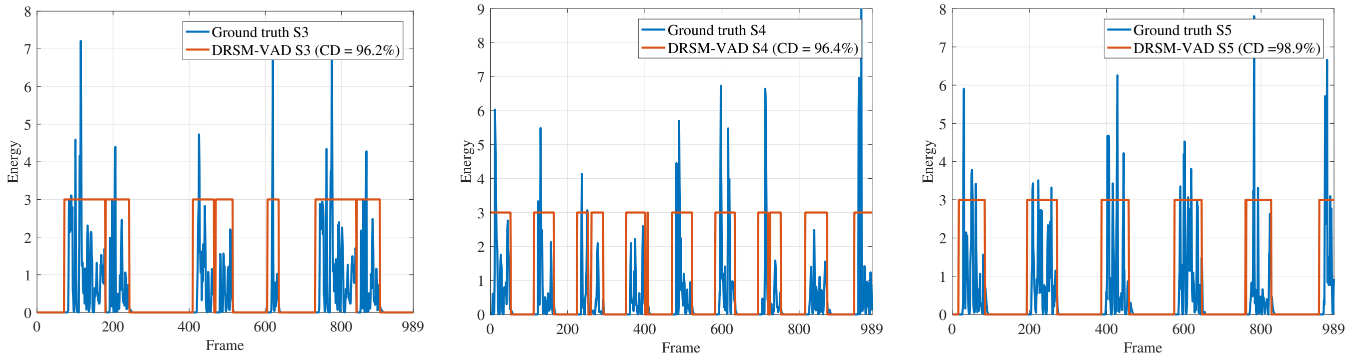
References

[1] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 1999, pp. 789–792.



(a) (b) (c)

Fig. 2: The impact of varying the degree of freedom ν on the outcome of the proposed distributed SRM-VAD in terms of (a) correct detection level, (b) false alarm rate, and (c) misdetection percentage.



(a) (b) (c)

Fig. 3: The acquired VAD patterns (red) using our SRM-VAD approach in the distributed setup for (a) S3, (b) S4, and (c) S5.

[2] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[3] S. M. Joseph and A. P. Babu, “Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding,” *Int. J. Speech Technol.*, pp. 1–14, 2016.

[4] S. J. Park, C. Lee, and D. H. Youn, “A residual echo cancellation scheme for hands-free telephony,” *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 397–399, 2002.

[5] H. Zhao, G. Wang, and X. Peng, “A novel voice activity detection method using energy statistical complexity,” in *IEEE Fifth Int. Conf. Bio-Inspired Comput.: Theories Appl. (BIC-TA)*, Sept. 2010, pp. 1175–1179.

[6] R. Tucker, “Voice activity detection using a periodicity measure,” *IEEE Proc. Commun., Speech Vision*, vol. 139, no. 4, pp. 377–380, Aug. 1992.

[7] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, “Comparative study on voice activity detection algorithm,” in *Int. Conf. Electr. Control Eng. (ICECE)*, June 2010, pp. 599–602.

[8] J. Sohn, N. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[9] J. H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[10] A. Bertrand and M. Moonen, “Energy-based multi-speaker voice activity detection with an ad hoc microphone array,” in *IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP*, 2010, pp. 85–88.

[11] ——, “Blind separation of non-negative source signals using multiplicative updates and subspace projection,” *Signal Process.*, vol. 90, no. 10, pp. 2877–2890, 2010.

[12] S. Maraboina, D. Kolossa, P. Bora, and R. Orglmeister, “Multi-speaker voice activity detection using ICA and beampattern analysis,” in *14th IEEE Eur. Signal Process. Conf. (EUSIPCO)*, 2006, pp. 1–5.

[13] M. Taghizadeh, P. Garner, H. Bourlard, H. Abutalebi, and A. Asaei, “An integrated framework for multi-channel multi-source localization and voice activity detection,” in *IEEE Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA)*, 2011, pp. 92–97.

[14] G. Chen, K. Kumatani, J. McDonough, and B. Raj, “Distant multi-speaker voice activity detection using relative energy ratio.”

[15] T. F. Bergh, I. Hafizovic, and S. Holm, “Multi-speaker voice activity detection using a camera-assisted microphone array,” in *IEEE Int. Conf. Syst., Signals, Image Process. (IWSSIP)*, 2016, pp. 1–4.

[16] 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 *Proc. 42nd IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 2017.

[17] M. H. Bahari, L. K. Hamaidi, M. Muma, J. Plate-Chaves, M. Moonen, A. M. Zoubir, and A. Bertrand, “Distributed multi-speaker voice activity detection for wireless acoustic sensor networks,” *IEEE Trans. Audio, Speech, Language Process., (submitted)*, 2017.

[18] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Adv. neural inf. process. syst.*, 2006, pp. 801–808.

[19] P. O. Hoyer, “Non-negative sparse coding,” in *Proc. 12th IEEE Workshop on Neural Netw. Signal Process.*, 2002, pp. 557–565.

[20] S. Chouvardas, M. Muma, K. Hamaidi, S. Theodoridis, and A. M. Zoubir, “Distributed robust labeling of audio sources in heterogeneous wireless sensor networks,” in *Proc. 40th IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP’15*, 2015, pp. 5783–5787.

[21] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, “Biclustering via sparse singular value decomposition,” *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010.

[22] N. Meinshausen and P. Bühlmann, “Stability selection,” *J. Roy. Stat. Soc.: Series B (Stat. Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[23] M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider, “Robust biclustering by sparse singular value decomposition incorporating stability selection,” *Bioinformatics*, vol. 27, no. 15, pp. 2089–2097, 2011.

[24] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, “Complex elliptically symmetric distributions: Survey, new results and applications,” *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, 2012.