Robust Distributed Multi-Source Detection and Labeling in Wireless Acoustic Sensor Networks

Vom Fachbereich 18
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von
Lala Khadidja Hamaidi, M.Sc.
geboren am 22.11.1989 in Algier (Algerien)

Referent:	Prof. Dr.-Ing. Abdelhak M. Zoubir (TU Darmstadt)
Korreferent:	Dr.-Ing. Michael Muma (TU Darmstadt)
Korreferent:	Asst. Prof. Dr.-Ing. Alexander Bertrand (KU Leuven)
Tag der Einreichung:	19.09.2017
Tag der mündlichen Prüfung:	13.12.2017

D 17
Darmstadt, 2018

'Gebt mir einen festen Punkt auf dem ich stehen kann und ich werde die Welt aus den Angeln heben!'

Archimedes

To my Family and Friends.

Acknowledgments

I would like to express my deepest gratitude to Prof. Dr.-Ing. Abdelhak Zoubir for his supervision and his enthusiastic encouragement. Prof. Dr.-Ing. Zoubir has been supportive of my career objectives and contributed actively in providing me with the protected academic time and research guidance to pursue those objectives. I wholeheartedly appreciate his advice and scientific feedback that challenged as much as motivated me to accomplish my doctoral degree.

I would like to show my gratitude to Dr.-Ing. Michael Muma for his co-supervision and valuable suggestions during the planning and development of my research work. Dr.-Ing. Muma has granted me with the essential and constructive scientific comments. I am grateful for his willingness to give his time so generously, which fosters my skills in Signal Processing.

I would like to offer my special thanks to my second co-supervisor Asst. Prof. Dr.-Ing. Alexander Bertrand from the Katholieke Universiteit Leuven for his support and readiness to share scientific material to promote research. His work inspires my research direction, as I have always been a follower of his scientific publications and manuscripts. Thank you for making the way from Belgium to especially attend my doctoral defense. I also take the chance to thank Prof. Dr.-Ing. Rolf Jakoby, Prof. Dr. mont. Mario Kupnik, and Prof. Dr.-Ing. Ulrich Konigorski for their hard work as Board of Examiners.

This work would not have been possible without the financial support of the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering (GSC CE) at Technische Universität Darmstadt. I also would like to express my very great appreciation for the people at the GSC CE, in particular Prof. Dr. rer. nat. Michael Schäfer, Dr. Markus Lazanowski, Dr.-Ing. Melanie Gattermayer, Carina Schuster, Christian Schmitt, and all my colleagues at the GSC CE. I especially thank Prof. Dr. Frank Aurzada from the Stochastics group at the Mathematical department of Technische Universität Darmstadt for his support.

I wish to thank Renate Koschella, Hauke Fath, Dr.-Ing. Nevine Demitri, Dr.-Ing. Sara Al-Sayed, Dr.-Ing. Michael Fauß, Mark Ryan Leonard, Dominik Reinhard, Di Jin, Adrian Šošić, Dr.-Ing. Jürgen Hahn, Dr.-Ing. Christian Weiss, Dr.-Ing. Simon Rosenkranz, Tim Schäck, Sergey Sukhanov, Ann-Kathrin Seifert, Patricia Binder, Dr.-Ing. Wassim Suleiman, Dr.-Ing. Christian Debes, Dr..-Ing. Stefano Fortunati, Dr.-Ing.

Fiky Suratman, Dr.-Ing. Gökhan Gül, Dr.-Ing. Feng Yin, Dr.-Ing. Mouhammad Alhumaidi, Dr.-Ing. Stefan Leier, Dr.-Ing. Roy Howard and ALL my colleagues and former colleagues of the Signal Processing Group at TU Darmstadt who provided me with a lively environment. In particular, my greatest thanks to my friends and office mates Dr.-Ing. Sahar Khawatmi and Freweyni Teklehaymanot. It has been a real pleasure to know you and gather some of the fondest memories in the journey of PhD.

I would also like to extend my thanks to my friends Edlira, Ewa, and Anissa who contributed greatly to my happiness during the PhD years in Darmstadt. I am particularly grateful to Mr. Gerhard and Mrs. Friederun Seichter, and to Mrs. Petra Thielmann for their constant cheering, discussions and the feeling of home I have had at full length of our time as neighbors.

Finally, I am most grateful to my parents Salah and Fatma Zohra, my sisters Soumia and Keltoum, and my brother Salah Eddine, for their unconditional love and support throughout my life. You gave me the strength to be who I am, which I always value. My sincere thanks go to my uncle Abdelkader Benrokia for believing in me and standing by my side.

Darmstadt, 03.01.2018

Kurzfassung

Die steigende Nachfrage nach komplexen Signalverarbeitungsverfahren in Verbindung mit niederenergetischen, großen, drahtlosen, akustischen Sensornetzwerken, sogenannten *wireless acoustic sensor networks* (WASNs) treibt den Wandel zu einem neuen Paradigma der Informations- und Kommunikationstechnologien (ICT) voran. Die aufkommende Forschungsrichtung strebt eine attraktive drahtlose Netzwerkkommunikation an, bei der mehrere heterogene Geräte mit unterschiedlichen Interessen an verschiedenen Signalverarbeitungsaufgaben kooperieren können. Im Englischen wird hierfür der Begriff *multiple devices cooperating in multiple tasks* (MDMT) verwendet. Diese Dissertation beschäftigt sich mit der verteilten Mehrquellen-Erkennung und -Kennzeichnung zur Verbesserung von Audiosignalen, die eine MDMT-gestützte, knotenspezifische Signalverstärkung in WASNs verfolgen. Tatsächlich ist eine genaue Erkennung und Kennzeichnung eine Grundvoraussetzung, um das MDMT-Paradigma zu verfolgen, bei dem die Knoten im WASN effektiv die Quellen ihres Interesses kommunizieren und somit mehrere Signalverarbeitungsaufgaben durch Zusammenarbeit verbessert werden können.

Zu Beginn wird ein neuartiges Rahmenwerk vorgestellt, das auf einem dominanten Quellenmodell in dezentralen WASNs zur Aktivitätsdetektion mehrerer aktiver Sprachsignalquellen in einer halligen und lauten Umgebung basiert. Eine vorläufige, multiplikative, nicht-negative, unabhängige Rank-eins-Komponentenanalyse (M-NICA) zur Extraktion dominanter Energiequellen anhand der zugehörigen Knotencluster wird vorgestellt. Algorithmen, die die mittlere absolute Abweichung und gewichtete mittlere absolute Abweichung innerhalb des Clusters minimieren, werden vorgeschlagen, um die Clusterzugehörigkeit der getrennten Energien zu bestimmen und so eine quellspezifische Sprachaktivitätserkennung zu erreichen.

Des Weiteren wird eine Verbesserung der Energie-Signaltrennung zur Vereinfachung der Mehrfachquellen-Aktivitäts-Diskriminierung angestrebt. Auf iterativen Rank-eins-Singulärwert-Zerlegungsebenen werden Regularisierungsterme angewandt, die Dünnbesetztheit induzieren. Anschließend wird mittels multiplikativer Aktualisierungen eine dünnbesetzte, nicht-negative, blinde Energietrennung vollzogen. Somit wird das Problem der Mehrfachquellenerkennung in eine dünnbesetzte, nicht-negative Quellenergie-Dekorrelation umgewandelt. Dünnbesetztheit stimmt die vermeintlich nicht aktiven Energiesignaturen exakt auf Null-Energien ab, sodass es einfacher ist, aktive Energien zu identifizieren, und ein Aktivitätsdetektor unkompliziert aufgebaut werden kann. In einem zentralisierten Szenario wird die Aktivitätsentscheidung von einem Fusionszentrum gesteuert, das die binäre Quellaktivitätsdetektion für jede teilnehmende En-

ergiequelle liefert. Diese Strategie liefert präzise Erkennungsergebnisse für eine kleine Anzahl von Quellen. Bei einer wachsenden Anzahl von Störquellen ist die verteilte Detektion vielversprechender. Gleichzeitig wird ein robuster, verteilter Energietrennungsalgorithmus für mehrere konkurrierende Quellen vorgeschlagen. Hierzu wird eine robuste und regularisierte $t_\nu M$-Schätzung der Kovarianzmatrix der gemischten Energien verwendet. Dieser Ansatz führt zu einer einfachen Aktivitätsentscheidung, bei der nur die robust getrennten Energiesignaturen der Quellen im WASN verwendet werden. Die Leistung des robusten Aktivitätsdetektors wird mit einem verteilten, adaptiven, knotenspezifischen Signalschätzverfahren zur Sprachverbesserung validiert. Im Gegensatz zur ursprünglichen M-NICA für die Quelltrennung verbessern die extrahierten binären Aktivitätsmuster im Zusammenspiel mit der robusten Energietrennung die knotenspezifische Signalschätzung signifikant.

Aufgrund der durch den zusätzlichen Schritt der Energiesignaltrennung verursachten, erhöhten Rechenkomplexität wird ein neuer Ansatz zur Lösung der Detektionsfrage von Mehrfachgeräte-Mehrfachquellen-Netzwerken vorgestellt. Stabilitätsselektion wird zur iterativen Extraktion robuster, rechts-singulärer Vektoren berücksichtigt. Die Unterabtastungs-Auswahlmethode sorgt für Transparenz bei der korrekten Auswahl der Regularisierungsvariablen im Lasso-Optimierungsproblem. Auf diese Weise bilden die stärksten dünnbesetzten, rechts-singulären Vektoren mit einer robusten ℓ_1-Norm und Stabilitätsselektion die Basisvektoren, die die Eingangsdaten effizient beschreiben. Sie werden mit einer robusten, unbeaufsichtigten Methode auf der Basis einer Norm ℓ_1 ermittelt. Die Klassifizierung der aktiven/nicht-aktiven Quellen erfolgt eines robusten Mahalanobis-Klassifikators. Hierzu wird ein robuster M-Schätzer der Kovarianzmatrix in der Mahalanobis-Distanz verwendet. Umfangreiche Auswertungen in zentralisierten und verteilten Szenarien werden durchgeführt, um die Effektivität des vorgeschlagenen Ansatzes zu bewerten. Die Überwindung der rechenintensiven Quellentrennung ist somit möglich, indem die robuste Stabilitätsselektion für die Extraktion von Multi-Energiemerkmalen genutzt wird.

Im Hinblick auf das Kennzeichnungsproblem verschiedener Quellen in einem WASN wird ein robuster Ansatz eingeführt, der die Einfallsrichtung der ankommenden Quellsignale ausnutzt. Ein auf der Kurzzeit-Fourier-Transformation basierendes Unterraumverfahren schätzt die Winkel von lokal stationären Breitbandsignalen mit Hilfe einer gleichförmigen linearen Sensorgruppe. Der Median der Winkel, die bei jedem Frequenzbereich geschätzt werden, wird verwendet, um den Gesamtwinkel für jede teilnehmende Quelle zu erhalten. Die Merkmale nutzen in diesem Fall die geräteübergreifende Ähnlichkeit in den jeweiligen Frequenzbereichen aus, die eine zuverlässige Schätzung der Ankunftsrichtung für jede Quelle liefern. Die Zuverlässigkeit

wird in Bezug auf den Median über die Frequenzen hinweg definiert. Alle quellspezifischen Frequenzbänder, die zur korrekten Schätzung der Winkel beitragen, werden ausgewählt. Für jede Quelle wird an jedem Gerät ein Merkmalsvektor gebildet, in dem die Indizes der Frequenzbereiche gespeichert werden, die innerhalb des oberen und unteren Intervalls der mittleren absoluten Abweichungsskala des geschätzten Winkels liegen. Die Kennzeichnung erfolgt durch ein verteiltes Clustering der extrahierten winkelbasierten Merkmalsvektoren mittels Konsensmittelung.

Abstract

The growing demand in complex signal processing methods associated with low-energy large scale wireless acoustic sensor networks (WASNs) urges the shift to a new information and communication technologies (ICT) paradigm. The emerging research perception aspires for an appealing wireless network communication where multiple heterogeneous devices with different interests can cooperate in various signal processing tasks (MDMT). Contributions in this doctoral thesis focus on distributed multi-source detection and labeling applied to audio enhancement scenarios pursuing an MDMT fashioned node-specific source-of-interest signal enhancement in WASNs. In fact, an accurate detection and labeling is a pre-requisite to pursue the MDMT paradigm where nodes in the WASN communicate effectively their sources-of-interest and, therefore, multiple signal processing tasks can be enhanced via cooperation.

First, a novel framework based on a dominant source model in distributed WASNs for resolving the activity detection of multiple speech sources in a reverberant and noisy environment is introduced. A preliminary rank-one multiplicative non-negative independent component analysis (M-NICA) for unique dominant energy source extraction given associated node clusters is presented. Partitional algorithms that minimize the within-cluster mean absolute deviation (MAD) and weighted MAD objectives are proposed to determine the cluster membership of the unmixed energies, and thus establish a source specific voice activity recognition.

In a second study, improving the energy signal separation to alleviate the multiple source activity discrimination task is targeted. Sparsity inducing penalties are enforced on iterative rank-one singular value decomposition layers to extract sparse right rotations. Then, sparse non-negative blind energy separation is realized using multiplicative updates. Hence, the multiple source detection problem is converted into a sparse non-negative source energy decorrelation. Sparsity tunes the supposedly non-active energy signatures to exactly zero-valued energies so that it is easier to identify active energies and an activity detector can be constructed in a straightforward manner. In a centralized scenario, the activity decision is controlled by a fusion center that delivers the binary source activity detection for every participating energy source. This strategy gives precise detection results for small source numbers. With a growing number of interfering sources, the distributed detection approach is more promising. Conjointly, a robust distributed energy separation algorithm for multiple competing sources is proposed. A robust and regularized $t_\nu M$-estimation of the covariance matrix of the mixed energies is employed. This approach yields a simple activity decision using only the robustly unmixed energy signatures of the sources in the WASN. The

performance of the robust activity detector is validated with a distributed adaptive node-specific signal estimation method for speech enhancement. The latter enhances the quality and intelligibility of the signal while exploiting the accurately estimated multi-source voice decision patterns. In contrast to the original M-NICA for source separation, the extracted binary activity patterns with the robust energy separation significantly improve the node-specific signal estimation.

Due to the increased computational complexity caused by the additional step of energy signal separation, a new approach to solving the detection question of multi-device multi-source networks is presented. Stability selection for iterative extraction of robust right singular vectors is considered. The sub-sampling selection technique provides transparency in properly choosing the regularization variable in the Lasso optimization problem. In this way, the strongest sparse right singular vectors using a robust ℓ_1-norm and stability selection are the set of basis vectors that describe the input data efficiently. Active/non-active source classification is achieved based on a robust Mahalanobis classifier. For this, a robust M-estimator of the covariance matrix in the Mahalanobis distance is utilized. Extensive evaluation in centralized and distributed settings is performed to assess the effectiveness of the proposed approach. Thus, overcoming the computationally demanding source separation scheme is possible via exploiting robust stability selection for sparse multi-energy feature extraction.

With respect to the labeling problem of various sources in a WASN, a robust approach is introduced that exploits the direction-of-arrival of the impinging source signals. A short-time Fourier transform-based subspace method estimates the angles of locally stationary wide band signals using a uniform linear array. The median of angles estimated at every frequency bin is utilized to obtain the overall angle for each participating source. The features, in this case, exploit the similarity across devices in the particular frequency bins that produce reliable direction-of-arrival estimates for each source. Reliability is defined with respect to the median across frequencies. All source-specific frequency bands that contribute to correct estimated angles are selected. A feature vector is formed for every source at each device by storing the frequency bin indices that lie within the upper and lower interval of the median absolute deviation scale of the estimated angle. Labeling is accomplished by a distributed clustering of the extracted angle-based feature vectors using consensus averaging.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Detection and Labeling in Wireless Acoustic Sensor Networks

'Avant donc que d'écrire, apprenez à penser.
Ce que l'on conçoit bien s'énonce clairement,
Et les mots pour le dire arrivent aisément.
Hâtez-vous lentement, et sans perdre courage.'

Nicolas Boileau

Traditional microphone arrays suffer from weak intelligibility of the recorded speech at the microphones due to local sampling of the sound field often at large distance from the sources. Together with their size and processing power limitations, microphone arrays are not sufficiently performant for many demanding applications [6]. Wireless acoustic sensor networks (WASN) consisting of spatially distributed wireless nodes (see Fig. 1.2) equipped with one or more microphones that are supplied with wireless communication and computation capabilities overcome these restrictions. Contrary to single-node, higher quality recordings are perceived in WASNs by taking advantage of the spatial diversity of the participating nodes in the network. This fact allows WASNs for improved speech enhancement algorithms compared to single-node methods [7, 8]. In spite of the challenges in designing WASNs due to the significant data traffic in the network, many researchers in the field have acknowledged the benefits of WASNs to improve the quality and/or the intelligibility of the observed speech signals corrupted by noise. WASNs are beneficial in a wide variety of research fields, such as hearing aids, echo cancellation and hands-free telephony [9], teleconferencing systems, automatic speech recognition (ASR) and speaker recognition [6, 10–15], speech coding systems [16], computer games, and speech enhancement [17–21]. In the centralized WASN configuration nodes transmit their observations to a fusion center (FC) that performs all processing. Recently, distributed speech enhancement algorithms, such as the distributed adaptive node-specific signal estimation (DANSE) algorithm, have been developed for WASNs [7, 8, 11, 22–33]. Distributed WASNs disseminate the computations among the nodes and thus do not depend upon a FC. The communication cost in this case is reduced by an exchange of information within neighboring nodes. Moreover,

distributed WASNs are robust against failures of the FC and scalable to larger networks at the cost of an increased computational effort at the different wireless nodes [7,8,34].

1.1 Multiple Devices for Multiple Tasks (MDMT) for WASNs

With the abrupt advances in the area of signal processing, ceaseless high-demanding constraints are perceived on wireless networks of microphone arrays. In most of the ongoing information and communication technologies (ICT), wireless communications using microphone arrays permit multi-functional sensor nodes that are low-cost, low-power, and small in size to communicate in an untethered fashion and collaborate as a group to achieve a signal processing task. These sensor nodes, equipped with sensing, data processing, and communication components, leverage the strength of the cooperative effort to supply increased sensing quality and higher precision in task fulfillment in time and space. In such a scenario, sensor nodes have the same specific intent and cooperation grants them the possibility to solve a unique signal processing task. This concept is still lagging behind compared to the non-stop increase in complexity of signal processing tasks and diversity of devices [35].

In fact, current research targets a new ICT paradigm that considers multiple heterogeneous devices with different interests, i.e., node-specific interests, cooperating in various signal processing tasks in a WASN. In this sense, an emerging new paradigm is that of multiple devices cooperating in multiple tasks (MDMT) [35]. This is different from the classical ICT strategies employed in wireless sensor network setups, in which multiple devices perform one single joint task. There is a potential to improve the network performance if multiple devices are cooperating in order to solve multiple signal processing tasks in a sensor network [35].

The MDMT technology encompasses aspects derived from various signal processing fields of study including, e.g., distributed adaptive speech enhancement, which offers a better node-specific audio signal enhancement [7,8,11,18] based on consistent, common, unique labeling of all relevant sources that are observed by the network, as well as precise voice activity detection (VAD) for the targeted multiple audio sources [1, 5, 34, 36, 37]. Researchers are targeting headway in establishing fundamental perimeters of this field that is experiencing a continuing growth with new applications [35]. The MDMT paradigm aims for a superior performance of the distinct signal processing tasks in case of diverse participating sensors. For seek of proof of the usefulness of the

MDMT paradigm, in this thesis an emphasis is put on robust distributed multi-source detection and labeling applied to audio enhancement scenarios aspiring a node-specific source-of-interest signal enhancement.

1.2 Motivation and Research Goals of This Doctoral Project

The labeling problem refers to providing a consistent identifier (label) to each of the sources over the WASN. In other words, the speech sources should have the same labels throughout the network. Furthermore, voice activity detection algorithms aim at deciding whether sources of interest are active or not. Many detection and labeling problems in statistical signal processing rely on specific assumptions, for instance, a specific parametric signal model, Gaussianity, independence, and stationarity. Often, these assumptions do not hold in practice, for example, the presence of non-Gaussian and impulsive noise has been confirmed in several measurement campaigns. In these situations, the performance of detection or labeling algorithms may drastically degrade. Conventional approaches for detection and labeling assume that all nodes have a single common underlying objective. In addition, distributed multi-source detection and labeling remains an unsolved problem. Hence, with regards to speech processing applications, the major interest of this PhD project is to solve two main problems:

- Distributed multi-source voice activity detection, see [Chapter 2, Chapter 3], and
- Distributed multi-source labeling of a known number of interfering sources in a challenging MDMT WASN, see [Chapter 4].

The fulfillment of the detection and labeling tasks for multiple interfering speech sources in a distributed environment is crucial. If an accurate detection and labeling is conduced, the nodes in the WASN are subsequently able to efficiently communicate their sources-of-interest and, consequently, any signal processing task can be enhanced via cooperation. Based on the multi-device multi-source WASN described by Fig. 1.2, Fig. 1.1 presents an intuitive sketch of the labeling/detection of active sources in a WASN. Notice that when targeting such practical use-case applications, no prior knowledge, such as the noise distribution, or the source and node positions and array orientations, is assumed to be available to the algorithms. These challenging points are mostly taken

as available preliminary knowledge in the conventional techniques treating the detection and source labeling questions, which makes the problem tractable and feasible to solve.

This work investigates the distributed multi-source detection/labeling framework from a generic point of view. Hence, a focal point is the extraction of robust low-dimensional feature vectors, which are required to lead to a unique labeling and an accurate activity detection. Furthermore, due to the non-stationarity of the speech signals and the absence of a FC, distributed and adaptive approaches are considered. Our aim is to develop robust distributed multi-source detection and labeling techniques in WASN scenarios with node-specific interests, which fulfills the MDMT concept. To achieve a node-specific decision, nodes cooperate locally within an ad hoc network structure. This requires detecting sources of interest for the different nodes in adverse environments taking into account robust measures. Distributed multi-source detection and labeling is a new research field and an important enabler for MDMT systems. Since the received speech signals are a mixture of multiple sources, a preliminary step of detection can be to unmix the received mixtures. A large focus is placed on developing unmixing methods for VAD throughout this thesis. This study intuitively provides an extracted set of unmixed features from the considered audio speech scenario depicted in Fig. 1.2. Precisely, our detection and labeling techniques should be robust with respect to interference and background noise environment uncertainties. Moreover, adaptiveness is an important key so as to cope with environment non-stationarities.

1.3 Multi-Source Multi-Device WASN Use-Case

Consider a WASN that is deployed in a public environment, such as an airport hall, a meeting room or a conference hall. Multiple speakers (sources) are simultaneously active in the network. In order to perform subsequent speech enhancement, source specific labels and voice activity patterns are required. Developing algorithms which do not necessitate a priori information, such as positions and orientations of the devices in the WASN, is essential.

Figure 1.2 shows a simulated public scenario with 7 active sources $\{S_1, \cdots, S_7\}$ and 20 nodes $\{D_1, \cdots, D_{20}\}$ that form the WASN in a reverberant 20×10 meter room. Data which follows such a setup has been generated within the HANDiCAMS[1] project and

[1]HANDiCAMS is an EU-funded project where researchers aim at developing new distributed and adaptive signal processing algorithms under a novel paradigm where multiple devices cooperate in multiple tasks (MDMT) to achieve superior performance in their node-specific interests. More information about this project can be found on: http://www.handicams-fet.eu/

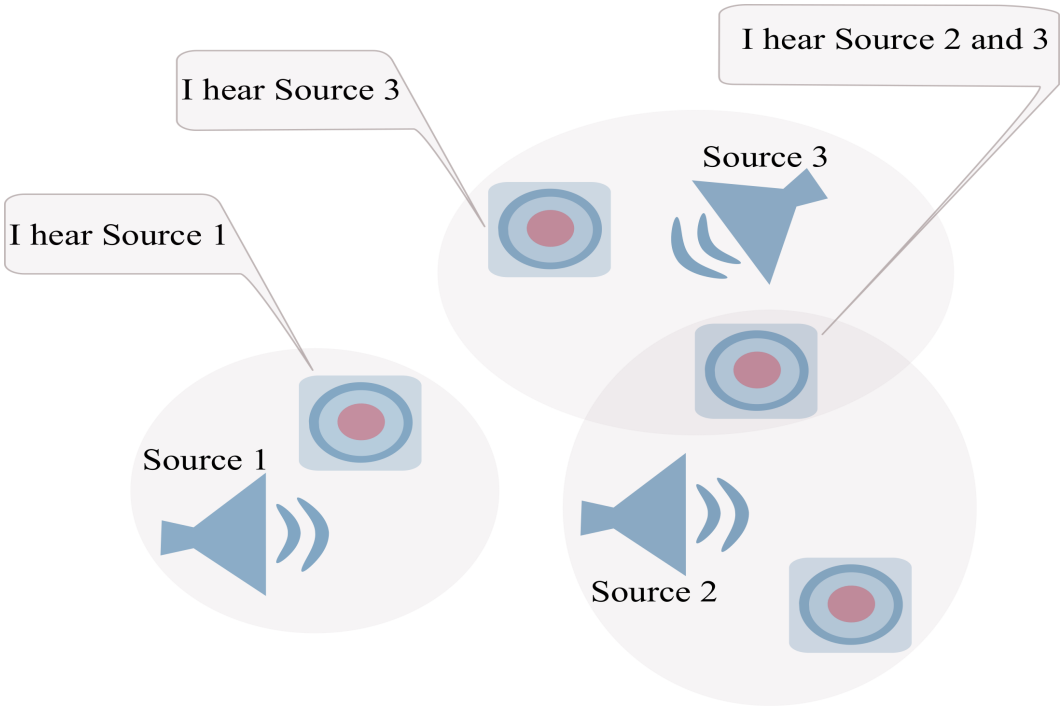


Figure 1.1: Nodes identifying well labeled active speech sources in a WASN.

is used for the evaluation of our proposed algorithms.

The sources (red) transmit speech signals that are recorded at the multiple sensor nodes (blue) of the WASN. The nodes of the network are heterogeneous and can be portable devices, such as mobile phones, or hearing aids. These are necessarily placed nearly to their owners (source-of-interest). For the sake of simplification, we consider a static use-case throughout this project. This means that the speech sources of the network do not move. Sensor nodes, colored in blue in Fig. 1.2, are composed of 3 microphones aligned as a uniform linear array (ULA). The distance between every two microphones is 1.5cm. Every source in the WASN represents a male or a female voice signal. We consider a language-independent speech use-case, in the sense that the targeted sources emit speech signals in different languages. The speakers in Fig. 1.2 emit signals that are recorded and sampled at the different microphone nodes with a sampling frequency of $f_s = 16$kHz. In such a speech scenario, sound signals are received as mixtures at the different microphones. Cross-talk and noise are components that appear with different powers at the multiple microphones. More precisely, the speech is affected by spatially independent additive white Gaussian noise (AWGN) or babble noise and the nodes are additionally disturbed by numerous interferers. Note that, in this setting, each sensor receives a delayed and filtered version of the signals based on the room-impulse-response.

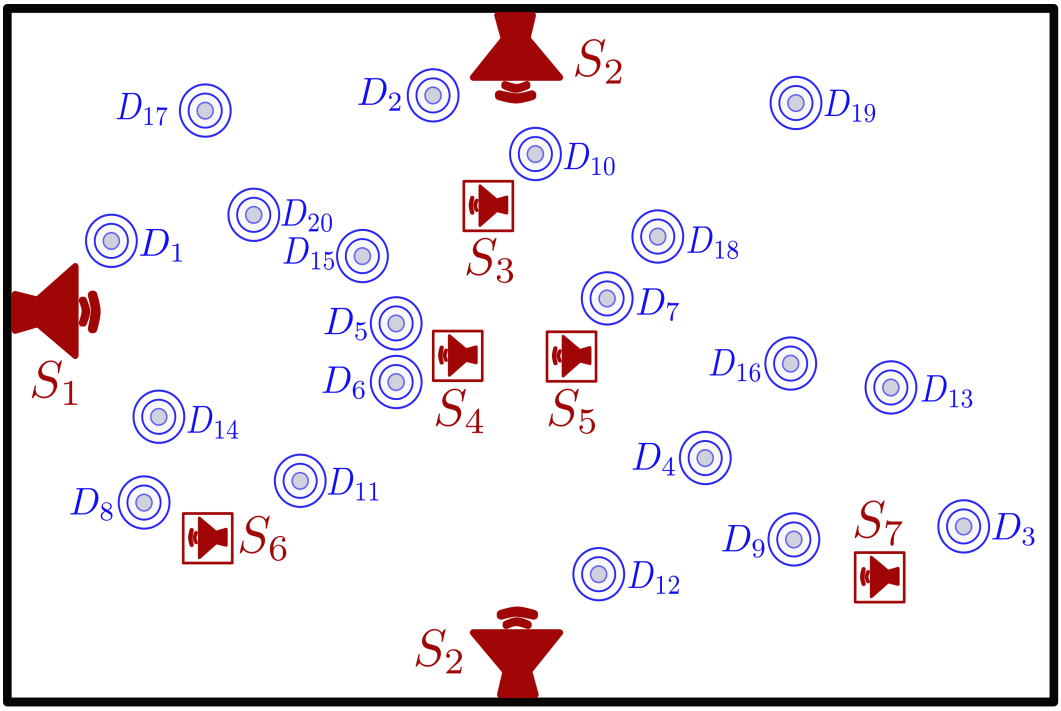


Figure 1.2: The speech use-case scenario: An example of a WASN observing seven speech sources (red) in a 20×10m room with reverberation time T60 $= 0.3$s. The microphone signals of the nodes (blue) are sampled at 16kHz. Source S_2 models a public address system playing an announcement broadcasted from two loudspeakers. Sources S_1, S_3, S_4, S_5, S_6 and S_7 are six different speech sources.

Each device may be interested in enhancing a certain source signal. For the described multiple source and multiple device WASN use-case, we propose, in this doctoral project, novel algorithms to robustly and collaboratively determine the labeling information and speech activity of the participating speech sources from the received mixtures.

1.4 Detection and Labeling: Related Works

1.4.1 Voice Activity Detection in Wireless Acoustic Sensor Networks

Voice activity detection (VAD) distinguishes periods of speech from periods containing only noise. VAD is ubiquitous in speech processing applications such as speech enhancement, speech coding, speaker and speech recognition. The VAD approaches

amount to a pair of research fields compiled in statistical feature extraction and generative/discriminative models. In this work, the former steps are building blocks for both detection and labeling tasks.

Research is still in full swing for efficient voice activity detectors, especially when multiple sources are active. Reliable automatic speech/non-speech detection is crucial for a number of different speech processing algorithms. This area of study affects a number of SP applications.

Conventional VADs are sensitive to a variably noisy environment, especially at low signal-to-noise ratio (SNR). This usually results in cutting off unvoiced regions of the speech signal and random oscillations of the detector's output. The classical single source VAD approaches are based on thresholding procedures [3, 38–42]. The thresholds are set based on a study that considers the different parameters of the voice signal to obtain simply structured VAD algorithms. However, these methods fail at finding the exact active/non-active transitions of a speech signal. Their immunity to noise is bounded, which makes them particularly not efficient under low SNR environments. Moreover, short-term energy-based detectors are proposed in [16, 43, 44]. Short-term features employed by these techniques are advantageous since they capture the local statistics of a signal at short periods where stationarity is assumed. In addition, these methods have the merit of tracking the noise statistics properly for improved threshold estimation. A front-end VAD technique is conveyed in [45] for speech signals buried in noisy and reverberant environments. The study is based on the modulation transfer function concept to reduce the ill effects of noise and reverberation for speech, and propose a robust VAD method. Empirical mode decomposition (EMD) together with modulation spectrum analysis (MSA) are employed in [46] for a robust VAD. EMD is used for reducing the background noise without estimating the SNR. Later, determining speech/non-speech periods is done using the MSA approach. The authors in [47] describe a unified approach meant for jointly solving the SP tasks related to: the underdetermined blind source separation (BSS), source activity detection, dereverberation and direction-of-arrival (DoA), through the estimation of the parameters of an overall generative model. The designed novel VAD algorithms in [48, 49] incorporate machine learning for detection where artificial neural network (ANN) classifiers are trained to assess superior speech discrimination. However, machine learning-based VADs require large data to properly learn the model. Determining speech activity based on time-frequency representations of noisy microphone signals are established in [50, 51]. The study of speech in the spectro-temporal space reveals a significant improvement of the VAD performance directed at solving various speech related applications, such as enhancing speech systems in low SNRs. More sophisticated approaches are based on the statistical modeling of the VAD problem and are proposed in [2, 52–62]. Statistical

methods for VAD report significant improvements in speech/non-speech discrimination accuracy over existing conventional VAD techniques. Nonetheless, they are based on more complex models and require strong assumptions, e.g. to compare a likelihood ratio with an appropriate thresholding function. Developing robust VAD methods that do not break down under very low SNRs is necessary. Effective VAD algorithms in terms of robustness and speech recognition performance in noisy environments are defined in [54, 63–66]. The proposed techniques, however, show a trade-off between detection accuracy and computational cost.

Even with these existing challenges, single-speaker speech detection can be considered as a well-studied problem. The multi-speaker VAD counterpart, however, remains an open question. Only few research work has targeted designing VAD approaches for multiple simultaneous speech sources [67–71]. Another framework presented in [4] considers the multi-speaker VAD as tracking the power of multiple simultaneous speech signals using an ad hoc microphone array with unknown microphone positions. By considering short-term energy-based recordings of the microphone signals, the multi-speaker VAD question can be converted into a non-negative blind source separation (NBSS) problem with non-negative sources, which can be solved efficiently with second-order statistics only.

1.4.1.1 Discussion

Regarding the single source detection case, the presented methods operate well under specific conditions, but their performance depends highly on the choice of thresholds, the efficient estimation of which remains an open question. In addition, these methods are single-channel detection solutions. Any further signal transmission using the same channel is considered as an interferer and will cause the detector to break down. In our proposed approaches, we consider a multi-source framework. A repeated sound source at a different position, obviously recorded with different power, interestingly collaborates in improving the quality of the proposed detectors in both single and multi-source schemes. On the other hand, results of the state-of-the-art methods relating to multi-source detection are not explicitly presented in the scientific manuscripts. Additionally, the multi-source use-case in some of these methods is defined such as, at a specific time instant, only one source is considered active while the others are interfering sources. From our perspective, we consider such problems as single source detection, because only a unique VAD pattern corresponding to the active source is generated after processing. A sequential processing of these methods is not feasible.

1.4.2 Source Labeling in Signal Processing

In contrast to centralized algorithms that confide the network's final decision to a central node, distributed learning for labeling is based on the cooperation between nodes over a network, which insures a minimal overall network risk. No coordinating FC that has access to the entire data from all nodes is needed. When data features are distributed across different nodes, relaying information to a centralized processing unit is discouraged due to the communication costs. A fully distributed scheme is favorable to solve the labeling task for the sake of scalability, reduced communication complexity, and robustness to isolated points of failure, i.e., a possible FC's failure.

Up until today, there is a growing research interest on the topic of distributed classification [72–79]. In particular, a distributed algorithm for supervised learning in the presence of a FC has been proposed in [77] and totally distributed schemes are treated in [74, 75]. Distributed algorithms for unsupervised learning have been proposed in [76, 80]. Nonetheless, to the best of our knowledge, prior to this doctoral project, the distributed labeling task has not yet been addressed for the case of multiple speech sources in a WASN. Recently, distributed labeling techniques have been introduced in [81–83] for labeling multiple objects in camera sensor networks.

1.5 Publications

The following publications have been produced during this doctoral project.

Internationally Refereed Journal Articles

- M. H. Bahari, L. K. Hamaidi, M. Muma, J. Plata-Chaves, M. Moonen, A. M. Zoubir, A. Bertrand, "Distributed Multi-Speaker Voice Activity Detection for Wireless Acoustic Sensor Networks", submitted to *IEEE Trans. Audio, Speech and Language Process.*, March 2017.

Internationally Refereed Conference Papers

- S. Chouvardas, M. Muma, L. K. Hamaidi, S. Theodoridis, A. M. Zoubir, "Distributed Robust Labeling of Audio Sources in Heterogeneous Wireless Sensor

Networks", In Proc. 40^{th} *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Brisbane, Australia, 2015.

- 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", In Proc. 42^{nd} *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, New Orleans, USA, 2017.

- L. K. Hamaidi, M. Muma, A. M. Zoubir, "Robust Distributed Multi-Speaker Voice Activity Detection Using Stability Selection for Sparse Non-Negative Feature Extraction", In the Proc. 25^{th} *European Signal Process. Conf. (EUSIPCO)*, Kos Island, Greece, 2017.

- L. K. Hamaidi, M. Muma, A. M. Zoubir, "Robust Distributed Sparsity-Constrained Non-Negative Source Separation and Multi-Speaker Voice Activity Detection for Distributed Speech Enhancement in Wireless Acoustic Sensor Networks", Submitted to the Proc. 2^{nd} *IEEE Int. Conf. Signals Syst. (ICSigSys)*, Bali, Indonesia, 2018.

1.6 Organization of This Doctoral Thesis

The organization of this doctoral thesis comes as follows:

Chapter 2 introduces a novel framework for the multi-source detection task in a distributed WASN. The suggested framework is based on three well stated steps: distributed source-specific node clustering [1], distributed rank-one source unmixing, and distributed clustering-based multi-source VAD. In fact, the detection task is based on partitional clustering of unmixed signals related to a dominant speech source. Two robust clustering techniques are presented for the aim of a compact multi-source detection method robust to outliers. Extensive simulation results on a speech scenario of 6 sources and 20 nodes prove the effectiveness of the detection technique under various noise conditions.

Chapter 3 focuses on developing an efficient energy-based multiple source separation method. The latter relies on a robust sparse modeling based on the ℓ_1-norm. Source-specific non-negative sparse features related to the right rotations of a sparse singular decomposition are extracted. This is considered as an initial energy separation phase that is post-processed with an actual robust separation

technique that uses multiplicative updates. Feature decorrelation is maximized with a median measure of central tendency used in the computation of the covariance matrix, which makes the proposed method, named sparse median-based multiplicative non-negative independent component analysis (SMM-NICA), robust to outlying energy signatures. A centralized scenario composed of 2 speech sources is considered to test the assessment of the proposed approach and apply it to a straightforward VAD classifier. Promising detection results are presented and discussed.

In addition, in Chapter 3 we derive a robust version of SMM-NICA for energy source unmixing. The multi-speaker VAD problem is converted into a robust and sparse blind source separation problem. The performance in terms of speech activity detection is further evaluated using a distributed and adaptive node-specific signal enhancement (DANSE), see [7,8], scheme where node-specific signal estimation is achieved based on the proposed robust sparsely estimated VAD patterns. A two-phased simulation setup is utilized to prove the accuracy of the suggested robust and sparse-promoting separation technique for VAD and speech enhancement in both centralized and distributed modes.

Furthermore, in Chapter 3, we demonstrate that the need for a complete speech separation technique for the fulfillment of the multi-source VAD can be diverted. In fact, a sparse constrained model based on stability selection is employed to extract robust source-specific feature vectors. Subsequently, robust classification techniques based on a robust $t_\nu M$-estimator are proposed for speech discrimination. Intensive experiments are conducted in both a centralized speech use-case of 2 sources, and an extended distributed approach to resolve speech decision for a challenging 7 source scenario buried in noise. VAD results based on this proposed technique are precise and promising.

Chapter 4 develops a two-step technique for robust multiple source labeling in WASNs. The non-hierarchical technique relies on the similarity information deduced from the frequency bins that generate reliable estimation of direction-of-arrivals (DoAs) at different nodes of the WASN. The labeling of the multiple sources resorts to exploiting these extracted DoA-based feature vectors in a distributed/cooperative unsupervised learning technique based on a similarity measure applied to the feature vectors. The labeling results for different experiments are presented and evaluated.

Chapter 5 settles some concluding points related to the presented multi-source detection and labeling methods. Moreover, some possible open questions, future research, and suggested schemes are identified.

Chapter 2

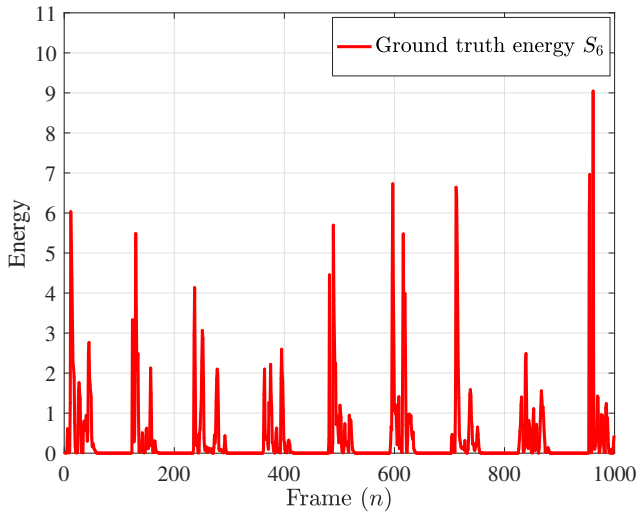
Distributed Multi-Speaker Clustering-Based Voice Activity Detection for Wireless Acoustic Sensor Networks

'Simplicity is the ultimate sophistication'

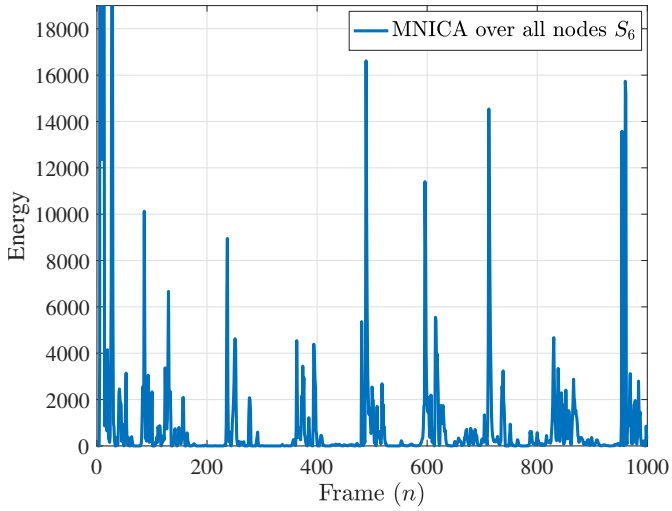
Leonardo da Vinci

In this chapter[1], the focus regards unsupervised learning based on clustering to solve the multi-source detection problem in terms of VAD. We distinguish two settings related to: the availability of data ahead of time where we perform batch multi-source VAD, and the setting where data is streaming-in and decision is made in real-time, which we call sequential multi-source VAD. The multi-source VAD question can be converted into a blind source separation problem with non-negative sources, which can be solved efficiently with second order statistics, by using short-term power measurements at different nodes [1, 4, 5]. Different centralized non-negative signal unmixing methods have been suggested in the literature, for instance, non-negative principal component analysis (NPCA) [84] and multiplicative non-negative independent component analysis (M-NICA) [85]. These algorithms are capable of producing separated source energy signals, from the nodes' observations [4]. However, they require a FC and their unmixing performance severely degrades with an increasing number of active sources, see for instance Fig. 2.1 (b) for an example of unmixing the energy of Source S_6 for the WASN with seven active speech sources that is displayed in Fig. 1.2. Obviously, nodes that are located in the proximity of a source observe the corresponding source signal with a higher power compared to other interfering source signals. Therefore, unmixing the energy signal of this specific source using the recorded signals at these nodes is much easier. Fig. 2.1 (c) depicts the improvement in performance when using M-NICA with only observations collected from devices D_8, D_{11}, and D_{14}, which are located around Source S_6, compared to using centralized M-NICA in Fig. 2.1 (b).

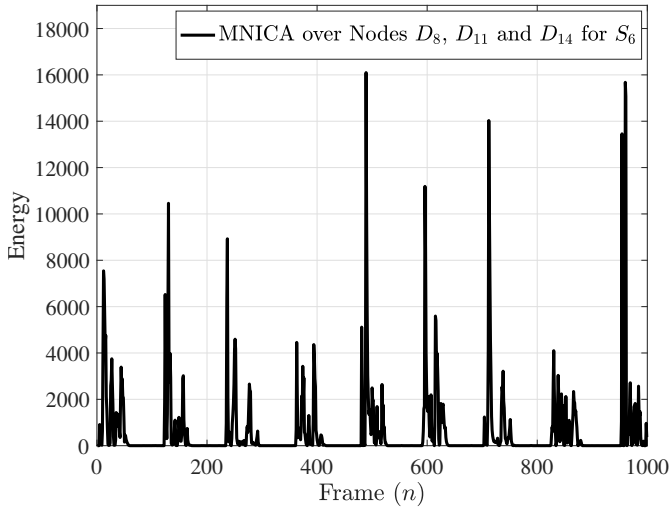
[1]This chapter is based on the journal article entitled: "Distributed Multi-Speaker Voice Activity Detection for Wireless Acoustic Sensor Networks", submitted to the IEEE Trans. Audio, Speech and Language Process. (T-ASL). Our major original contributions are in Section III. D, Section III. E, and Section IV. D of this journal submission.



(a)



(b)



(c)

Figure 2.1: The unmixing result for Source S_6 in the scenario of Fig. 1.2 using (b) M-NICA over the observation of all devices and (c) M-NICA with the observations of devices D_8, D_{11}, and D_{14}.

2.1 Introduction

2.1.1 Background

While single-speaker single-node VAD is a well researched problem [2, 3, 43, 50, 53, 56, 57, 61, 62, 66, 86], to the best of our knowledge, no distributed multi-speaker voice activity detection (DM-VAD) method is available in the literature. Even for centralized WASNs, the literature is sparse [4, 5, 68–71]. The VAD method introduced by Bertrand *et al.* for multiple-concurrent-speakers in centralized WASNs [4] performs a multi-speaker energy pattern extraction by designing an efficient energy unmixing algorithm in a WASN. Nevertheless, after energy separation, no implicit VAD is performed in [4]. Moreover, in [71], 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. This approach is computationally demanding as it operates in the time-frequency domain. An integrated centralized multi-source speaker localization and multi-channel VAD framework is introduced in [69]. The work exploits the behavior of the spatial gradient steered response power function using the phase transform method. While in [68], identifying a single target speaker from multiple speakers in a centralized fashion 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 centralized VAD method [70] 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 our recent work [5].

Many distributed speech enhancement algorithms, such as the distributed adaptive node-specific signal estimation (DANSE) algorithm [7, 8, 11, 27], or the distributed speech enhancement based on multi-channel Wiener filtering (MWF) in [18] require a DM-VAD method to estimate the speech and noise covariances. Therefore, in this chapter, an original DM-VAD for WASNs is proposed. The proposed method neither requires a FC nor prior knowledge about the node positions, microphone array orientations or the number of observed sources.

Exploiting the WASN topology to develop a DM-VAD technique is a promising idea, yet challenging. In fact, it requires a distributed method to locate the nodes around each source (LONAS) [1]. LONAS solves a node clustering problem, where nodes in the vicinity of each source are grouped into a cluster. Only one cluster of nodes should

be formed for every source [1]. LONAS defines a unified framework that answers the source enumeration and the node clustering problems based on adaptive distributed eigenvalue decomposition (EVD) [87–91]. In this chapter, the distributed WASN setting is resolved using LONAS to determine the clusters of nodes for which a single source is dominant. The subsequent VAD in the presence of multi-speakers is then achievable in a distributed mode by applying algorithms that are proposed in this chapter.

2.1.2 Problem Formulation and Signal Model

We analyze an ad hoc WASN accommodating Q speakers and $k = [1, \ldots, K]$ devices. Each device k comprises a uniform linear array (ULA) equipped with M_k microphone sensing elements indexed by $m = [1, \ldots, M_k]$. In our setup, we assume an identical number of microphones at every active node k. The overall number of microphones throughout the network is $M = \sum_{k=1}^{K} M_k$. Fig. 1.2 sketches the studied audio scenario. A speaker q generates signals $\tilde{s}_q[\eta]$, $\eta = [1, \ldots, T]$, where η denotes the sample time index. Let $\tilde{\mathbf{s}}_q$ describe the column vector of all emitted signals from source q, i.e. $\tilde{s}_q[\eta]$, at every time instant $\eta = [1, \ldots, T]$. The speech sources $[\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_Q]^\top$ are mutually independent and uniquely labeled using the algorithm presented in [34]. We assume statistical second-order stationarity for blocks of length L and define the instantaneous power of a signal $\tilde{s}_q[\eta]$ at each block as

$$s_q[n] = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{s}_q[nL + l]^2, \tag{2.1}$$

where $n = [1, \ldots, N]$ is the frame index. The $s_q[n]$ are stacked in a Q-dimensional vector $\mathbf{s}[n]$. The instantaneous noisy signal power at the mth microphone of the kth device is

$$y_{k,m}[n] = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{y}_{k,m}[nL + l]^2, \quad m \in \{1, \ldots, M_k\}, \tag{2.2}$$

where $\tilde{y}_{k,m}$ denotes the observed signal at the mth microphone of the kth device. By assuming mutually independent source signals and neglecting the reverberation effects over time segments [4], in a centralized network, the system-wide non-negative $y_{k,m}[n]$

of all devices k are stacked in a M-dimensional vector $\mathbf{y}[n]$. The mixture is modeled by

$$\mathbf{y}[n] \approx \mathbf{A}\mathbf{s}[n] + \boldsymbol{\omega}[n], \quad n = 1, \ldots, N, \tag{2.3}$$

with

$$\mathbf{y}[n] \triangleq \left[(\mathbf{y}_1[n])^\top, \ldots, (\mathbf{y}_k[n])^\top, \ldots, (\mathbf{y}_K[n])^\top\right]^\top \tag{2.4}$$

$$\mathbf{y}_k[n] \triangleq [y_{k,1}[n], \ldots, y_{k,M_k}[n]]^\top \tag{2.5}$$

$$\mathbf{s}[n] \triangleq [s_1[n], \ldots, s_Q[n]]^\top, \tag{2.6}$$

where $\mathbf{A} \in \mathbb{R}^{M \times Q}$ is the mixing matrix that describes the power attenuation between speaker q and microphone m. The additive noise term $\boldsymbol{\omega}[n]$ follows the same design introduced in Eqs. (2.1)-(2.2). In the centralized setup, as in [4], the instantaneous linear mixtures in Eq. (2.3) allow for the estimation of the unknown signal powers $\mathbf{s}[n]$.

2.2 Overview of the Proposed DM-VAD Scheme

In order to elucidate the framework that the proposed DM-VAD undergoes, the steps of the overall DM-VAD mechanism are illustrated in Fig. 2.2. The figure provides an overview of the proposed DM-VAD algorithm, which consists of three main steps

(1) Locating Nodes around sources (LONAS) [1],

(2) distributed source-specific energy signal unmixing,

(3) energy signal based voice activity detection.

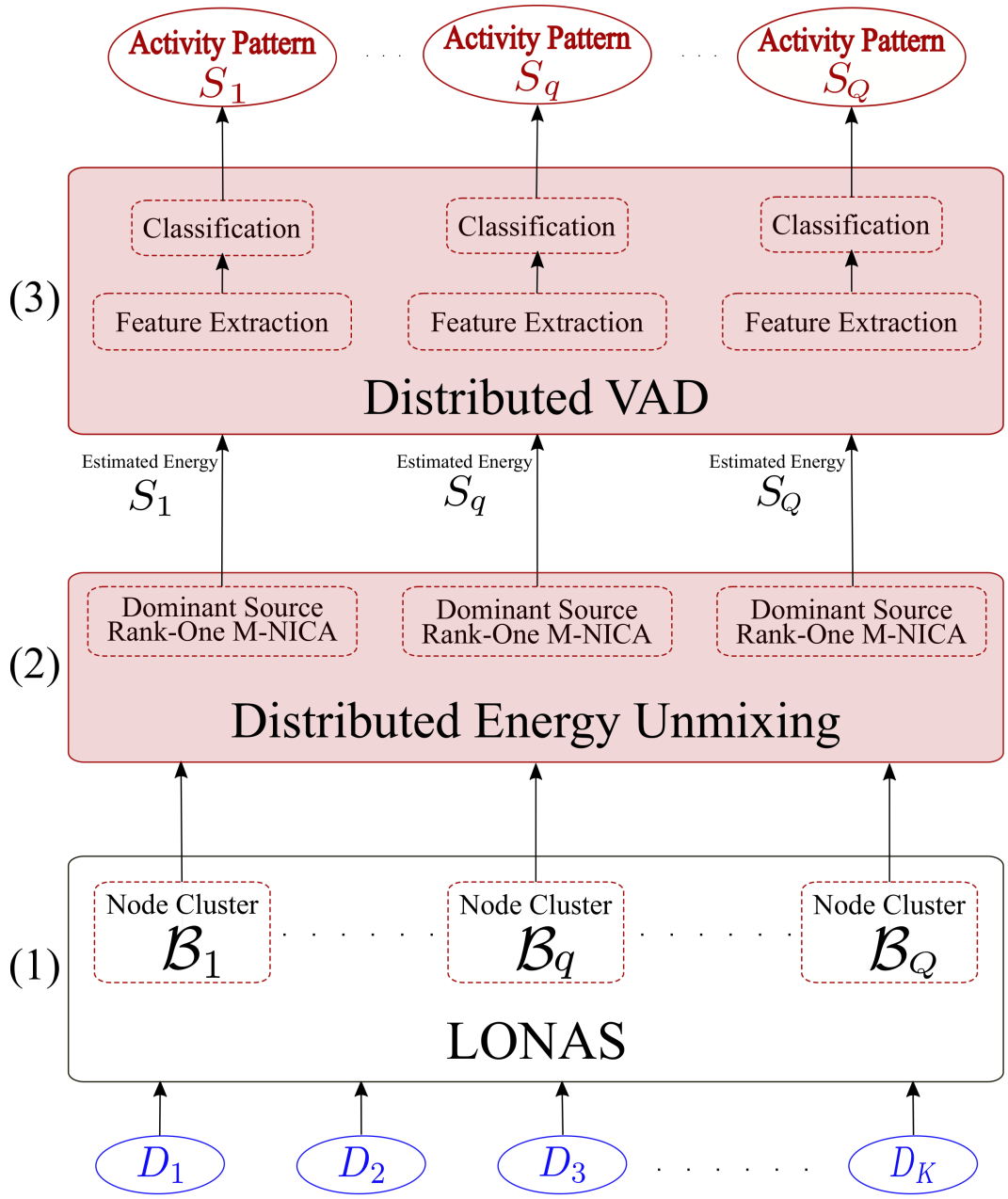


Figure 2.2: Block-diagram of the proposed DM-VAD framework. The input signals are energy mixtures received at every device $D_k \in \{D_1, \ldots, D_K\}$ and the output of the proposed system are VAD patterns relative to the energy sources $S_q \in \{S_1, \ldots, S_Q\}$.

Node clustering around their unique dominant source-of-interest is performed using the LONAS method [1], which is presented in Section 2.3.2. LONAS is able to identify Q node clusters $\mathcal{B}_q, q = [1, \ldots, Q]$, which are composed such that $\mathcal{B}_q$ observes source q as the dominant speech source. This technique endures a distributed source enumeration method to obtain an estimate of Q, which we denote as $\hat{Q}$. Once a distributed node clustering is created using LONAS, a consecutive distributed source energy un-

mixing is enabled at every cluster dominated by a single energy signature related to a well-labeled speech source. This implies that each cluster applies independently a rank-one M-NICA to separate its dominant source's energy from the remaining signal and noise content. In fact, the computationally efficient blind source separation (BSS) technique, namely M-NICA [85], rapidly loses its performance when the number of targeted sources increases in a centralized scenario with a FC. A distributed extraction of dominant source-specific energy signals from the mixed observations hugely outperforms the estimation performance of the existing centralized M-NICA approach [1,5]. It is to note that estimating the energy signal of a single-source using the observations of the nodes around it is a much easier task for M-NICA compared to estimating Q energy signals simultaneously given the observations of all the nodes in the WASN. The scalability issue for large Q and K in the M-NICA method is solved by the divide-and-conquer strategy, namely LONAS [1]. Finally, to determine voice activity, partitional clustering algorithms are applied for which low-dimensional features are extracted from the unmixed source energies to distinguish the pause from the active speech frames for each source separately. At this stage, we propose a robust weighted partitional algorithm that has higher clustering accuracy, mainly for speech energy signatures. Exhaustive simulations are performed on consecutive real world multi-variate speech data to show the efficiency of the proposed distributed multi-speaker VAD framework.

2.2.1 Original Contributions in This Chapter

We introduce a novel clustering-based DM-VAD method for distributed WASNs. Our original approach to the DM-VAD includes:

- Contributing to and proposing a DM-VAD framework for WASNs.
- Designing and evaluating a voice activity detector based on rank-one M-NICA energy unmixing using the dominant source model, see Section 2.3.3.
- Proposing activity detection via a clustering approach based on energy features, see Section 2.4.

2.3 Distributed Unmixing of Source Energy Signals

In this section, we first present fundamentals of the centralized M-NICA algorithm based on [4, 85]. We find it essential explaining the centralized M-NICA since our proposed techniques all through this thesis are based on energy features extracted from

different improved versions of M-NICA. Centralized unmixing of the energies based on Eq. (2.3) can be performed, for example, using non-negative principal component analysis (NPCA) [84] and multiplicative non-negative independent component analysis (M-NICA) [4, 85]. We then briefly describe LONAS [1]. Finally, the proposed distributed source dominant rank-one M-NICA approach is explained in Section 2.3.3.

2.3.1 The Centralized M-NICA Algorithm

A long-standing research question is how to find a suitable representation of data. Independent component analysis (ICA) seeks essential structures in statistical data. The identified data components are both statistically independent and non-Gaussian [92–97]. The separation of the observed mixed energies turns into a NBSS that can be solved using non-negative ICA (NICA) methods.

For a self-contained thesis, we introduce in the following the basic M-NICA algorithm. M-NICA is proposed in [85] with the aim to reconstruct the original non-negative signals from the observed linear mixtures based on a multiplicative update rule. A multiplicative update preserves the non-negativity constraint of the signals and does not depend on a user-defined learning rate as opposed to gradient based updates [98].

Let $\bar{\mathbf{Y}}$ denote the $M \times N$ non-negative matrix of all received noisy energies at every microphone m and frame n collected from Eq. (2.4). The goal of the centralized M-NICA is to find a $Q \times N$ non-negative matrix $\bar{\mathbf{S}}$ such that the rows of the recovered matrix $\bar{\mathbf{S}}$ are uncorrelated and only contain non-negative numbers. The centralized M-NICA is a fixed-point type algorithm that is used to generate this matrix. The centralized M-NICA pre-processes the energy separation step by a singular value decomposition (SVD) step in Eq. (2.7). The SVD decomposes the energy signal $\bar{\mathbf{Y}}$ into the left $\mathbf{U} \in \mathbb{R}^{M \times M}$ and right $\mathbf{V} \in \mathbb{R}^{N \times M}$ right rotations of singular vectors, and a scaling matrix $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ of M singular values on its diagonal. This step is meant to substitute the matrix $\bar{\mathbf{Y}}$ by its smoothed best rank approximation via an SVD operation as shown in Eq. (2.8). This means only the largest $Q < M$ singular values are considered in the computation of the new filtered matrix $\bar{\mathbf{Y}}$. More specifically, the product in Eq. (2.8) generates a smoothed matrix $\bar{\mathbf{Y}} \in \mathbb{R}^{M \times N}$ that is formed from the multiplication of the matrices $\bar{\mathbf{U}} \in \mathbb{R}^{M \times Q}$, $\bar{\mathbf{V}} \in \mathbb{R}^{N \times Q}$, and $\bar{\mathbf{\Sigma}} \in \mathbb{R}^{Q \times Q}$. The remaining singular values, i.e. $Q+1, \ldots, M$ are regarded as noise. It is to notice that this step is capable of removing some noise from the observations. This initialization step is pursued by a centralized signal decorrelation step using Eq. (2.9) of Algorithm 1, derived in [85]. In Eq. (2.9), the elements of the matrix $\bar{\mathbf{S}}$ are updated to decrease the mutual correlation between

the rows of $\bar{\mathbf{S}}$. Due to the fact that $\bar{\mathbf{S}}$ is initialized with non-negative elements, the decorrelation process in Eq. (2.9) will preserve the non-negativity because of its multiplicative nature. However, the rows of the constructed matrix $\bar{\mathbf{S}}$ are no longer in the signal subspace defined by the rows of $\bar{\mathbf{Y}}$. Consequently, the matrix $\bar{\mathbf{S}}$ is projected to the row space of $\bar{\mathbf{Y}}$ using Eq. (2.14). A summary of the centralized M-NICA procedure is given in Algorithm 1 and in [85].

Algorithm 1 Centralized M-NICA [4, 85]

Input

1: $\bar{\mathbf{Y}} = (\mathbf{y}[1], \cdots, \mathbf{y}[N]) \in \mathbb{R}_+^{M \times N}$ based on Eq. (2.4)

Initialization

2: $\forall q = 1, \ldots, Q$, $\forall n = 1, \ldots, N$: $[\bar{\mathbf{S}}]_{qn} \longleftarrow [\bar{\mathbf{Y}}]_{qn}$

3: Replace $\bar{\mathbf{Y}}$ by its best rank approximation by means of the singular value decomposition (SVD), i.e.

4:
$$\{\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}\} \longleftarrow \mathrm{SVD}(\bar{\mathbf{Y}}) \tag{2.7}$$

5:
$$\bar{\mathbf{Y}} \longleftarrow \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^\top \tag{2.8}$$

where $\bar{\mathbf{\Sigma}}$ is the $Q \times Q$ diagonal matrix containing the Q largest singular values of $\bar{\mathbf{Y}}$ on its diagonal, and the corresponding left and right singular vectors are stored in the columns of $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$, respectively.

6: **Decorrelation Step**

$\forall q = 1, \ldots, Q$ $\forall n = 1, \ldots, N$

$$[\bar{\mathbf{S}}^*]_{q,n} \leftarrow [\bar{\mathbf{S}}]_{q,n} \left[\frac{\grave{\mathbf{S}}_q \bar{\mathbf{S}}_q^\top \mathbf{\Lambda}_1^{-1} \bar{\mathbf{S}}_q + \bar{\mathbf{S}}_q \bar{\mathbf{S}}_q^\top \mathbf{\Lambda}_1^{-1} \grave{\mathbf{S}}_q + \mathbf{\Lambda}_2 \bar{\mathbf{S}}_q}{\grave{\mathbf{S}}_q \bar{\mathbf{S}}_q^\top \mathbf{\Lambda}_1^{-1} \grave{\mathbf{S}}_q + \bar{\mathbf{S}}_q \bar{\mathbf{S}}_q^\top \mathbf{\Lambda}_1^{-1} \bar{\mathbf{S}}_q + \mathbf{\Lambda_2} \grave{\mathbf{S}}_q} \right]_{q,n} \tag{2.9}$$

$$\grave{\mathbf{S}} = \frac{1}{N} \{\bar{\mathbf{S}}_n\} \mathbf{1}_N^\top, \forall n = 1, \ldots, N \tag{2.10}$$

$$\mathbf{C}_{\bar{\mathbf{S}}} = (\bar{\mathbf{S}} - \grave{\mathbf{S}})(\bar{\mathbf{S}} - \grave{\mathbf{S}})^\top \tag{2.11}$$

$$\mathbf{\Lambda}_1 = D\{\mathbf{C}_{\bar{\mathbf{S}}}\} \tag{2.12}$$

$$\mathbf{\Lambda}_2 = D\{(\mathbf{\Lambda}_1^{-1} \mathbf{C}_{\bar{\mathbf{S}}})^2\} \tag{2.13}$$

where $\mathbf{1}_N$ denotes an N-dimensional column vector in which each entry is 1, and $D\{\mathbf{X}\}$ denotes the operator that sets all off-diagonal elements of $\mathbf{X}$ to zero.

7: **Signal Subspace Projection Step**

$\forall q = 1, \ldots, Q$, $\forall n = 1, \ldots, N$:

$$[\bar{\mathbf{S}}]_{q,n} \leftarrow \max([\bar{\mathbf{S}}^* \bar{\mathbf{V}} \bar{\mathbf{V}}^\top]_{q,n}, 0) \tag{2.14}$$

8: Return to Step 6.

Once a fixed point of Eqs. (2.9)-(2.14) is achieved, the elements in each row of the non-negative matrix $\bar{\mathbf{S}}$ correspond to frames of the unmixed signal $s_q[n]$. The mixing matrix $\hat{\mathbf{A}}$ related to a row of $\bar{\mathbf{S}}$ can then be calculated using

$$\hat{\mathbf{A}} = \bar{\mathbf{Y}}\bar{\mathbf{S}}^{\top}(\bar{\mathbf{S}}\bar{\mathbf{S}}^{\top})^{-1}. \tag{2.15}$$

By applying the centralized M-NICA algorithm, there remains always a permutation and scaling ambiguity between the columns of $\hat{\mathbf{A}}$ and the estimated energy source signals in $\bar{\mathbf{S}}$. Since we are interested in solving the multi-speaker VAD problem based on well-labeled energy signatures in $\bar{\mathbf{S}}$, the permutation problem of the source signals resulting from the centralized M-NICA algorithm is solved using a distributed labeling algorithm that we propose in [34]. The labeling method is able to assign with high accuracy the energy signatures to their corresponding speech sources in the WASN. In this manner, the resulting energies in the network, and thus the speakers, are well identified. Multi-source energy labeling, or speaker identification based on energy signatures, is discussed in detail in Chapter 4.

2.3.1.1 Computational Cost of the Centralized M-NICA

The complexity of the centralized M-NICA algorithm is similar to that of the non-negative principal component analysis (NPCA). Notice that the M-NICA converges slowly compared to NPCA, but achieves more accurate unmixing results. Especially, when using small sample sizes M-NICA outperforms NPCA significantly. In addition, M-NICA does not rely on a user-defined step size parameter, as opposed to NPCA. This data-driven parameter should be tuned by the user to ensure the convergence of NPCA, and thus obtain relevant unmixing results. Having a number of samples greater than the number of targets, i.e. $N \gg Q$, the overall complexity of the M-NICA algorithm is $\mathcal{O}(Q^2N)$, which is the same as the NPCA algorithm [85].

2.3.2 Locating Nodes Around Sources (LONAS)

Node clustering around their unique dominant source-of-interest is performed using the LONAS method [1]. LONAS allows to identify Q node clusters $\mathcal{B}_q, q = [1, \ldots, Q]$, which are composed such that $\mathcal{B}_q$ observes source q as the dominant speech source. To achieve a node clustering, LONAS applies a distributed eigenvalue decomposition

(EVD). Details on LONAS and a method to estimate the numbers of sources in the network are given in [1] and the outcoming node clustering results are utilized within this research. The results from applying the LONAS to cluster the nodes around their dominant sources of interest in the WASN described by Fig. 1.2 are summarized in [1] and shown in Fig. 2.3.

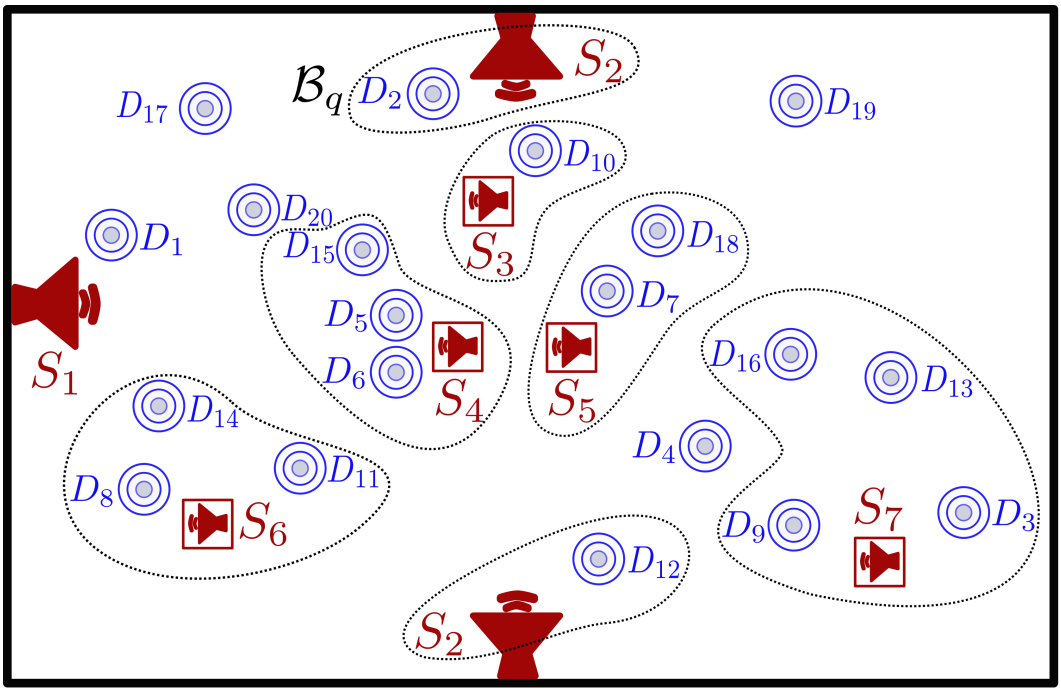


Figure 2.3: Results of the distributed clustering of nodes around their unique dominant sources of interest using LONAS [1] in a WASN of $Q = 7$ speech sources (red) and $K = 20$ devices (blue). Clusters of nodes, i.e., $\mathcal{B}_q$ are represented with black dashed lines for every source q.

2.3.3 Proposed Distributed Rank-One M-NICA for Cluster Dominant Source Estimation

Since different speakers have different positions, the design of a non-negative BSS, such as M-NICA, can rely on spatial information collected by multiple microphones. With a growing number of interfering energy sources in the WASN, the mixture of energies recorded at different microphones is markedly affected by interference. The noise signal can be environmental noise and/or a speaker that interferes with the target speaker (the microphone's source-of-interest). Therefore, a centralized blind source separation algorithm of type M-NICA becomes unreliable in scenarios with higher number of speakers. This is because crucial assumptions, such as the well-groundedness [4, 85]

are violated when many sources are active. In order to overcome the restraints of the centralized M-NICA, we derive a distributed rank-one M-NICA based on distributed node clustering achieved by LONAS. Let $\mathcal{B}_q$ denote the set of nodes that are assigned to the qth source by LONAS, and $\#(\mathcal{B}_q) > 0$ denotes its cardinality, i.e., the number of nodes assigned to the qth source. Further, analogously to Eq. (2.4), let $\mathbf{y}_{\mathcal{B}_q}[n] \in \mathbb{R}^{\#(\mathcal{B}_q)\times 1}$ contain the instantaneous energies of the microphone signals of all nodes $k \in \mathcal{B}_q$ at time segment n. Then, assuming that $s_q[n]$ is the dominant source for the nodes in $\mathcal{B}_q$ we define

$$\mathbf{y}_{\mathcal{B}_q}[n] \approx \mathbf{a}_{\mathcal{B}_q} s_q[n] + \boldsymbol{\omega}_{\mathcal{B}_q}[n], \quad q \in \{1, \ldots, \hat{Q}\}, \tag{2.16}$$

where $\mathbf{a}_{\mathcal{B}_q}$ is a $\#(\mathcal{B}_q)$-dimensional mixing vector that describes the power attenuation between the qth source and the nodes within $\mathcal{B}_q$. Based on Eq. (2.16), each cluster $\mathcal{B}_q, q \in \{1, \ldots, \hat{Q}\}$ uses a source-specific rank-one M-NICA algorithm to determine s_q. A rank-one M-NICA algorithm applied on distributed clusters is capable of recovering a single dominating energy signal q at every cluster of nodes $\mathcal{B}_q$ based on the integration of the cooperative information collected from the set of elementary microphones at a cluster level $\mathcal{B}_q$. Instead of a single centralized M-NICA that assumes Q sources, this implies that, the proposed distributed unmixing approach based on LONAS performs Q rank-one M-NICA algorithms for the Q formed clusters of nodes. For this reason, the performance of the source energy recovery based on the proposed rank-one M-NICA, performed in each cluster, no longer depends on Q, and in principle, Q can grow arbitrarily large.

For the sake of an easy exposition, the ensuing section is concerned with the substantial derivations carried out in order to map the design of a centralized M-NICA into a distributed rank-one M-NICA for a distributed energy-based BSS. Hence, assume that we collect a $\#(\mathcal{B}_q) \times N$ data matrix $\bar{\mathbf{Y}}_{\mathcal{B}_q}$, at cluster $\mathcal{B}_q$, that contains N samples $\mathbf{y}_{\mathcal{B}_q}[n], n = [1, \ldots, N]$, in its columns based on Eq. (2.4). In a distributed M-NICA, we aim at finding a single $1 \times N$ vector $\mathbf{s}_q$ per cluster, such that $\mathbf{s}_q = (s_q[1], \ldots, s_q[N])$, and $\mathbf{s}_q$ contains non-negative values and corresponds to the refined extracted energy for a dominant source q in a cluster $\mathcal{B}_q$. Thus, the derived steps outlined in Algorithm 2 corresponding to the distributed fixed-point rank-one type M-NICA algorithm are used to generate such a vector related to a source-of-interest at every cluster q. In Algorithm 2, the initialization of the vector $\mathbf{s}_q$ is given in Step 2. As a next step, the SVD-based pre-processing of the input matrix $\bar{\mathbf{Y}}_{\mathcal{B}_q}$ is established to replace $\bar{\mathbf{Y}}_{\mathcal{B}_q}$ with its best first rank approximation of the same size $\mathbb{R}^{\#(\mathcal{B}_q)\times N}$. In this case, the left and right

rotations, i.e., $\mathbf{U} \in \mathbb{R}^{\#(\mathcal{B}_q)\times\#(\mathcal{B}_q)}$ and $\mathbf{V} \in \mathbb{R}^{N\times\#(\mathcal{B}_q)}$, as well as, the scaling matrix $\mathbf{\Sigma} \in \mathbb{R}^{\#(\mathcal{B}_q)\times\#(\mathcal{B}_q)}$ are extracted in Eq. (2.18). Afterwards, Eq. (2.19) substitutes $\bar{\mathbf{Y}}_{\mathcal{B}_q}$ by its first rank approximated smoothed version matrix of size $\#(\mathcal{B}_q) \times N$. The decorrelation step for the distributed rank-one M-NICA comes with modifications imposed on the numerator and the denominator compared to Eq. (2.9). In particular, the computation of the covariance matrix $\mathbf{C}_{\bar{\mathbf{S}}} \in \mathbb{R}^{Q\times Q}$ in Eq. (2.11) reduces to computing the variance $c_{\mathbf{s}_q}$ of the signal $\mathbf{s}_q$ in Eq. (2.22). This entails that the diagonal matrices $\mathbf{\Lambda}_1 \in \mathbb{R}^{Q\times Q}$ and $\mathbf{\Lambda}_2 \in \mathbb{R}^{Q\times Q}$ used to derive Eq. (2.9) reduce, in the proposed distributed rank-one M-NICA, to $\grave{\lambda}_1$ and $\grave{\lambda}_2$ corresponding to weighting scalars equal to the computed variance of the signal $\mathbf{s}_q$ and the value one, respectively. Consequently, the proposed decorrelation function for the case of a distributed rank-one M-NICA algorithm is given in Eq. (2.20) of Algorithm 2. Since the term $\grave{\lambda}_2$ reduces to one, the decorrelation formulation in Eq. (2.20) can be written as

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\mathbf{s}_q + \mathbf{s}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\grave{\mathbf{s}}_q + \mathbf{s}_q}{\grave{\mathbf{s}}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\grave{\mathbf{s}}_q + \mathbf{s}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\mathbf{s}_q + \grave{\mathbf{s}}_q}\right]_n \tag{2.17}$$

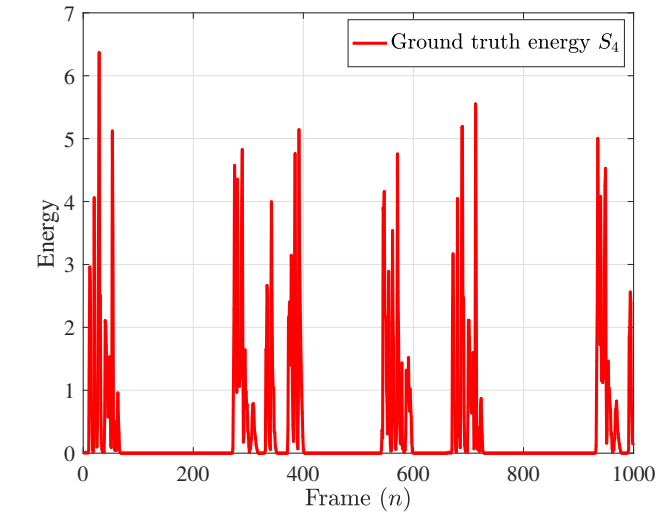
The signal subspace projection for the distributed M-NICA is performed and introduced in Eq. (2.25). Equations (2.20)-(2.25) are repeatedly implemented until a fixed point of the algorithm is achieved. Likewise, the mixing vector $\mathbf{a}_{\mathcal{B}_q} \in \mathbb{R}^{\#(\mathcal{B}_q)\times 1}$ is estimated using

$$\mathbf{a}_{\mathcal{B}_q} = \bar{\mathbf{Y}}_{\mathcal{B}_q}\mathbf{s}_q^\top(\mathbf{s}_q\mathbf{s}_q^\top)^{-1}. \tag{2.26}$$

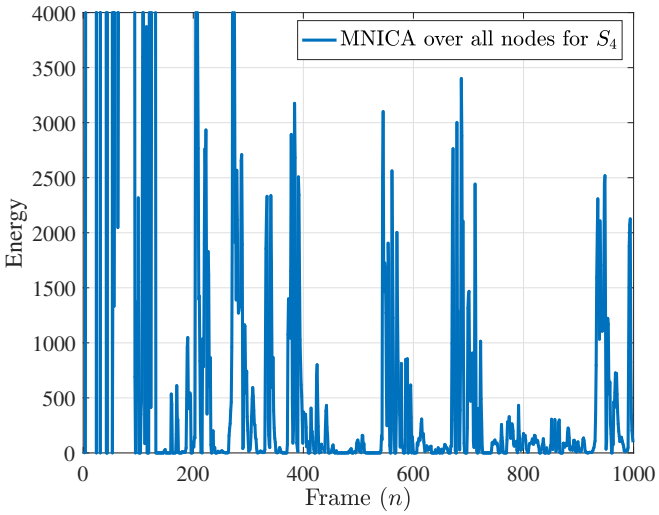
Clearly, the distributed rank-one M-NICA algorithm for the estimation of the signals $\mathbf{s}_q$ separately at every cluster does not issue a permutation problem. Obviously, the recovered unique signal $\mathbf{s}_q$ per cluster q corresponds to the dominant source at that cluster q. A summary of the derived distributed rank-one M-NICA approach is given in Algorithm 2. Figures. 2.4 and 2.5 show the unmixing results when applying a distributed rank-one M-NICA that uses node clusters around dominant sources, compared to the energy unmixing of a centralized M-NICA over all nodes for Sources S_4 and S_5, respectively.

2.3.3.1 Computational Cost of the Distributed Rank-one M-NICA

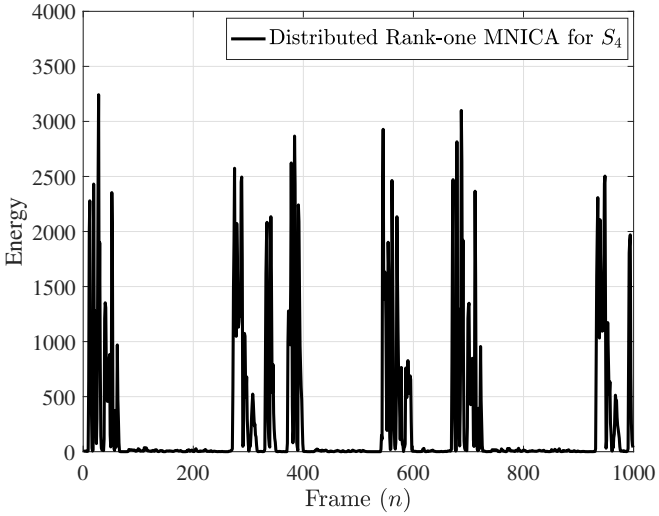
The above mentioned divide-and-conquer approach, related to deriving a distributed unmixing based on the M-NICA algorithm (see Algorithm 2), reduces to a problem



(a)

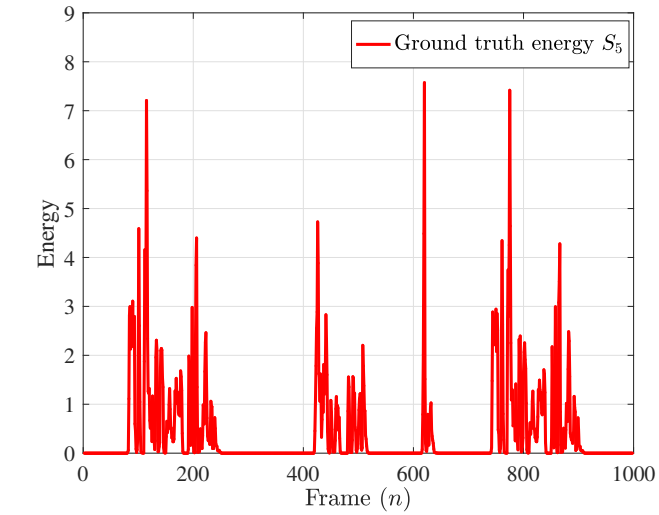


(b)

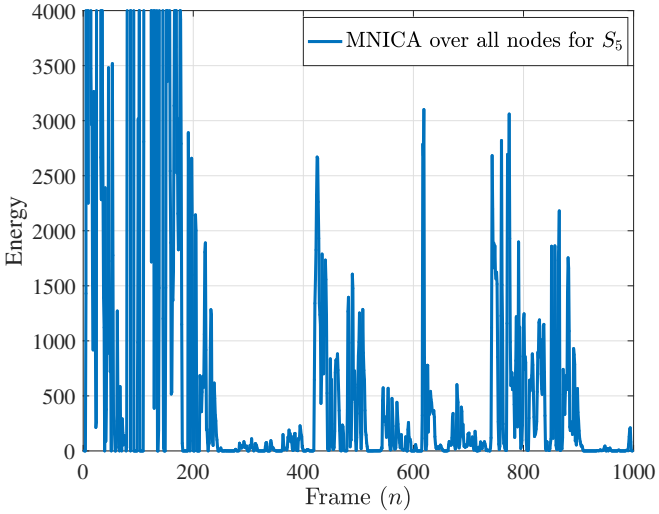


(c)

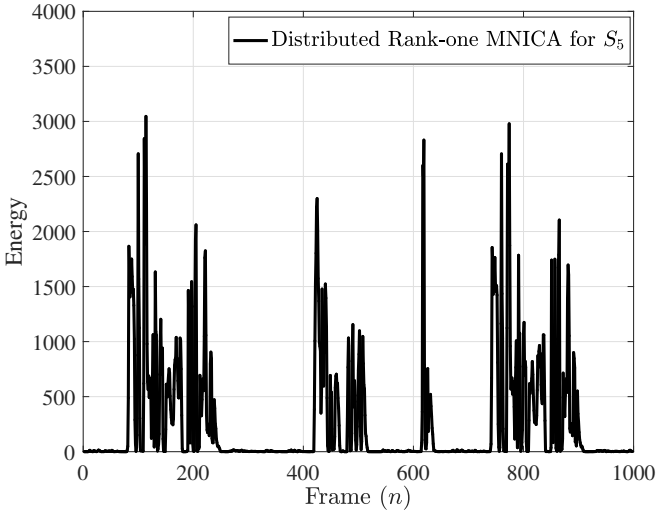
Figure 2.4: The unmixing results for Source S_4 using (b) M-NICA over all nodes and (c) Distributed source dominant rank-one M-NICA.



(a)



(b)



(c)

Figure 2.5: The unmixing results for Source S_5 using (b) M-NICA over all nodes and (c) Distributed source dominant rank-one M-NICA.

Algorithm 2 Distributed rank-one M-NICA

1: **Input**
$\bar{\mathbf{Y}}_{\mathcal{B}_q} = (\mathbf{y}_{\mathcal{B}_q}[1], \cdots, \mathbf{y}_{\mathcal{B}_q}[N]) \in \mathbb{R}_+^{\#(\mathcal{B}_q)\times N}$ based on Eq. (2.16)

2: **Initialization**
For a unique dominant source q, $\mathbf{s}_q \longleftarrow [\bar{\mathbf{Y}}_{\mathcal{B}_q}]_{1n}\ \forall n = 1, \ldots, N$
Replace $\bar{\mathbf{Y}}_{\mathcal{B}_q}$ by its best first rank approximation by means of the singular value decomposition (SVD), i.e.

$$\{\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}\} \longleftarrow \mathrm{SVD}(\bar{\mathbf{Y}}_{\mathcal{B}_q}) \tag{2.18}$$

$$\bar{\mathbf{Y}}_{\mathcal{B}_q} \longleftarrow \bar{\mathbf{u}}\bar{\sigma}\bar{\mathbf{v}}^\top \tag{2.19}$$

where $\bar{\sigma}$ is the largest singular value of $\bar{\mathbf{Y}}_{\mathcal{B}_q}$, $\bar{\mathbf{u}} \in \mathbb{R}^{\#(\mathcal{B}_q)\times 1}$ corresponds to the left singular vector of length $\#(\mathcal{B}_q)$ stored in $\mathbf{U} \in \mathbb{R}^{\#(\mathcal{B}_q)\times\#(\mathcal{B}_q)}$, and $\bar{\mathbf{v}}^\top$ is the right singular vector of $\mathbf{V} \in \mathbb{R}^{N\times\#(\mathcal{B}_q)}$ with $\bar{\mathbf{v}} \in \mathbb{R}^{N\times 1}$.

3: **Decorrelation Step**
$\forall n = 1, \ldots, N$

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\mathbf{s}_q + \mathbf{s}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\grave{\mathbf{s}}_q + \grave{\lambda}_2\mathbf{s}_q}{\grave{\mathbf{s}}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\grave{\mathbf{s}}_q + \mathbf{s}_q\mathbf{s}_q^\top\grave{\lambda}_1^{-1}\mathbf{s}_q + \grave{\lambda}_2\grave{\mathbf{s}}_q}\right]_n \tag{2.20}$$

$$\grave{\mathbf{s}}_q = \frac{1}{N}\{\mathbf{s}_q\}\mathbf{1}_N^\top \tag{2.21}$$

$$c_{\mathbf{s}_q} = (\mathbf{s}_q - \grave{\mathbf{s}}_q)(\mathbf{s}_q - \grave{\mathbf{s}}_q)^\top \tag{2.22}$$

$$\grave{\lambda}_1 = c_{\mathbf{s}_q} \tag{2.23}$$

$$\grave{\lambda}_2 = (\grave{\lambda}_1^{-1}c_{\mathbf{s}_q})^2 = 1 \tag{2.24}$$

4: **Signal Subspace Projection Step**
For a given source q, $\forall n = 1, \ldots, N$:

$$[\mathbf{s}_q]_n \leftarrow \max([\mathbf{s}_q^*\bar{\mathbf{v}}\bar{\mathbf{v}}^\top]_n, 0) \tag{2.25}$$

5: Return to Step 3

with linear complexity. Specifically, the steps of the distributed M-NICA are applied in parallel to every identified cluster of nodes in order to estimate a unique energy source. This means, the computational cost of the distributed algorithm is independent of the overall number of participating sources, but only depends on the number of observations. At every cluster, we assume $N \gg 1$. The value one here pertains to the unique source-of-interest q of the formed cluster $\mathcal{B}_q$. Based on this, we infer that the complexity of the distributed M-NICA at a cluster $\mathcal{B}_q$ increases linearly with N, i.e., $\mathcal{O}(N)$. The implementation of the distributed M-NICA is parallelized, as to recover

the Q energy sources, without any additional time complexity.

2.4 Distributed Clustering-Based Multi-Speaker Voice Activity Detection (DM-VAD)

The final step of the proposed algorithm distinguishes the active and the non-active speech segments for each source by means of efficient partitional clustering algorithms [99, 100]. These algorithms determine the class membership of each time segment, depending on its distance to the estimated cluster centroids. The idea of our approach is to transform the VAD problem into a clustering task by extracting features from the estimated energy signatures. There are two clusters that correspond to the active and non-active speech clusters. Unique labels of the speech sources throughout the network are available from the distributed labeling algorithm presented in [34] and Chapter 4 of this thesis.

Rational decision is involved to answer the question: "Which center corresponds to which speech case?". Apparently, taking the minimum center to be a descriptive measure of the non-active speech and vice versa seems to be a practical assumption for our speech discrimination purpose. We utilize this straightforward assumption all along the upcoming proposed approach for voice activity decision. Running a K-means type algorithm in the case of known two-centers based clustering enables us to estimate two centers, one related to the active speech while the other describes the non-active speech.

Based on the source energies extracted using the proposed distributed framework presented earlier, the forthcoming subsections consist of

1. Robust low-dimensional short-term energy feature extraction
2. Non-stationary speech discrimination based on two proposed robust objective functions for clustering.

2.4.1 Robust Low-Dimensional Short-Term Energy Features

Let $\hat{s}^{(n)}_{\mathcal{B}_q}, \quad q = [1, \ldots, \hat{Q}]$ denote the estimated source-specific energy signals $s_q[n]$ in Eq. (2.25). Source-specific voice activity patterns for each source $q = [1, \ldots, \hat{Q}]$ are

determined by extracting features from $\hat{s}_{\mathcal{B}_q}^{(n)}$ locally within each node cluster $\mathcal{B}_q$, allowing for a distributed computation. The unique labels of $\hat{s}_{\mathcal{B}_q}^{(n)}$ throughout the network are available from the distributed labeling algorithm presented in [34]. The feature vector

$$\boldsymbol{v}_q^{(n)} \triangleq [v_{q,1}^{(n)}, v_{q,2}^{(n)}, v_{q,3}^{(n)}]^\top \tag{2.27}$$

is formed from three different features. The selected features are the result of an empirical study that contained a larger set of features which we do not elaborate on for the sake of conciseness. The selected features are computed as follows

1. **Short-term arithmetic average**

$$v_{q,1}^{(n)} = \frac{1}{W} \sum_{i=n-W}^{n+1} \hat{s}_{\mathcal{B}_q}^{(i)}, \quad n \in \{W+1, \cdots, N\} \tag{2.28}$$

2. **Short-term standard deviation**

$$v_{q,2}^{(n)} = \sqrt{\frac{1}{W} \sum_{i=n-W}^{n} (\hat{s}_{\mathcal{B}_q}^{(i)} - v_{q,1}^{(i)})^2}, \quad n \in \{W+1, \cdots, N\} \tag{2.29}$$

3. **First-order energy difference**

$$v_{q,3}^{(n)} = \hat{s}_{\mathcal{B}_q}^{(n)} - \hat{s}_{\mathcal{B}_q}^{(n+1)}, \quad n \in \{W, \cdots, N-1\} \tag{2.30}$$

Figure 2.6 gives an illustrating example, where each point corresponds to one feature vector $\boldsymbol{v}_q^{(n)}$, which either belongs to the active speech (blue crosses) or the pause class (red dots). From Fig. 2.6, it can be seen that the distribution of the data in the feature space is non-symmetric and non-Gaussian. The speech clusters are highly overlapping and do not generate spherical clusters. Based on these facts, relying on a standard K-means for speech activity detection is unsuitable. Following our previously presented arguments, we propose two alternative robust objective functions for clustering. We also use some existing robust variations of the K-means algorithm for comparison. One such variation is the K-medians, where the component-wise sample median of each cluster is used to determine its centroid. This results in minimizing the error over all clusters with respect to the ℓ_1-norm distance metric. A further robust variation is the K-medoids, which can be used with arbitrary distance metrics, and is based on the

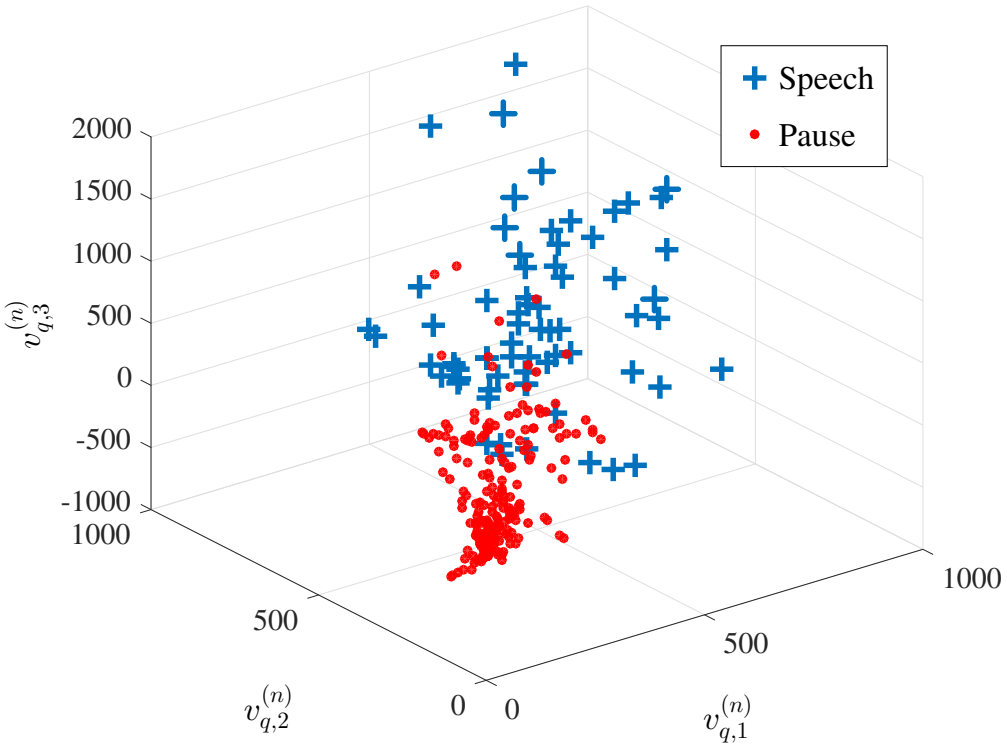


Figure 2.6: Example of the extracted feature vectors for Source S_2.

medoid, which is the instance from the dataset for which the average dissimilarity to all the objects in the cluster is minimal. In the following subsections, we explain the distributed VAD decision based on the original K-means, then we look at the proposed robust objective functions in clustering to perform speech activity discrimination.

2.4.2 K-means Type Clustering Criteria for Distributed Non-Stationary Multiple Speech Discrimination

Cluster analysis is one of the main analytical methods in data mining and signal processing. Clustering discovers reasonable grouping of raw data while searching hidden patterns that may exist in datasets. The data points in the resulting clusters are similar when they belong to the same cluster and differ from a cluster to another. The quality of a cluster is usually measured using the variance. We briefly discuss the standard K-means algorithm [99] and propose a clustering approach that provides an improvement for our considered use-case.

The K-means clustering, commonly named the Lloyd's algorithm, is a greedy algorithm which is guaranteed to converge to a local minimum while minimizing its score

function. K-means is a numerical, unsupervised, non-deterministic, iterative method. Finding optimal K-means centers is an NP-hard problem. Therefore, heuristics are often used. The algorithm consists of two separate phases. The first phase regards selecting the centroids randomly and the second phase concerns assigning each data object to the nearest center. The Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. Once a first clustering of all data points is performed using randomly initialized cluster centroids, the first step is completed. This provides an initial grouping of the data. Next, a re-calculation of the centers of every cluster is done by taking the average of the data within every formed cluster. This iterative process continues repeatedly until reaching a minimum of the criterion function, which is a minimization of the within-cluster sum of squares (WCSS) defined as

$$\arg\min_{\mathbf{C}} \sum_{j=1}^{2} \sum_{\boldsymbol{v}_q^{(n)} \in \mathbf{C}_j} \|\boldsymbol{v}_q^{(n)} - \hat{\mathbf{c}}_j^{(q)}\|_2^2 \tag{2.31}$$

Equation (2.31) describes the objective function of the K-means algorithm that, applied to our speech use-case, aims to partition N observations of the features $\boldsymbol{v}_q^{(n)}$ into $j = 1, 2$ sets $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2\}$. The mean of points in $\mathbf{C}_j$ is the center of the cluster and is denoted with $\hat{\mathbf{c}}_j^{(q)} \triangleq [\hat{c}_{j,1}^{(q)}, \hat{c}_{j,2}^{(q)}, \hat{c}_{j,3}^{(q)}]^\top \in \mathbb{R}^{3\times 1}, \forall j \in \{1, 2\}$. In addition, we use the K-medoids technique [101] to estimate the speech cluster centroids for the active and non-active speech clusters. The K-medoids is a variation of the simple framework given by the K-means algorithm. The modification that it brings to the K-means relies on choosing the actual data points as representative prototypes for the clusters which makes the K-medoids more resilient to noise and outliers in the data compared to K-means. Moreover, for a comparative purpose, we also use the K-medians [102] as an alternative technique for estimating cluster centroids. The K-medians is a robust version of the K-means that relies on the median measure, which is less sensitive to outliers, for the update of the centers instead of the average mean employed by K-means. In the sequel, we subsume the K-means, K-medoids, K-medians and the proposed feature weighted K-MAD algorithms under the category K-means type algorithms.

2.4.2.1 Proposed K-MAD and Weighted K-MAD Clustering Algorithms

We propose a variation of K-means that we call *Feature Weighted* K-MAD. Here, a weighted objective function, in this case, the mean absolute deviation to the mean is minimized using an iterative procedure. The weighting ensures that each feature is

given equal importance, whereas the MAD is less sensitive to outliers compared to the squared Euclidean distance.

The motivation for using a weighted K-MAD (WK-MAD) is that the spread in each dimension of the feature space varies strongly, as shown in Fig. 2.6. From Fig. 2.6, it can be observed that there are no evident partitions in the feature space. A linear distance measure can fail clustering highly non-separated data as it is the case for speech energies. This can occur when the used energy-based features are non-Gaussian distributed and overlapping. Overlapping features imply non-spherical representation of features in the space. Thus, applying a stand-alone linear Euclidean distance for clustering does not provide satisfactory results. A weight function that ensures each feature is given equal weight in the K-MAD procedure is then introduced. Moreover, the proposed WK-MAD algorithm relaxes the assumption of clustering while maintaining comparable cluster variances. This holds only for the case where the clusters are harmoniously spherical, well separated, have similar number of elements and finally similar clusters' volume. Hence, the proposed WK-MAD algorithm aims at minimizing a robust weighted objective function, in this case, the mean absolute deviation to the mean. In the proposed objective function, outliers has less weight compared to a squared Euclidean distance. The WK-MAD clustering method is efficiently computable and sufficiently accurate for the VAD purpose. In the following, an explanation of the working principle of the proposed algorithms is presented.

As initialization, a centroid $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ associated to a source q is chosen randomly at the iteration $\zeta = 0$ as one of the existent data points. Next, we compute for each dimension of $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ separately the Euclidean distance to each of the feature vectors entries. This is done using

$$\|v_{q,f^\star}^{(n)} - \hat{c}_{j,f^\star}^{(q)}\|_2, \quad j = 1, 2, \quad n = 1, \ldots, N, \quad f^\star = 1, \ldots, 3, \tag{2.32}$$

where $f^\star$ represents the features' index. In this way, for each $f^\star = 1, \ldots, 3$, we obtain a matrix of measured distances $\boldsymbol{L}_{f^\star} \in \mathbb{R}^{N \times 2}$ whose entries are given in Eq. (2.32). Next, we compute the respective average distance for each feature $f^\star$

$$\mu_{\boldsymbol{L}_{f^\star}} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{j=1}^{2} \|v_{q,f^\star}^{(n)} - \hat{c}_{j,f^\star}^{(q)}\|_2, \quad f^\star = 1, \ldots, 3. \tag{2.33}$$

The feature related weights are defined as

$$w_{f^\star} = \frac{3/\mu_{\boldsymbol{L}_{f^\star}}}{\sum_{f^\star=1}^{3} 1/\mu_{\boldsymbol{L}_{f^\star}}}, \quad f^\star = 1, \ldots, 3. \tag{2.34}$$

Combining Eq. (2.32) and Eq. (2.34), the weighted Euclidean distance is readily obtained as

$$\sum_{f^\star=1}^{3} w_{f^\star} \| v_{q,f^\star}^{(n)} - \hat{c}_{j,f^\star}^{(q)} \|_2, \quad j = 1, 2, \quad n = 1, \ldots, N. \tag{2.35}$$

In the subsequent iteration, each feature vector $\boldsymbol{v}_q^{(n)}$ is assigned to a cluster $\boldsymbol{C}_1^\zeta$ or $\boldsymbol{C}_2^\zeta$. Once the clusters are formed, the centroid estimates are updated by evaluating

$$\hat{\mathbf{c}}_j^{(q)} = \frac{\sum_{\boldsymbol{v}_q^{(n)} \in \boldsymbol{C}_j^\zeta} \boldsymbol{v}_q^{(n)}}{\#(\boldsymbol{C}_j^\zeta)}, \quad j = 1, 2, \tag{2.36}$$

where $\#(\boldsymbol{C}_j^\zeta)$ is the cardinality of the cluster $\boldsymbol{C}_j^\zeta$. The steps defined in Eqs. (2.32)-(2.36) are then repeated until the centroids do not change their values or an alternative relaxed convergence criterion to a local minimum is met. The objective function, which is minimized by the Feature Weighted K-MAD, is

$$\sum_{j=1}^{2} \sum_{n=1}^{N} \sum_{f^\star=1}^{3} w_{f^\star} \| v_{q,f^\star}^{(n)} - \hat{c}_{j,f^\star}^{(q)} \|_2, \quad j = 1, 2, \quad n = 1, \ldots, N. \tag{2.37}$$

Algorithm 3 summarizes the steps applied in the proposed feature-based K-MAD clustering technique where the minimized objective function is the within-cluster sum defined in Eq. (2.38). In a similar way, Algorithm 4 outlines the necessary steps for the computation of the proposed WK-MAD partitional algorithm.

Let $\hat{\mathbf{c}}_j^{(q)} \triangleq [\hat{c}_{j,1}^{(q)}, \hat{c}_{j,2}^{(q)}, \hat{c}_{j,3}^{(q)}]^\top \in \mathbb{R}^{3\times 1}, \forall j \in \{1, 2\}$ denote the estimated centroids of the K-means type algorithms for source $q = [1, \ldots, Q]$, and let $\hat{\mathbf{c}}_1^{(q)}$ correspond to the pause class, which is easily identified by $\min(\hat{c}_{j,1}^{(q)})$ for $j \in \{1, 2\}$ since the short-term average energy of this class is smaller than for the active speech class, and then $\hat{\mathbf{c}}_2^{(q)}$ corresponds to the active speech class. The cluster memberships are then determined from

Algorithm 3 Proposed K-MAD clustering for Source q.

Initialization

Initial guess for the centers $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$

Classification Phase

repeat

Use the centers to classify the data points $\boldsymbol{v}_q^{(n)}$ into clusters using Eq. (2.32).

Update the centroids $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ using Eq. (2.36)

until $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ do not change, or convergence to a local minimum is met.

Output

Minimized mean absolute deviation-based objective function defined by

$$\arg\min_{\mathbf{C}} \sum_{j=1}^{2} \sum_{\boldsymbol{v}_q^{(n)} \in \mathbf{C}_j} \|\boldsymbol{v}_q^{(n)} - \hat{\mathbf{c}}_j^{(q)}\|_2 \tag{2.38}$$

Estimate of the centroids $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$

Estimated clusters $\mathbf{C}_j, j = \{1, 2\}$.

Algorithm 4 Proposed WK-MAD clustering for Source q.

Initialization

Initial guess for the centers $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$

Classification Phase

repeat

Use the centers to classify the data points $\boldsymbol{v}_q^{(n)}$ into clusters using Eq. (2.32).

Compute the respective average distance for each feature $f^\star$ using Eq. (2.33)

Calculate the weights for every feature $f^\star$ using Eq. (2.34)

Obtain the weighted Euclidean distance by applying Eq. (2.35)

Update the centroids $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ using Eq. (2.36)

until $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ do not change, or convergence to a local minimum is met.

Output

Minimized objective function defined in Eq. (2.37)

Estimate of the centroids $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$

Estimated clusters $\mathbf{C}_j, j = \{1, 2\}$.

$$t_j(\boldsymbol{v}_q^{(n)}) = \|\boldsymbol{v}_q^{(n)} - \hat{\mathbf{c}}_j^{(q)}\|_2^2, \quad n \in \{W+1, \ldots, N\}, \tag{2.39}$$

based on which the binary voice activity decision rule for multiple speech sources in the distributed use-case is formed by

$$\delta_q^{(n)} = \begin{cases} 0 & \text{if } t_1(\boldsymbol{v}_q^{(n)}) < t_2(\boldsymbol{v}_q^{(n)}) \quad \text{(pause)}, \\ 1 & \text{otherwise} \quad \text{(active speech)}. \end{cases} \tag{2.40}$$

2.4.2.2 Computational Complexity of the Proposed K-MAD and WK-MAD

In general, the problem of finding the global optimum of the partitional K-means objective function is NP-hard. However, the standard implementation of K-means only approximates a local optimum of its objective function. At this stage, assuming a fixed number τ of running iterations of the K-means algorithm requires $\mathcal{O}(\tau\#(\mathbf{C})N\varphi)$. Herein, $\#(\mathbf{C}) = 2$ is the number of clusters, and $\varphi = 3$ the dimension of the data. Likewise, the suggested variations, namely K-MAD and WK-MAD are similar in complexity to the standard K-means. Hence, the time complexity of both of these proposed robust algorithms is of $\mathcal{O}(\tau\#(\mathbf{C})N\varphi)$, with $j = \{1, \ldots, \varphi\}$.

2.4.3 Energy Classification-Based Hangover Scheme

The decision rule introduced in Eq. (2.40) might result in the misclassification of some low power data points. For this reason, a hangover scheme is utilized for correction. Because of its practical usefulness, we report on a simple and optional correction step. This step aims at reducing the misdetection rate by reassigning the labels for some low power data points that were falsely assigned to the pause class in $\delta_q^{(n)}$ by the decision rule defined in Eq. (2.40). Given the assignments from Eq. (2.40), let $\mathcal{M}_0$ denote the set of $n \in \{W+1, \ldots, N\}$ for which $\delta_q^{(n)} = 0$ and let $\mathcal{M}_1$ denote the set where $\delta_q^{(n)} = 1$. Then we calculate

$$\hat{\sigma}_0 \triangleq \mathrm{mad}(\{v_{q,1}^{(n)}\}), \quad \forall n \in \mathcal{M}_0 \tag{2.41}$$

and

$$\hat{\sigma}_1 \triangleq \mathrm{mad}(\{v_{q,1}^{(n)}\}), \quad \forall n \in \mathcal{M}_1 \tag{2.42}$$

where $\mathrm{mad}(\mathcal{X})$ is the median absolute deviation of a dataset $\mathcal{X}$, and $v_{q,1}^{(n)}$ refers to the short-term arithmetic average feature presented previously in Eq. (2.28). By defining

$$D_0(n) \triangleq |v_{q,1}^{(n)} - \hat{\sigma}_0| \tag{2.43}$$

and

$$D_1(n) \triangleq |v_{q,1}^{(n)} - \hat{\sigma}_1|, \tag{2.44}$$

the voice activity decision $\delta_q^{(n)}$ may be corrected as follows:

$$\delta_{q,\text{new}}^{(n)} = \begin{cases} 1 & \text{if } D_0(n) > D_1(n), \\ \delta_q^{(n)} & \text{otherwise.} \end{cases} \tag{2.45}$$

The correction step in Eq. (2.45) is intended to solve the misclassification of the attenuated energies that still represent active speech. Figure 2.7 and Fig. 2.8 illustrate the effect of the correction step on the empirical distribution function of the energies $v_{q,1}^{(n)}$ that are associated to the speech class. In the top graph of Fig. 2.7, the histogram based on the assignments of Eq. (2.40) is displayed. The middle graph shows the distribution of speech obtained for the case of the (unavailable) ground truth assignments. The lower graph of Fig. 2.7, on the other hand, depicts the speech distribution after applying the correction step of Eq. (2.45). It is noticed that the speech distribution after the correction step becomes more similar to the one obtained from the ground truth assignments. A shift in the speech distribution mode to the left is observed in the bottom subplot. This is explained by the correct reassignment of elements to the speech distribution. The positive effect of this correction step is also noticed in the real data experiments, see Tabs. 2.1-2.4 and Tab. 2.8.

2.4.4 Batch-Mode DM-VAD Algorithm

The proposed VAD algorithm, which is run locally, e.g., by a unique node at each node cluster $\mathcal{B}_q, q = [1, \ldots, Q]$ can be operated on batches of data (batch-mode VAD), or for streaming data (sequential VAD). The batch mode VAD algorithm is summarized in Algorithm 5.

Algorithm 5 Batch-mode VAD algorithm for Source q evaluated locally within $\mathcal{B}_q$.

Input

Set a window value W

Batch VAD procedure

for $n = W + 1, \ldots, N$ **do**

    Compute the features $\boldsymbol{v}_q^{(n)}$ using Eq. (2.27).

end for

Estimate the centroids $\hat{\mathbf{c}}_j^{(q)}, j = \{1, 2\}$ using
K-means, K-medians, K-medoids, or the proposed K-MAD/WK-MAD.

Label $\min(\hat{c}_{j,1}^{(q)}), j = \{1, 2\}$ and $\max(\hat{c}_{j,1}^{(q)}), j = \{1, 2\}$
as pause and active speech centroids, respectively.

Decide $\forall n \in \{W + 1, \ldots, N\}$ based on Eq. (2.40)

Output

VAD patterns in $\delta_q^{(n)}, \forall n \in \{W + 1, \ldots, N\}$.

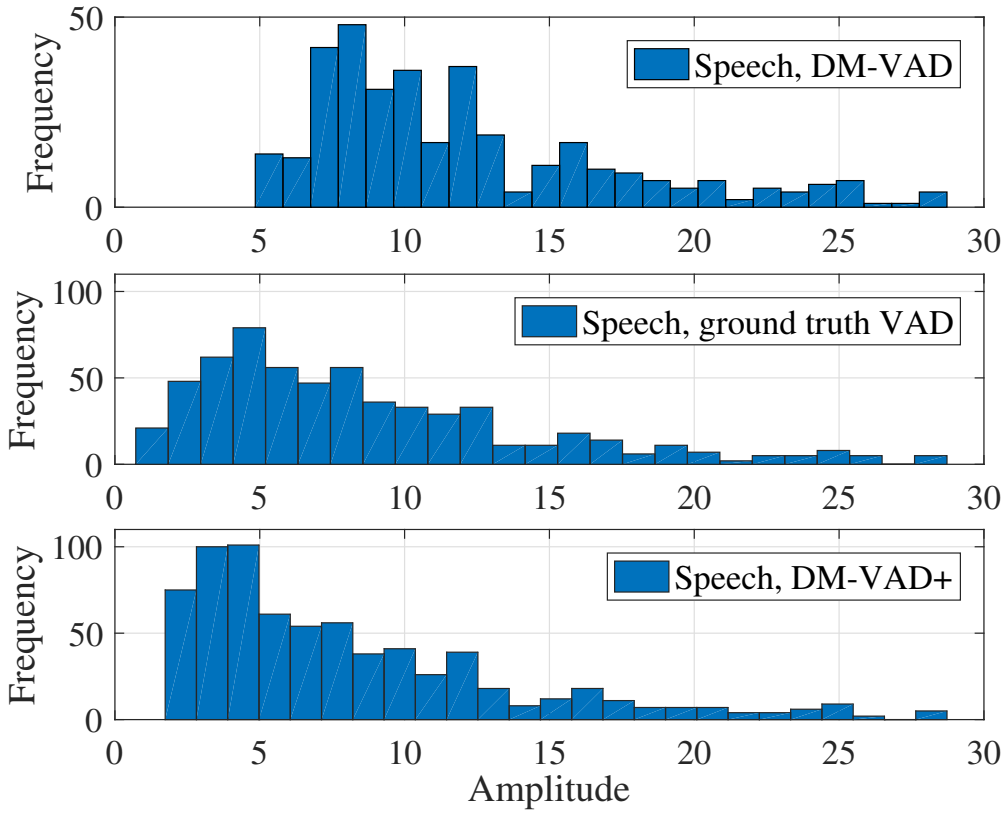


Figure 2.7: Example of the histogram of $v_{q,1}^{(n)}$, $n \in \{W+1,\ldots,N\}$ for Source S_2 for the active speech class using the distributed multi-speaker VAD (DM-VAD) approach before (top), i.e. DM-VAD, and after (bottom), i.e. DM-VAD+, applying the correction step defined in Eq. (2.45), and the ground truth histogram for Source S_2 is shown in the middle.

2.4.5 Sequential-Mode DM-VAD Algorithm

Incremental clustering, as opposed to traditional batch-mode clustering, for VAD has the ability to process new streaming data features without performing a full re-clustering, i.e. a full calculation of the decision pattern. The proposed sequential-mode DM-VAD algorithm allows for a dynamic tracking and incremental decision updates to the database during the clustering procedure. In the sequential DM-VAD algorithm, the VAD decision is made immediately as data streams in.

In the sequential VAD algorithm, the feature vector $\boldsymbol{v}_q^{(n)}$ is calculated sequentially for streaming-in unmixed energy signals $\hat{s}_{\mathcal{B}_q}^{(n)}, n = [W+1,\ldots,N]$ which can be computed with the adaptive M-NICA algorithm as described in [4]. The proposed sequential mode DM-VAD algorithm uses a growing window so as to incorporate all past information. In principle, a sliding window implementation is also possible, however, the window must be chosen large enough so as to capture both active speech and pause segments. The

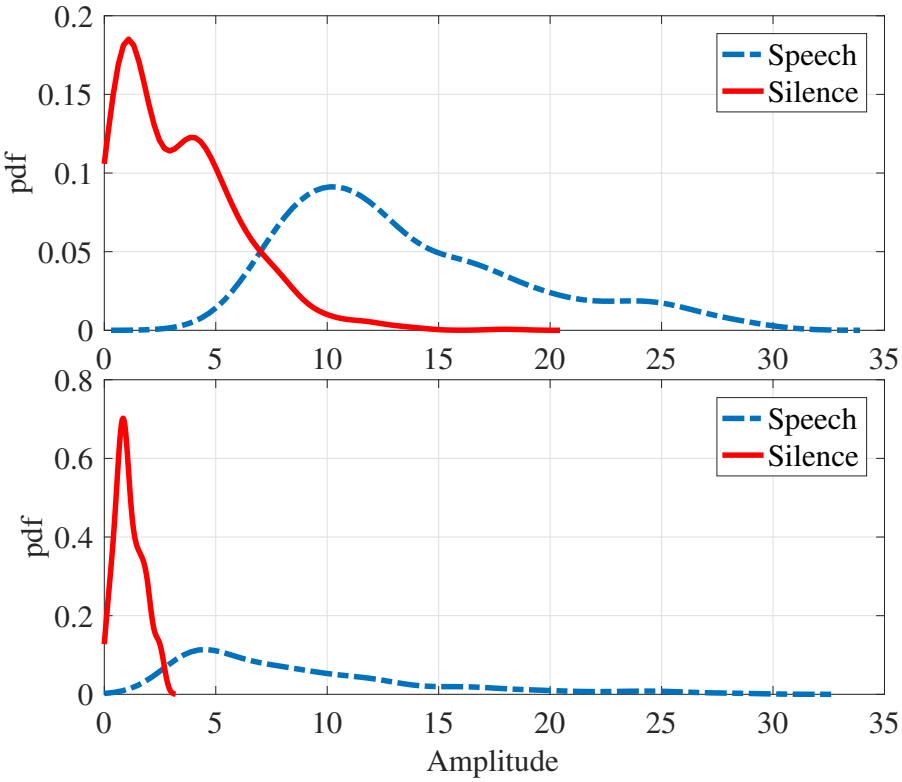


Figure 2.8: Example of the probability density function estimate (pdf) for the active speech region (blue) and the non-active speech region (red) before improving the misdetection rate (top) and after applying the correction step in Eq. (2.45) (bottom).

initial size W^0 of the growing window $W^{(n)}$ at the first iteration $n = W^0 + 1$ should be chosen sufficiently large to reliably extract $\boldsymbol{v}_q^{(n)}$. In this case, the instantaneous feature vectors are obtained by evaluating Eq. (2.28), Eq. (2.29), and Eq. (2.30) for all time segments $\leqslant W^0 + 1$. The features at each time segment n are collected as in Eq. (2.27). All further steps are the same as in the batch mode algorithm, given the available data, except that the random initialization of the centroids in the sequential VAD algorithm is performed only once. Then the sequential VAD uses the previous value of the centroid estimates as initialization. The sequential VAD algorithm is summarized in Algorithm 6.

After operating Algorithm 5 related to the batch-mode clustering-based DM-VAD, the information about the extracted VAD patterns for every energy source q is relayed and shared within and between node clusters. Pertaining to Algorithm 6 that solves the sequential-mode clustering-based DM-VAD problem, the process of sharing the VAD patterns for every source q is accomplished for every real-time voice activity decision.

Algorithm 6 Sequential VAD algorithm for Source q evaluated locally within $\mathcal{B}_q$.

Input

Initialize the window size by $W^{(n)} = W^0$

Randomly pick $\hat{\mathbf{c}}_j^{(q)}, \forall j = 1, 2$ centroids from the data.

Sequential VAD procedure

while $n \leq N$ **do**

  Estimate the centroids $\hat{\mathbf{c}}_j^{(q)}, j = 1, 2$, using
  K-means, K-medians, K-medoids, or the proposed K-MAD/WK-MAD.

  Label $\min(\hat{c}_{j,1}^{(q)}), j = 1, 2$, and $\max(\hat{c}_{j,1}^{(q)}), j = 1, 2$
  as pause and active speech centroids, respectively.

  Decide for $n = N_{\text{current}}$ based on Eq. (2.40)

  $W^{(n)} = W^{(n)} + 1$

end while

Output

Streaming data VAD patterns in $\delta_q^{(n)}, \forall n \in \{W + 1, \ldots, N\}$.

2.5 Detection Simulation Results

In this section, numerical experiments are conducted to assess the performance of our proposed DM-VAD. The system is evaluated and compared to existing benchmarks. The accuracy of the proposed detection method is verified for single-speaker and multi-speaker scenarios by considering the WASN displayed in Fig. 1.2.

2.5.1 Batch-Mode Voice Activity Detection for Single-Speaker Scenario

The performance for single-speaker VAD is benchmarked against two existing single-node methods, i.e., the VAD-1 [2] and the VAD-2 [3] given observations from Node D_2 for Source S_2 and Node D_9 for Source S_7, respectively, of the scenario depicted in Fig. 1.2. The distributed multi-speaker VAD (DM-VAD) refers to the proposed VAD approach based on K-medoids and without post-processing (see Section 2.4.3), whereas DM-VAD+ includes this step. Tables 2.1-2.4 summarize the results of the comparative study for Sources S_2 and S_7 under Gaussian and babble noise conditions of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$. The babble noise sequences at each microphone are created by taking non-overlapping excerpts from a long babble noise process. It is therefore spatially independent. In the ensuing tables, the values in bold are indicators of a superior performance attained by our proposed VAD technique. The performance metrics are: correct decision (CD), missed detection (MD), false alarm (FA), equal error rate (EER), and cost of log-likelihood ratio ($C_{\text{llr}}^{\min}$). The EER reports the measure between the

frame-level speech and non-speech detections and $C_{\text{llr}}^{\min}$ measures the quality of the log-likelihood ratio detection output. In both cases, a small value corresponds to a highly accurate VAD. DM-VAD+ outperforms its single-node competitors by leveraging upon the WASN via a distributed M-NICA based on LONAS and achieves $> 92\%$ correct VAD in all cases.

Metric	VAD Results for Source S_2			
	DM-VAD	DM-VAD+	VAD-1	VAD-2
CD	64.8	**92.3**	89.5	63
MD	34.7	5.7	3	0
FA	**0.5**	2	7.5	37
EER	0.06	**0.06**	0.4	0.4
$C_{\text{llr}}^{\min}$	0.2	**0.2**	0.9	0.9

Table 2.1: Comparison of our approach with different benchmark algorithms, referred to as VAD-1 [2] and VAD-2 [3], for a single active Source S_2 and additive white Gaussian noise of variance $\sigma_{\omega}^2 = 0.01$.

For the same noisy environment and a different single Source S_7, Tab. 2.2 shows that DM-VAD+, VAD-1 and VAD-2 provide nearly perfect detection results.

Metric	VAD Results for Source S_7			
	DM-VAD	DM-VAD+	VAD-1	VAD-2
CD	80.3	96.2	94.6	96.3
MD	19.7	3.8	3.3	3.1
FA	**0**	**0**	2	0.6
EER	**0.01**	**0.01**	0.35	0.35
$C_{\text{llr}}^{\min}$	**0.03**	**0.04**	0.85	0.85

Table 2.2: Comparison of our approach with different benchmark algorithms [2, 3], for a single active source S_7 and additive white Gaussian noise of variance $\sigma_{\omega}^2 = 0.01$.

In Tabs. 2.3 and 2.4, we consider the single speech sources S_2 and S_7 corrupted with babble noise. Results show that VAD-1 and VAD-2 are more sensitive to babble noise since they lose in detection performance while the decisions in DM-VAD and DM-VAD+ remain stable.

Metric	VAD Results for Source S_2			
	DM-VAD	DM-VAD+	VAD-1	VAD-2
CD	65	**92.7**	88.2	61.9
MD	34.5	5.2	2.2	0
FA	**0.5**	2.1	29.6	38.1
EER	0.06	**0.06**	0.4	0.4
$C_{\text{llr}}^{\text{min}}$	**0.2**	**0.2**	0.9	0.9

Table 2.3: Comparison of our approach with different benchmark algorithms [2, 3], for a single active source S_2 and babble noise of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$.

Metric	VAD Results for Source S_7			
	DM-VAD	DM-VAD+	VAD-1	VAD-2
CD	80.3	**96.2**	94.6	57.6
MD	19.7	3.8	3.3	20.8
FA	**0**	**0**	2	21.6
EER	0.01	**0.01**	0.35	0.35
$C_{\text{llr}}^{\text{min}}$	0.03	**0.04**	0.85	0.85

Table 2.4: Comparison of our approach with different benchmark algorithms [2, 3], for a single active source S_7 and babble noise of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$.

2.5.2 Batch-Mode Distributed Multi-Speaker Voice Activity Detection

The performance of the proposed detector in batch-mode (see Algorithm 5) is evaluated on the challenging multi-speaker scenario with six active sources, as given in Fig. 1.2, for different variations of K-means and an additive white noise, e.g. AWGN or babble noise, of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$. For the considered multi-source multi-device speech scenario corrupted with additive noise of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$, the Signal-to-Noise-Ratio (SNR) and the Signal-to-Interference-Plus-Noise-Ratio (SINR) are computed. Table 2.5 summarizes the SNR results in dB computed using Eq. (2.46), such that

$$\text{SNR}_k^q = 10\log_{10}\left(\frac{\sigma_{k,q}^2}{\sigma_{\boldsymbol{\omega}}^2}\right), \quad k = [1, \cdots, K], q = [1, \cdots, Q], \tag{2.46}$$

where SNR_k^q describes the SNR value at device k related to source q, $\sigma_{k,q}^2$ is the qth signal power recorded at device k, and $\sigma_{\boldsymbol{\omega}}^2$ is the variance of the additive noise. Similarly, Tab. 2.6 outlines the SINR values in dB that are computed as

$$\mathrm{SINR}_k^q = 10\log_{10}\left(\frac{\sigma_{k,q}^2}{\sum_{q'}\sigma_{k,q'}^2 + \sigma_{\boldsymbol{\omega}}^2}\right), q \neq q', \tag{2.47}$$

where $\sigma_{k,q}^2$ is the power of the source-of-interest q recorded at device k, $\sum_{q'}\sigma_{k,q'}^2$ is the power of all interfering sources $q' \neq q$ excluding source q.

Device	**Sources**					
	S_2	S_4	S_5	S_6	S_7	S_3
D_1	-11.4589	-17.6404	-16.6260	-13.6959	-20.7077	-15.6372
D_2	0.0934	-11.6536	-11.5104	-18.6551	-17.2166	-2.1003
D_3	-10.9801	-18.1935	-16.0995	-22.5626	-7.6164	-17.4065
D_4	-3.5013	-9.9627	-8.4591	-16.2553	-11.0875	-12.2566
D_5	-6.9445	20.5279	-0.8062	-15.1585	-13.4586	-6.8904
D_6	-6.9354	20.5279	-0.8036	-15.0987	-13.3724	-7.5072
D_7	-6.9359	-1.3016	21.0934	-16.4537	-12.0481	-6.8898
D_8	-9.2737	-15.2052	-15.2053	24.0060	-18.6392	-14.6985
D_9	-9.4454	-16.1539	-14.1710	-20.4702	26.1795	-14.7554
D_{10}	-2.7619	-8.9624	-8.4145	-17.4757	-14.5563	26.4637
D_{11}	-8.2539	-12.4152	-13.4559	-4.7990	-18.9410	-13.0758
D_{12}	0.0967	-12.4637	-10.9967	-16.3664	-10.6964	-14.0706
D_{13}	-10.1892	-16.9620	-15.1463	-21.2894	-2.6891	-15.1536
D_{14}	-10.5095	-15.6839	-15.3758	-7.0405	-19.4056	-14.9124
D_{15}	-6.8680	-2.2273	-5.7471	-15.3000	-14.7592	-4.2921
D_{16}	-8.4652	-14.2714	-11.3672	-19.1954	-4.4415	-12.7729
D_{17}	-7.9013	-14.6474	-14.6394	-16.1804	-20.0453	-10.4947
D_{18}	-5.6399	-7.4222	-4.5985	-18.2256	-12.6457	-2.2425
D_{19}	-7.9034	-15.7171	-13.3606	-21.0486	-13.2083	-10.4876
D_{20}	-6.9695	-6.1520	-8.9937	-14.2016	-16.4199	-7.0315

Table 2.5: SNR for the multi-device ($K = 20$) multi-source ($Q = 6$) speech setup.

Tables 2.7 and 2.8 summarize the outcome of DM-VAD and DM-VAD+ for AWGN

Device	**Sources**					
	S_2	S_4	S_5	S_6	S_7	S_3
D_1	-11.9412	-18.3286	-17.2974	-14.2887	-21.4281	-16.2879
D_2	-2.4306	-16.0336	-15.8867	-23.1208	-21.6741	-5.5113
D_3	-11.9022	-19.3368	-17.2115	-23.7380	-8.1977	-18.5399
D_4	-4.9764	-12.3934	-10.7851	-18.8734	-13.5750	-14.7888
D_5	-27.5528	16.8864	-21.3907	-35.7733	-34.0729	-27.4986
D_6	-27.5428	16.9332	-21.3871	-35.7126	-33.9857	-28.1155
D_7	-28.0973	-22.4451	17.6041	-37.6211	-33.2142	-28.0511
D_8	-33.2988	-39.2318	-39.2320	23.1208	-42.6661	-38.7251
D_9	-35.6364	-42.3458	-40.3628	-46.6623	25.3210	-40.9472
D_{10}	-29.2385	-35.4430	-34.8949	-43.9574	-41.0378	23.7840
D_{11}	-10.0023	-14.4232	-15.4972	-5.9848	-21.0693	-15.1060
D_{12}	-0.9878	-15.9844	-14.4732	-19.9515	-14.1618	-17.6250
D_{13}	-12.3031	-19.2732	-17.4308	-23.6329	-3.4241	-17.4382
D_{14}	-11.6403	-17.0172	-16.7027	-7.7912	-20.7883	-16.2289
D_{15}	-10.4857	-5.0310	-9.2489	-19.2381	-18.6904	-7.5833
D_{16}	-10.3259	-16.4194	-13.4199	-21.4101	-5.6394	-14.8800
D_{17}	-8.6637	-15.8522	-15.8440	-17.4186	-21.3297	-11.5149
D_{18}	-9.0523	-11.0127	-7.8620	-22.1206	-16.4705	-4.9614
D_{19}	-8.7596	-17.0307	-14.6117	-22.4226	-14.4541	-11.5955
D_{20}	-9.0848	-8.1547	-11.3042	-16.7312	-18.9862	-9.1544

Table 2.6: SINR for the multi-device ($K = 20$) multi-source ($Q = 6$) speech setup.

of variance $\sigma_{\omega}^2 = 0.01$. Comparable detection results are achieved when alternating between the variants of the K-means algorithm, and the post-processing step is most useful for Source S_2, where the original speech signal is a noisy PA announcement. The worst-case CD for DM-VAD+ is $> 84\%$ in this challenging scenario.

Method	Metric	DM-VAD					
		S_2	S_4	S_5	S_6	S_7	S_3
K-means	CD	61.1	92.3	93.6	61.7	86.7	92.4
	MD	**38.1**	7.4	5.5	38.3	13.2	6.7
	FA	**0.8**	0.3	0.9	**0**	**0**	0.9
	EER	0.22	**0.02**	**0.02**	0.17	0.03	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.63	**0.09**	0.11	0.41	0.12	0.18
K-medians	CD	**70.4**	**93.1**	**95.2**	**86.6**	**88.7**	**93.8**
	MD	27.5	**6.6**	**3.8**	**13.5**	**11.3**	**5**
	FA	2.12	0.3	1	**0**	**0**	1.2
	EER	**0.14**	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.55	**0.09**	0.11	**0.06**	**0.03**	0.14
K-medoids	CD	62.7	85	82.1	74.7	80.3	88.1
	MD	36.3	14.9	17.6	25.3	19.7	11.2
	FA	1	**0.1**	**0.3**	**0**	**0**	**0.7**
	EER	0.15	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	**0.5**	**0.09**	**0.1**	**0.06**	**0.03**	**0.12**
K-MAD	CD	61.1	92.3	93.6	61.7	86.8	92.4
	MD	**38.1**	7.4	5.5	38.3	13.2	6.7
	FA	**0.8**	0.3	0.9	**0**	**0**	0.9
	EER	0.22	**0.02**	**0.02**	0.17	0.03	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.63	**0.09**	0.11	0.4	0.12	0.18
WK-MAD	CD	63.4	92.4	93.2	62.1	87	92.7
	MD	35.5	7.3	5.9	37.9	13	6.4
	FA	1.1	0.3	0.9	**0**	**0**	0.9
	EER	0.2	**0.02**	**0.02**	0.17	0.03	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.6	**0.09**	**0.1**	0.39	0.12	0.18

Table 2.7: Proposed batch-mode DM-VAD using different clustering methods for signals corrupted by additive white Gaussian noise of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$.

2.5.3 Sequential-Mode Distributed Multi-Speaker Voice Activity Detection

The performance of the proposed detector in the sequential-mode, i.e., SDM-VAD+ is evaluated on the challenging multi-speaker scenario with six active sources, as given in Fig. 1.2, for K-medoids.

Table 2.9 displays the VAD results when using a growing window $W^{(n)}, \quad n = W^0 + 1, \ldots, N$, for AWGN of variance $\sigma_{\boldsymbol{\omega}}^2 = 0.01$. A performance loss of maximally 6%, compared to the batch-mode is experienced. Figure 2.9 depicts the convergence for

Method	Metric	DM-VAD+					
		S_2	S_4	S_5	S_6	S_7	S_3
K-means	CD	85.2	96.2	97	89.9	**96.2**	**94.8**
	MD	5.1	**0.8**	0.9	10.1	**3.8**	**2.2**
	FA	9.7	3	**2.1**	**0**	**0**	**2.9**
	EER	**0.15**	**0.02**	**0.02**	**0.01**	0.03	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.5	**0.09**	**0.1**	0.07	0.12	0.18
K-medians	CD	**86.3**	**96.3**	97	**93.6**	**96.2**	**94.8**
	MD	**3.5**	**0.8**	0.9	**6.4**	**3.8**	**2.2**
	FA	10.1	**2.9**	**2.1**	**0**	**0**	**2.9**
	EER	**0.15**	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	**0.49**	**0.09**	0.11	**0.06**	**0.04**	**0.14**
K-medoids	CD	84.6	**96.3**	**97.1**	**93.6**	**96.2**	**94.8**
	MD	6	**0.8**	**0.8**	**6.4**	**3.8**	**2.2**
	FA	**9.4**	**2.9**	**2.1**	**0**	**0**	**2.9**
	EER	**0.15**	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.5	**0.09**	**0.1**	**0.06**	**0.04**	**0.14**
K-MAD	CD	85.2	96.2	97	89.9	**96.2**	**94.8**
	MD	5.1	**0.8**	0.9	10.1	**3.8**	**2.2**
	FA	9.7	3	**2.1**	**0**	**0**	**2.9**
	EER	**0.15**	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.5	**0.09**	**0.1**	**0.06**	**0.04**	**0.14**
WK-MAD	CD	84.8	96.2	**97.1**	89.8	**96.2**	**94.8**
	MD	5.5	**0.8**	**0.8**	10.2	**3.8**	**2.2**
	FA	9.7	3	**2.1**	**0**	**0**	**2.9**
	EER	**0.15**	**0.02**	**0.02**	**0.01**	**0.01**	**0.04**
	$C_{\text{llr}}^{\text{min}}$	0.5	**0.09**	**0.1**	**0.06**	**0.04**	**0.14**

Table 2.8: Proposed batch-mode DM-VAD+ using different clustering methods for signals corrupted by additive white Gaussian noise of variance $\sigma_{\omega}^2 = 0.01$.

the different speech sources to their associated VAD decision. Clearly, the transient behavior of the SDM-VAD+ is source-dependent. When using the growing window technique described in Section 2.4.5, SDM-VAD+ for this setup achieves a steady state performance after approx. $300 - 500$ speech frames of 30ms duration each.

The performance is next analyzed using a fixed size moving window, i.e., the buffer of the past speech data included in the decision is limited to 400 frames. Table 2.10 summarizes the fixed sliding window SDMVAD+ results, while Fig. 2.10 displays the convergence of the SDM-VAD+ for different sources.

Source	Metric				
	CD	MD	FA	EER	$C_{\text{llr}}^{\text{min}}$
S_2	80.2	5.7	14.1	0.2	0.63
S_4	92.9	1.7	5.4	0.07	0.32
S_5	90.8	**1**	8.2	0.06	0.22
S_6	90.6	7.4	2	0.07	0.31
S_7	**94**	4.3	**1.7**	**0.04**	**0.22**
S_3	89.5	2.5	8	0.07	0.27

Table 2.9: Proposed sequential-mode VAD (SDM-VAD+) with K-medoids using a growing window for signals corrupted by additive white Gaussian noise with variance $\sigma_{\omega}^2 = 0.01$.

Source	Metric				
	CD	MD	FA	EER	$C_{\text{llr}}^{\text{min}}$
S_2	78.8	6.1	15.1	0.2	0.65
S_4	92.8	1.6	5.6	0.1	0.39
S_5	90.7	**0.9**	8.4	0.06	0.2
S_6	90.3	7.2	2.5	0.09	0.37
S_7	**93.9**	4.7	**1.4**	**0.04**	**0.23**
S_3	88.9	2.3	8.8	0.06	0.26

Table 2.10: Proposed sequential-mode VAD (SDM-VAD+) with K-medoids using a fixed sliding window for a mixture of energies corrupted by additive white Gaussian noise with variance $\sigma_{\omega}^2 = 0.01$.

2.6 Conclusions

The current chapter describes a novel technique that solves the challenging multi-source VAD in a distributed WASN. The proposed method, i.e. DM-VAD, is validated on real speech data consisting of multiple simultaneously interfering source signals in a reverberant and noise conditioned environment.

The presented DM-VAD requires mainly two information about the data. Well-labeled source signals and a known number of interfering sources. However, the labeling and source enumeration information can be obtained beforehand using existing techniques [34, 103].

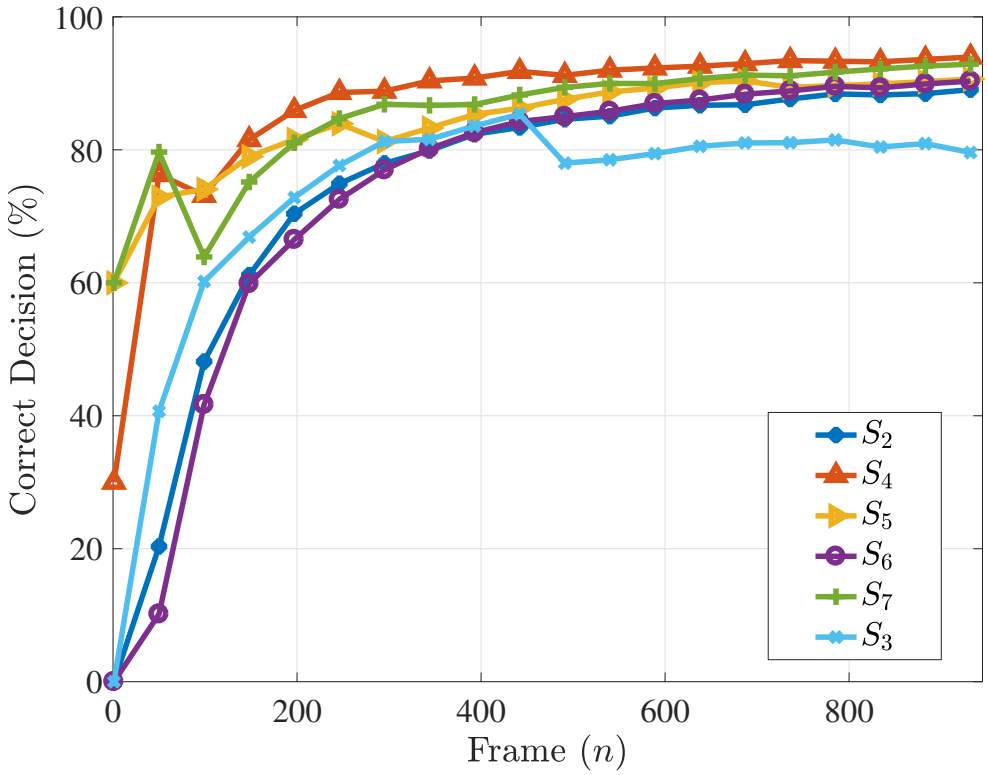


Figure 2.9: Sequential decision of SDM-VAD+ using a growing window.

The key idea of this chapter is to operate a voice activity decision function after well-separating the energies of the mixture of received signals locally at node cluster level. This step is mandatory because we target a source-specific speech discrimination, meaning that, our voice activity system should output speech activity pattern for every contributing speech source at every node of the network. M-NICA provides quality non-negative energy features that can be used for VAD for small numbers of conflicting speech sources. However, it quickly loses its performance with increasing speaker source numbers and an amplified environmental noise. Following this reasoning, we have extended the centralized multiple non-negative energy separation to a distributed energy separation problem. A distributed setup is related to forming clusters of nodes around every source-of-interest using the LONAS algorithm in [1]. The logic behind considering this built-in framework pertains to estimating a rank-one M-NICA algorithm at every cluster of nodes to extract a unique energy signal related to a well-labeled source. To this end, the multiple source separation problem in a centralized WASN boils down to executing multiple rank-one M-NICA algorithms in parallel, in a distributed manner, at every well-defined cluster of nodes to estimate a cleaner unique dominant energy source.

Consequently, the DM-VAD technique for WASNs does not require a FC or prior

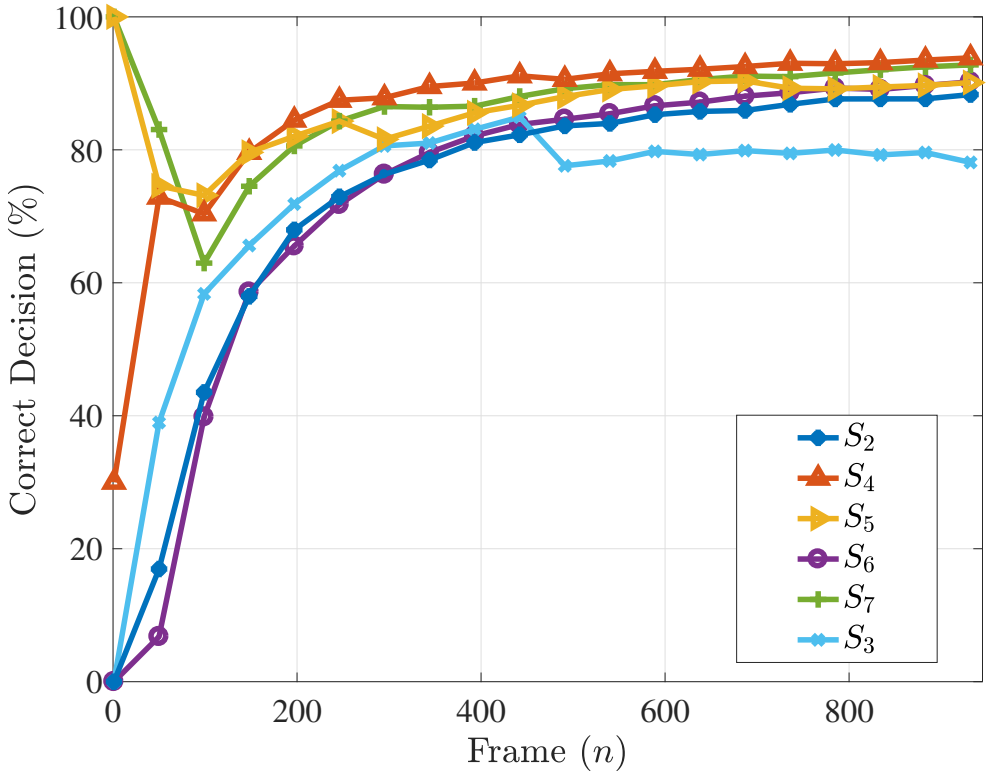


Figure 2.10: Sequential decision of SDM-VAD+ using a fixed sliding window for the scenario given in Fig. 1.2.

knowledge about the node positions, microphone array orientations or the number of observed sources. The multi-speaker VAD has then been approached by extracting further robust low-dimensional short-term features from the unmixed energy signals by applying robust K-means type clustering algorithms. Additionally, robust clustering approaches, namely K-MAD and WK-MAD, have been presented. These partitional robust techniques provide comparable clustering results with built-in robust metrics for the estimation of the clusters' centroids. We have also introduced in this chapter a comparative study of our approach in contrast to state-of-the-art methods. The results of this study demonstrate the capabilities of the proposed approach in both single-source and multiple-source VAD. Hence, the steps of the proposed clustering-based DM-VAD method show promising performance compared to existing benchmark methods. More than 85% of correct detection in the worst case is obtained for a challenging scenario where 20 nodes observe 6 sources in a simulated reverberant rectangular room illustrated in Fig. 1.2. The proposed method is also able to operate for streaming data taking into account a small performance loss compared to the batch-mode case.

Chapter 3

Robust Distributed Sparse Constrained Non-Negative Blind Energy Separation for Multi-Speaker Voice Activity Detection in Wireless Acoustic Sensor Networks

'I learned that courage was not the absence of fear, but the triumph over it'

Nelson Mandela

In various speech processing applications, short-term features reflect the speech statistical parameters. In this PhD project, short-term energy-based feature analysis is accomplished based on shifting short-time window, see [4]. The chosen window size is of 30 msec where short-term stationarity of the speech signal is assumed. No window overlapping is used, which implicates a window step size of 30 msec. Considering the energy level as an indicator for speech activity follows intuition and is especially useful for the multi-speaker setting, where the energies associated to the sources must be separated from a received mixture in order to obtain speaker-specific VAD patterns [1, 5, 34, 36, 85]. Energy separation is computationally less expensive compared to the source unmixing task for time-series. The combination of short-term energy features with the energy unmixing process is investigated in depth to solve the multi-source VAD problem.

In many real-world applications, such as biomedical imaging, energy consumption analysis, pixel intensities data and spectral analysis, the sources to recover are non-negative. The positive sources have physical meanings, which favor the processing of the non-negative data under non-negativity constraints. Several authors have proposed methods for solving the noise-free separation problem under non-negativity constraint, such as the non-negative matrix factorization (NMF) [98, 104, 105], or the NICA approaches [106–109]. In this thesis, we assume statistically independent positive latent energy sources, then we consider solving energy-based NICA with the intent to obtain a better voice activity detector. More recent NICA techniques, such as [85], further assume well-grounded random variables with a non-vanishing probability density function (pdf) in the positive neighborhood of zero. Under these conditions, the minimization

of the inter-source correlation is a way to recover the hidden sources up to permutation and scaling ambiguities. However, these standard constraints are not always sufficient to guarantee the uniqueness of the solution. We have mentioned previously in Chapter 2, Section 2.3.3, that deriving the distributed M-NICA approach overcomes the permutation indetermination issue. In the challenging noise-embedded NICA modeling that we assume in Eq. (2.3), we believe that improving the energy separation based on the noisy instantaneous mixtures in $\mathbf{y}[n]$ for VAD is a feasible procedure. Hence, in this work, sparseness constraints are incorporated in the noise-embedded NICA to improve the energy recovery of the multiple contributing speech sources and reduce the range of admissible solutions. This, indeed, favors a particular type of solution for VAD, where sparsity plays a role in tuning noisy energies to exactly (non-active) zero-valued data points.

3.1 Introduction

ICA is a well established technique that is capable of separating independent sources that are linearly mixed, e.g., in a wireless sensor network (WSN) [92,95,96,110]. Given a multivariate observed data, ICA characterizes the model for which some unmixed latent variables and a mixing system are unknown and subject to estimation. The latent variables are the source signals that determine the independent components of the observed data. A vast amount of research explores ICA, (see, e.g. [94, 107]), particularly its powerful performance compared to other methods such as principal component analysis (PCA) [4, 84, 85]. Many applications require ICA for their data analysis, including image denoising [111] and face recognition [112] in digital images or speech enhancement [113] and voice activity detection (VAD) [1, 4, 85].

In this chapter, we consider the application of multi-speaker VAD for a WASN. This involves dealing with mixtures of simultaneously recorded speech signals at spatially distributed microphones. ICA is used to extract unmixed (non-negative) energy signals, based on which speaker specific VAD is performed. Non-negative ICA algorithms (NICA) are presented in [106, 108]. Similar representations that are tailored to the statistics of non-negative data exist in the literature, such as the NMF introduced in [98, 114]. In noisy environments, the lack of robustness is very problematic for NICA. In fact, the majority of the NICA methods assume a noise-free model in order to keep the problem tractable. However, this assumption is unrealistic in speech scenarios. Consequently, we assume an embedded-noise NICA model for which the proposed energy source unmixing can be solved using second order statistics only. The

latter assumption makes the proposed NICA-based unmixing approach computationally efficient compared to ICA algorithms that use higher order statistics (HOS) [5].

A variety of non-negative data representation problems take advantage of the ℓ_1-norm regularization in order to obtain a sparse representation of the solution [115]. This is known as non-negative sparse coding (SC) and is proposed in [104,105,116–119]. SC is widely applicable in signal analysis [116,120]. It defines a representational scheme that favors a desired degree of sparseness by means of a data-driven sparseness measure. The idea of enhancing the energy features with a penalized ℓ_1-norm model is relevant, as it readily brings about a straightforward zero-threshold VAD, which detects speech activity. Conjointly, despite the abundance of studies that focus on NICA, to the best of our knowledge none has targeted multi-speaker VAD with sparseness constraints. We show that our proposed sparse algorithm better reconstructs the unknown non-negative speech energies as compared to the standard M-NICA along with solving the speech/non-speech energy discrimination for a VAD purpose. Higher correct detection results are achieved even in challenging noisy and reverberant WASN conditions such as the cocktail party scenario illustrated in Fig. 1.2.

In spite of its efficiency, one can think of performing VAD without having to compute a complete sparse-based unmixing process. Solving the NICA problem with an incorporated multiplicative gradient descent updates is of "supplementary" computational load for the multi-source VAD task. Hence, in the current chapter, we also target developing a multi-speaker VAD technique with a built-in semi separation procedure that does not consume a complete energy unmixing step, such as M-NICA [85] or the proposed sparse blind energy separation [5]. The technique aims at achieving a multi-source VAD of lower complexity by only considering a stability-based penalized sum-of-squares criterion minimization with an ℓ_1-regularized term. The latter empowers sparse SVD layers with an automatic selection of the sparseness parameters, which produce decently separated non-negative energy features suitable for the multi-source VAD task. A detailed list of contributions provided in this chapter is given in Section 3.2.

3.2 Contributions

In this chapter[1], a sparse median-based M-NICA (SMM-NICA) approach is proposed that provides an improvement to the M-NICA algorithm in [85] by integrating spar-

[1]This chapter is based on the following conference articles:

- "Multi-Speaker Voice Activity Detection by an Improved Multiplicative Non-Negative Inde-

sity constraints in the embedded-noise NICA model. Sparsity is introduced by using a sparse singular value decomposition (SSVD) as an initial step for the multiplicative update. We initialize our algorithm by projecting the non-negative data onto the right rotation matrix subspace on which we impose sparsity. We propose an improved version of the M-NICA algorithm. We examine a challenging NICA application in a noise-embedded multi-speaker VAD setup. We present a novel approach that includes sparsity constraints to solve the energy separation problem with independent source signals. A sparse feature extraction step is performed to project the non-negative signals onto a dimension-reduced subspace and identify sparse principal components. Then, we maximize the signal decorrelation by employing a median measure of central tendency in the computation of the covariance matrix that contributes in robustness against outliers. Moreover, our approach supplies a straightforward multi-speaker VAD, for which no empirical thresholding or other ad hoc decision rule is required. Instead, an active voice frame simply corresponds to a non-zero value of the separated energy signal. Numerical experiments using real data validate the superior performance of the proposed technique.

In addition, a robust multi-source VAD is presented. Robustness is achieved by integrating a $t_\nu M$-estimator of the covariance matrix in the multi-source separation algorithm, resulting in the proposed $t_\nu M$-SMM-NICA algorithm. The robust energy separation approach in the presence of multi-sources is tested in a centralized and a distributed scenario. We further improve the sparse multi-source separation algorithm, i.e. $t_\nu M$-SMM-NICA, by taking into account a regularized $t_\nu M$-estimator. The robust estimators for the computation of the covariance matrix in the NICA-based decorrelation procedure provide valuable results in terms of correct detection rate. Hence, the suggested robust and sparse energy separation algorithms serve well the VAD process. Moreover, we measure the quality of the estimated voice activity for every source, by performing speech enhancement based on the generated VAD patterns. The estimated node-specific speech signals of the WASN using the robust sparse VAD algorithm are compared to the estimated signals with an M-NICA-based VAD.

Moreover, an alternative approach for robust multi-speaker VAD in WASNs is pre-

pendent Component Analysis With Sparseness Constraints", in Proc. 42nd IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP).

- "Robust Distributed Multi-Speaker Voice Activity Detection Using Stability Selection for Sparse Non-Negative Feature Extraction", in Proc. 25th IEEE Eur. Signal Process. Conf. (EUSIPCO).

- "Robust Distributed Sparsity-Constrained Non-Negative Source Separation and Multi-Speaker Voice Activity Detection for Distributed Speech Enhancement in Wireless Acoustic Sensor Networks", Submitted to the Proc. 2nd IEEE Int. Conf. Signals Syst. (ICSigSys).

sented in this chapter. We improve upon [5] and [1] with a two-step robust solution to the multi-speaker VAD problem by exploiting SC, see [116, 120]. 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. 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 VAD. Additionally, a subsequent robust classification step that uses robust Mahalanobis distance based on M-estimation is performed. Hence, separating active speech frames from silent frames is done via the proposed robust Mahalanobis classifier that is based on an M-estimation of the covariance matrix. A sub-sampling-based variable selection method for SC combined with a robust Mahalanobis classifier appropriately addresses the multiple speech activity detection task and makes it unnecessary to use a posterior energy unmixing method, as proposed by the SMM-NICA method [5]. In this work, stability selection is, in fact, exploited to extract well-separated energy speech signatures.

Both centralized and distributed multi-speaker VAD is considered in this chapter, and in both cases highly accurate VAD results are obtained. The assessment of our methods is actually tested on a challenging WASN of 20 nodes observing 6 sources in a reverberant environment. Highly accurate VAD results are obtained, which makes the blind energy separation techniques for VAD of great potential in the multi-source multi-device speech signal processing.

3.3 Multi-Speaker Voice Activity Detection by an Improved Multiplicative Non-Negative Independent Component Analysis With Sparseness Constraints

In this section[2], a new algorithm for energy source separation based on sparse modeling is presented. Original ICA methods approach the separation problem with a noise-free assumption on the model. In particular, our approach aims at improving the M-NICA method that relies on multiplicative updates while assuming a noise-embedded environment and a sparse modeling of the received noisy mixtures.

[2]This section is based on our work presented in the conference article entitled: "Multi-Speaker Voice Activity Detection by an Improved Multiplicative Non-Negative Independent Component Analysis With Sparseness Constraints", in Proc. 42$^{\text{nd}}$ IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP).

3.3.1 Proposed Median-Based M-NICA With Sparsity Constraints (SMM-NICA)

A useful property that could be added to the M-NICA algorithm is the capability of producing a sparse representation. A sparse M-NICA solution can encode much of the data using few active components, which better reflects reality and makes the encoding easy to interpret. Applied to VAD for instance, the concept of sparse coding introduces a representational scheme where only few units out of a large population are effectively used. The active units can be interpreted as active speech, while the zero elements, are non-active speech. Many sparseness measures are proposed in literature. With regards to non-negative representation, we suggest using the ℓ_1 regularization. In the following subsections, we explain how sparse coding is integrated into the M-NICA algorithm for the sake of better signal estimation and an enhanced VAD procedure in a noisy environment.

3.3.1.1 Sparse Singular Features

We define $\mathbf{Y} \in \mathbb{R}_+^{M\times N}$ as the matrix containing all vectors $\mathbf{y}[n]$ from Eqs. (2.3)–(2.5) with $n = 1, \ldots, N$. The standard M-NICA algorithm pre-processes the data using a singular value decomposition step (SVD). The latter can be seen as a PCA technique in itself that extracts the first principal components [121]. Transforming $\mathbf{Y}$ using the SVD projects the signal onto the sub-spaces

$$\mathrm{SVD}(\mathbf{Y}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{3.1}$$

where the left orthogonal matrix $\mathbf{U} \in \mathbb{R}^{M\times M}$ represents the principal directions, $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times N}$ is the scaling matrix of singular values, and $\mathbf{V}^\top \in \mathbb{R}^{N\times N}$ is the right rotation orthogonal matrix of singular vectors. The matrix product $\boldsymbol{\Sigma}\boldsymbol{V}^\top \in \mathbb{R}^{M\times N}$ embodies the principal components. The linear transformation $\boldsymbol{\Sigma}$ controls the speech energies by a scale factor that is the same in all directions. Omitting this factor does not deteriorate the signal shape. In addition, based on [122], the criterion of orthogonality for the vectors in $\mathbf{U}$ forces the right vectors in $\mathbf{V}$ to be a mixture of sources. We suggest using the matrix of right singular vectors as a feature for the subsequent separation step. We employ a sparse decomposition (SSVD) in lieu of an SVD to extract sparse features. We impose sparsity on the right rotation matrix $\mathbf{V}$ and seek a lower rank representation of the matrix $\mathbf{Y}$ with the requirement that the right singular vectors $\mathbf{v}$

are sparse for every source $q = 1, \ldots, Q$. This means that these vectors may have many zero entries. Conjointly, we believe that adding sparsity to these components yield a more parsimonious representation, clearly emphasizing a unique feature contribution to the subsequent multiplicative update rule. In this chapter, we use the individual sparse right singular matrix as some scaled features of the principal components obtained via SSVD. Our features are attractive since, as described previously, the space of right singular vectors is considerably smaller than the observation space. As a summary, our main task is to recover sparse and well unmixed right rotations $\mathbf{v}$ for every source $q = 1, \ldots, Q$ that we use as features for the multi-speaker VAD procedure.

3.3.1.2 Sparse Right Singular Vectors Subspace Projection

First, a rank-one SVD layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ is the best approximation of $\mathbf{Y}$ if it solves

$$\arg\min_{\sigma,\mathbf{u},\mathbf{v}} \|\mathbf{Y} - \sigma\mathbf{u}\mathbf{v}^\top\|^2, \tag{3.2}$$

where $\mathbf{u}$ is a unit M-vector and $\mathbf{v}$ is a unit N-vector. In order to obtain a sparse vector $\mathbf{v}$, we add sparsity-inducing penalties on $\mathbf{v}$ in the optimization objective in Eq. (3.2). We thus can expand Eq. (3.2) with an ℓ_1 regularization penalty term to formulate a sparsity promoting optimization problem. Specifically, we minimize with respect to the triplet $(\sigma, \mathbf{u}, \mathbf{v})$ the following penalized sum-of-squares criterion

$$\|\mathbf{Y} - \sigma\mathbf{u}\mathbf{v}^\top\|^2 + \lambda_\mathbf{v}\Phi(\sigma\mathbf{v}), \tag{3.3}$$

with $\Phi(\sigma\mathbf{v})$ being the ℓ_1 regularization function

$$\Phi(\sigma\mathbf{v}) = \sum_{n=1}^{N} |\sigma v[n]| \tag{3.4}$$

and $\lambda_\mathbf{v}$ being the non-negative penalty parameter. If $\lambda_\mathbf{v} = 0$ than Eq. (3.3) is equivalent to Eq. (3.2). The selection of $\lambda_\mathbf{v}$ corresponds to selecting the degree of sparsity of $\mathbf{v}$. The latter is the number of zero components in $\mathbf{v}$ or, based on [118], the number of n elements that satisfy

$$g(\lambda_{\mathbf{v}}) = \#\left(\left\{n \in \{1, \cdots, N\} : [\mathbf{Y}^{\top}\mathbf{u}]_n \sigma > \frac{\lambda_{\mathbf{v}}}{2}\right\}\right) \tag{3.5}$$

for a fixed $\mathbf{u}$. Here, $g(\lambda_{\mathbf{v}})$ is the degree of sparsity function and $\#(\cdot)$ represents the cardinality symbol. Moreover, [118] and [123] suggest the use of the Bayesian information criterion (BIC), from [124], to estimate the optimal number of non-zero coefficients

$$\mathrm{BIC}(\lambda_{\mathbf{v}}) = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{mn\hat{\varrho}^2} + \frac{\log(mn)}{mn} g(\lambda_{\mathbf{v}}) \tag{3.6}$$

with $\hat{\varrho}^2$ denoting the ordinary least squares estimate of the error variance in Eq. (3.3). In order to reach a sparse $\mathbf{v}$, the minimization of Eq. (3.3) with respect to $\sigma\mathbf{v}$ is iterated until convergence. A closed-form solution for minimizing $\sigma\mathbf{v}$ in Eq. (3.3) is detailed in Appendix (A.1). Consequently, it follows that the sparse representation of the vector $\mathbf{v}$ is obtained using a component-wise update that computes every element $v[n]$ of $\mathbf{v}$ based on

$$v[n] = \frac{1}{\sigma}\left[\mathrm{sgn}\left\{[\mathbf{Y}^{\top}\mathbf{u}]_n\right\}\left(|[\mathbf{Y}^{\top}\mathbf{u}]_n| - \frac{\lambda_{\mathbf{v}}}{2}\right)\right], \tag{3.7}$$

with $\lambda_{\mathbf{v}}$ being the minimiser of Eq. (3.6).

3.3.1.3 The Proposed SMM-NICA Algorithm

Algorithm 7 summarizes the steps of our method. Given $\mathbf{Y}$, we iterate Eq. (3.2)–(3.10) to build a matrix of sparse singular vectors $\mathbf{V}^{\mathcal{S}}$. Specifically, the matrix $\mathbf{V}^{\mathcal{S}}$ is first initialized as an empty matrix in Algorithm 7. Step 2. Then, at every qth iteration of Algorithm 7. Eq. (3.2)–(3.10), the computed sparse vector $\mathbf{v}$ relative to source q is concatenated to $\mathbf{V}^{\mathcal{S}}$ using Algorithm 7, Step 7. After Q iterations of Eq. (3.2)–(3.10), the matrix $\mathbf{V}^{\mathcal{S}}$ is of dimensions $\mathbb{R}_{+}^{Q\times N}$ and contains all the Q computed sparse vectors $\mathbf{v}$ in its rows. We use the final state of the sparse matrix $\mathbf{V}^{\mathcal{S}}$ to initialize the proposed SMM-NICA as shown in Eq. (3.11). Then, Eqs. (3.12)-(3.16) are reiterated to retrieve an invariant estimate of $\mathbf{S}$. According to [4,85], the nature of the multiplicative update introduced in Eq. (3.16) conserves the non-negativity of the matrix $\mathbf{S}$. The function $D\{\cdot\}$ in Eqs. (3.14)-(3.15) sets all off-diagonal elements to null. In Eq. (3.12), we use the median central measure instead of the mean suggested in [4, 85]. A precise

descriptive measure depends highly on the shape of the data distribution. The median mid-point outperforms the mean in terms of accuracy for heavy tailed distributions since the mean can strongly be influenced by a small number of outliers [125]. Fig. 3.1 shows the right-skewed histogram for the energy of source S_2 considered in Fig 1.2. Obviously, the mean characterizes the relatively high but infrequent values. For our purpose, the median is a better summary of the typical value.

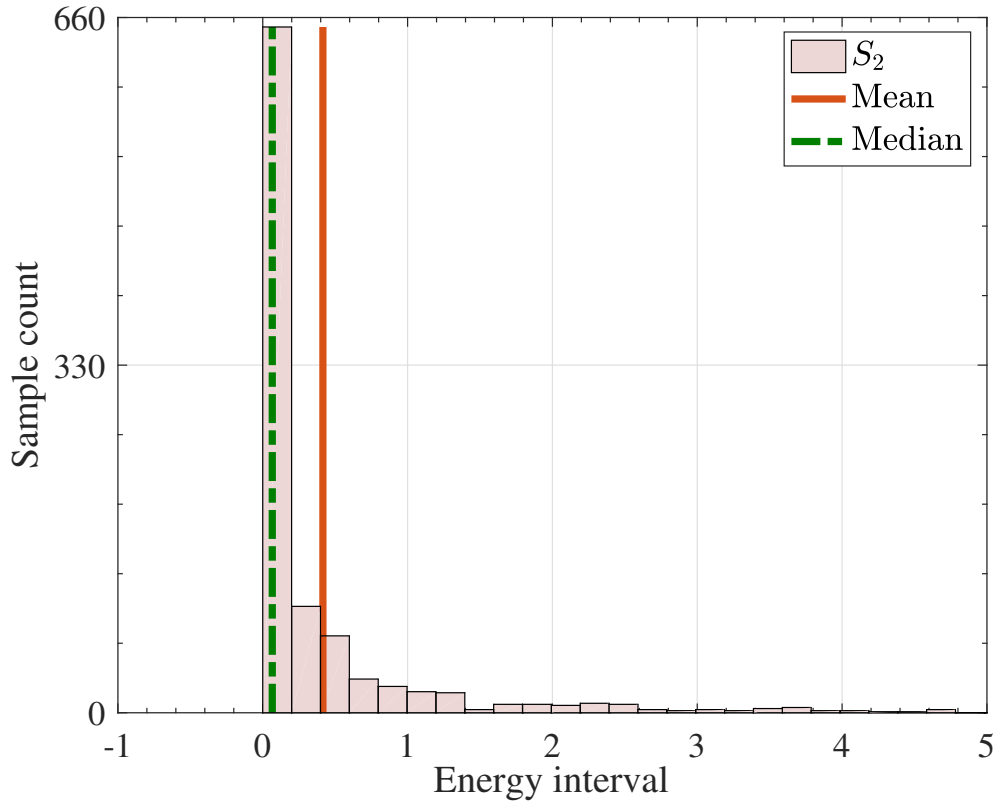


Figure 3.1: Right-skewed histogram for the ground truth energies of S_2 with the mean (red line) and median (dashed green) speech energy central values.

3.3.1.4 Computational Complexity of the Proposed SMM-NICA

Since the SMM-NICA method is composed of two phases, namely the pre-processing phase and a subsequent decorrelation step, we study the complexity of the two phases separately. First, the SVD decomposition fulfilled in the pre-processing step requires time complexity of the order of $\mathcal{O}(\min(M^2N, N^2M))$. A subsequent step is based on iteratively extracting SVD layers upon which we impose sparsity. The step of sparse optimization problem based on Lasso described in Eq. (3.3) is obviously the computationally most expensive. In fact, the sum of squares and the Lasso penalty are both convex. However, the Lasso loss function is not strictly convex. This is the reason why

Algorithm 7 The SMM-NICA algorithm

Input

$\mathbf{Y} = (\mathbf{y}[1], \cdots, \mathbf{y}[N]) \in \mathbb{R}_{+}^{M \times N}$ based on Eq. (2.3)

$\mathbf{V}^{\mathcal{S}} \triangleq \emptyset$

Initialization

for $q = 1, \ldots, Q$ **do**

Extract rank-one SVD layer $(\sigma, \mathbf{u}, \mathbf{v})^{\top}$ from $\mathbf{Y}$ that solves Eq. (3.2)

Minimize Eq. (3.3) with respect to $\mathbf{v}$

Update the sparse right singular vector $\mathbf{v}$ using Eq. (3.7)

Construct the sparse matrix $\mathbf{V}^{\mathcal{S}} \triangleq \mathbf{V}^{\mathcal{S}} \parallel \mathbf{v}^{\top}$, with $\parallel$ being the concatenation symbol.

Set

$$\sigma = \mathbf{u}^{\top}\mathbf{Y}\mathbf{v} \tag{3.8}$$

Compose a sparse lower-rank matrix

$$\mathbf{Y}_{\text{SSVD}} = \sigma \mathbf{u}\mathbf{v}^{\top} \tag{3.9}$$

Matrix subtraction

$$\mathbf{Y} = \mathbf{Y} - \mathbf{Y}_{\text{SSVD}} \tag{3.10}$$

end for

Define

$$[\mathbf{S}]_{q,n} \leftarrow |[\mathbf{V}^{\mathcal{S}}]_{q,n}|, \forall q = 1, \ldots, Q, \forall n = 1, \ldots, N. \tag{3.11}$$

repeat

$$\ddot{\mathbf{S}} = \underset{(q) \in Q}{\text{median}}\{\mathbf{S}_q\}\mathbf{1}_N^{\top}, \forall q = 1, \ldots, Q \tag{3.12}$$

$$\mathbf{C}_S = (\mathbf{S} - \ddot{\mathbf{S}})(\mathbf{S} - \ddot{\mathbf{S}})^{\top} \tag{3.13}$$

$$\mathbf{\Lambda}_1 = D\{\mathbf{C}_S\} \tag{3.14}$$

$$\mathbf{\Lambda}_2 = D\{(\mathbf{\Lambda}_1^{-1}\mathbf{C}_S)^2\} \tag{3.15}$$

Minimize the correlation in $[\mathbf{S}]_{q,n}$

$$[\mathbf{S}]_{q,n} \leftarrow [\mathbf{S}]_{q,n} \left[\frac{\ddot{\mathbf{S}}\mathbf{S}^{\top}\mathbf{\Lambda}_1^{-1}\mathbf{S} + \mathbf{S}\mathbf{S}^{\top}\mathbf{\Lambda}_1^{-1}\ddot{\mathbf{S}} + \mathbf{\Lambda}_2\mathbf{S}}{\ddot{\mathbf{S}}\mathbf{S}^{\top}\mathbf{\Lambda}_1^{-1}\ddot{\mathbf{S}} + \mathbf{S}\mathbf{S}^{\top}\mathbf{\Lambda}_1^{-1}\mathbf{S} + \mathbf{\Lambda}_2\ddot{\mathbf{S}}} \right]_{q,n} \tag{3.16}$$

until reaching a fixed-point of Eqs. (3.12)-(3.16)

there may be no global solution, but there are multiple local solutions that minimize the Lasso loss function. There are many fast algorithms meant to solve the Lasso regression problem in the general form given in Eq. (3.3). This optimization problem can be solved using quadratic programming yielding an overall upper-bound for the worst-case complexity of $\mathcal{O}(3^M)$. Meanwhile, the least angle regression and shrinkage (LARS) can solve the Lasso optimization problem, which leads to a computational complexity of $\mathcal{O}(MN\min(M,N))$. This computational time is required for every iteration q, until reaching a maximum extracted SVD loads, which is defined as the number of the treated targets Q in Algorithm 7. With regards to the decorrelation step, which is based on the M-NICA algorithm, the complexity is of the order of M-NICA. This means, the computational cost of the median-based decorrelation step of the SMM-NICA is $\mathcal{O}(Q^2N)$.

3.3.2 Experimental Results and Discussion

In this section, we provide simulation results for the multi-speaker energy separation based on our proposed SMM-NICA technique. We consider the scenario depicted in Fig. 1.2 with two active speech sources S_2 and S_3 affected by a reverberant environment. We compare the performance of the proposed algorithm with the original M-NICA based on diverse performance metrics in different noise variance environments. Table 3.1 outlines the overall separation results when a mixture of two active speech sources (S_2, S_3) is considered. These mixtures are corrupted with noise of two variance levels, i.e. $\sigma_{\boldsymbol{\omega}}^2 = \{0.1, 0.5\}$. Tables 2.5 and 2.6 contains an example of the SNR and SINR values in dB, for an additive babble noise of power $\sigma_{\boldsymbol{\omega}}^2 = 0.01$ to the mixed multi-source scenario illustrated in Fig. 1.2, recorded at each microphone and computed using Eqs. (2.46) and (2.47), respectively. In a first experiment, an additive white Gaussian noise (AWGN) is considered. The proposed method reduces the root mean square error (RMSE) considerably. We further assess our results in terms of the signal correlation ρ. The proposed SMM-NICA is capable of enhancing the signal correlation for S_2, as shown in Tab. 3.1. The distance between the estimated energies and the ground truth is evaluated through ℓ_1 and ℓ_2 norms, respectively. The developed SMM-NICA technique displays remarkably small distances outperforming M-NICA in all cases. Moreover, we analyse the normalized RMSE that omits the energy scaling in the performance assessment. Besides an accurate energy separation performance, the efficiency of SMM-NICA remains stable as the variance of the noise increases to $\sigma_{\boldsymbol{\omega}}^2 = 0.5$. Figure 3.2 (a) and Fig. 3.3 (a) illustrate the ground truth energies for S_2 and S_3, respectively. The corresponding unmixed energies produced by M-NICA are depicted in Fig. 3.2 (b) and Fig. 3.3 (b). It can be seen that some erroneous

energy spikes, appear in the M-NICA result. For example, the energies in Fig. 3.2 (b) experience a cross-talk in the frame interval around $k = [450, \cdots, 550]$, which obviously belongs to the alternative source S_2. On the other hand, Fig. 3.2 (c) shows a high accuracy in the sense that the cross-talk is attenuated and most of the supposedly zero-energies are indeed attenuated to zero and thus properly unmixed. We further compare the performance of M-NICA to our proposed technique for S_3 where the SMM-NICA in Fig. 3.3 (c) precisely tunes the energies describing the pause regions to zero. As a second case, we also study the performance of the proposed method with an additive babble noise. The results are summarized in Tab. 3.3. In addition, we compare the median-based M-NICA, or MM-NICA, to the standard Mean-based M-NICA and our proposed variants, i.e., SMeM-NICA and SMM-NICA. Again, both SMeM-NICA and SMM-NICA outperform M-NICA. In terms of energy separation, the MM-NICA performs better than the original M-NICA in the used speech use-case sketched in Fig. 1.2. The energy unmixing results of MM-NICA for the centralized scenario are depicted in Tab. 3.2. Regarding the VAD performance, we exploit the sparse estimated energies in the VAD procedure. Hence, a simple detector that does not require a threshold is implemented. Our VAD step simply assigns the estimated zero-energies to the non-active speech region and vice versa. Higher detection is obtained, as shown in Tab. 3.4 and Tab. 3.5 for both Gaussian and babble noise environments. SMM-NICA achieves a significant 99.4% correct decision for S_3 in the babble noise case with variance $\sigma_{\omega}^2 = 0.5$, see Tab. 3.5.

3.4 Robust Distributed Sparsity-Constrained Regularization Model-Based Multi-Speaker Voice Activity Detection for Speech Enhancement in Wireless Acoustic Sensor Networks

When the data departs from a nominal distribution, i.e. a well defined probability distribution with fixed set of parameters, applying robust methods may yield a better performance. The aim of robust signal processing is to design methods that are not unduly affected by modeling errors and outliers/heavy tailed noise. On the other hand, robust methods perform nearly optimal when the distributional assumptions are exactly fulfilled [125]. In this section[3], a new robust sparse-constrained VAD system for

[3]This section is based on the following conference article: "Robust Distributed Sparsity-Constrained Non-Negative Source Separation and Multi-Speaker Voice Activity Detection for Distributed Speech Enhancement in Wireless Acoustic Sensor Networks", Submitted to the Proc. 2nd IEEE Int. Conf. Signals Syst. (ICSigSys)

VAD results Case 1: Additive white Gaussian noise							
Variance	**Source**	**Method**	**Performance measure**				
			NRMSE	RMSE	ρ	ℓ_1-norm	ℓ_2-norm
$\sigma^2_{\boldsymbol{\omega}} = 0.1$	S_2	M-NICA	0.974	97.1	0.78	4.6×10^4	3.1×10^3
		SMeM-NICA	**0.972**	**0.97**	0.77	403.45	30.79
		SMM-NICA	**0.972**	**0.97**	**0.83**	**401.89**	**30.76**
	S_3	M-NICA	0.97	1.7×10^3	0.8	6×10^5	5.2×10^4
		SMeM-NICA	0.97	**0.97**	0.8	**321.46**	**30.8**
		SMM-NICA	0.97	**0.97**	0.8	321.78	**30.8**
$\sigma^2_{\boldsymbol{\omega}} = 0.5$	S_2	M-NICA	0.97	180.3	0.78	9.22×10^4	5.7×10^3
		SMeM-NICA	0.97	0.973	0.78	403.32	30.79
		SMM-NICA	0.97	**0.97**	**0.83**	**401.51**	**30.76**
	S_3	M-NICA	0.97	1.7×10^3	0.8	6.49×10^5	5.34×10^4
		SMeM-NICA	0.97	0.974	0.8	**321.47**	**30.8**
		SMM-NICA	0.97	**0.97**	0.8	321.79	**30.8**

Table 3.1: Comparison of the energy separation performance of the original M-NICA algorithm and the proposed approaches: the sparse mean-based M-NICA (SMeM-NICA), and the median-based M-NICA (SMM-NICA) for two sources (S_2 and S_3). Case 1: Additive white Gaussian noise of variance $\sigma^2_{\boldsymbol{\omega}} = 0.1$ and $\sigma^2_{\boldsymbol{\omega}} = 0.5$.

Method	**Variance**	**Source**	**Performance measure**				
			NRMSE	RMSE	ρ	ℓ_1-norm	ℓ_2-norm
MM-NICA	$\sigma^2_{\boldsymbol{\omega}} = 0.1$	S_2	99.8	10	0.78	4.69×10^3	317.35
		S_3	2.7×10^6	1.6×10^3	0.8	5.5×10^5	5.2×10^4
	$\sigma^2_{\boldsymbol{\omega}} = 0.5$	S_2	104.2	10.2	0.78	5.21×10^3	324.26
		S_3	2.7×10^6	1.6×10^3	0.8	5.52×10^5	5.20×10^4

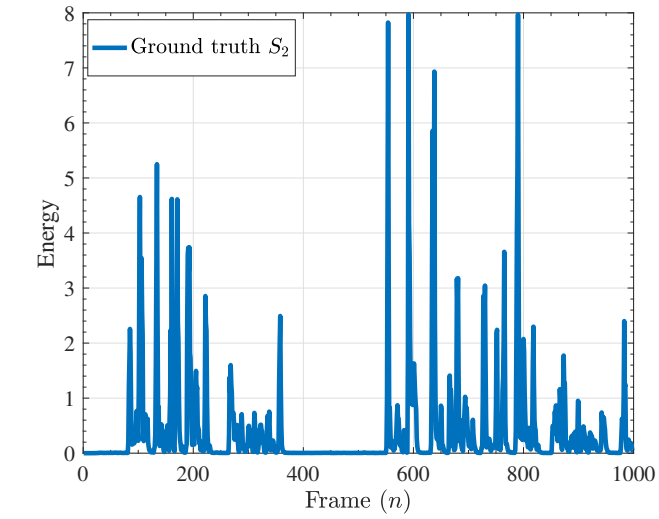
Table 3.2: Energy separation performance of the median-based M-NICA (MM-NICA) algorithm for two sources (S_2 and S_3) buried in additive white Gaussian noise of variance $\sigma^2_{\boldsymbol{\omega}} = 0.1$ and $\sigma^2_{\boldsymbol{\omega}} = 0.5$.

Case 2: Background babble noise							
Variance	**Source**	**Method**	**Performance measure**				
			NRMSE	RMSE	ρ	ℓ_1-norm	ℓ_2-norm
$\sigma^2_{\boldsymbol{\omega}} = 0.1$	S_2	M-NICA	0.974	10	0.78	4.7×10^3	316
		SMeM-NICA	**0.972**	0.974	**0.83**	**401.3**	**30.74**
		SMM-NICA	**0.972**	**0.973**	**0.83**	402.6	30.76
	S_3	M-NICA	0.974	1.6×10^3	0.8	5.5×10^5	5.2×10^4
		SMeM-NICA	**0.973**	**0.97**	**0.81**	**321.4**	**30.78**
		SMM-NICA	**0.973**	**0.97**	0.8	321.7	30.79
$\sigma^2_{\boldsymbol{\omega}} = 0.5$	S_2	M-NICA	0.973	78.1	0.78	4×10^4	2.5×10^3
		SMeM-NICA	**0.972**	**0.97**	**0.84**	**401**	**30.74**
		SMM-NICA	**0.972**	**0.97**	0.83	401.9	30.76
	S_3	M-NICA	0.97	1.7×10^3	0.8	6×10^5	5.3×10^4
		SMeM-NICA	0.97	**0.973**	0.8	**321.5**	30.8
		SMM-NICA	0.97	0.974	0.8	321.7	**30.79**

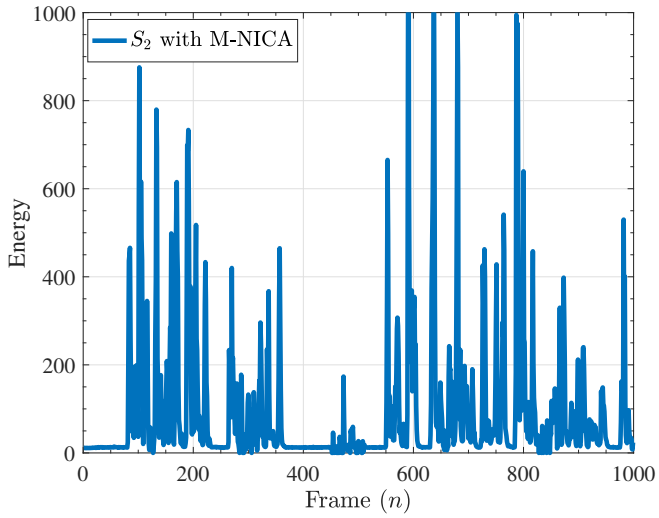
Table 3.3: Comparison of the energy separation performance of the original M-NICA algorithm and the proposed approaches: the sparse mean-based M-NICA (SMeM-NICA), and the median-based M-NICA (SMM-NICA) for two sources (S_2 and S_3). Case 2: Babble noise of variance $\sigma^2_{\boldsymbol{\omega}} = 0.1$ and $\sigma^2_{\boldsymbol{\omega}} = 0.5$.

VAD results Case 1: Additive white Gaussian noise				
Variance	**Source**	**VAD-based Methods**		
		M-NICA VAD (%)	SMeM-NICA VAD (%)	SMM-NICA VAD (%)
$\sigma^2_{\boldsymbol{\omega}} = 0.1$	S_2	63.3	**92.8**	**92.8**
	S_3	26.1	**82**	**82**
$\sigma^2_{\boldsymbol{\omega}} = 0.5$	S_2	62.9	**89.6**	**89.6**
	S_3	26.1	**81.4**	**81.4**

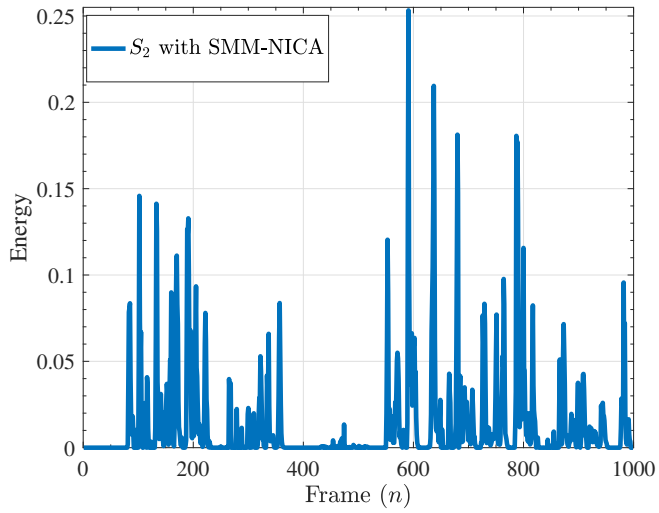
Table 3.4: Comparison of VAD performance of the original M-NICA algorithm and the proposed approaches: the sparse mean-based M-NICA (SMeM-NICA), and the sparse median-based M-NICA (SMM-NICA) for two sources (S_2 and S_3). Case 1: Additive white Gaussian noise of variance $\sigma^2_{\boldsymbol{\omega}} = 0.1$ and $\sigma^2_{\boldsymbol{\omega}} = 0.5$.



(a)

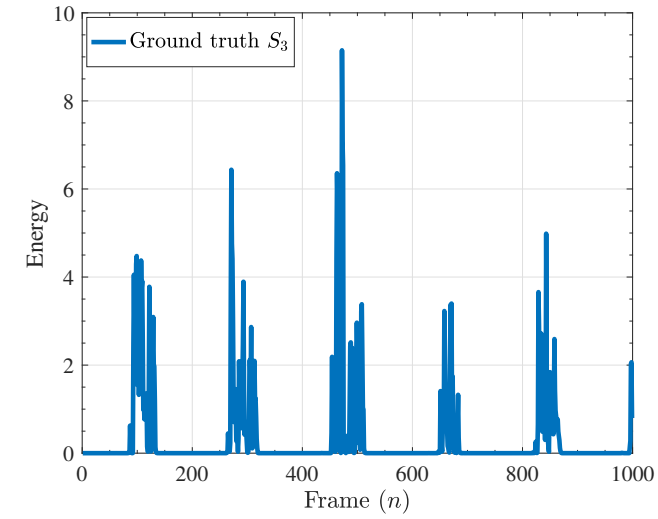


(b)

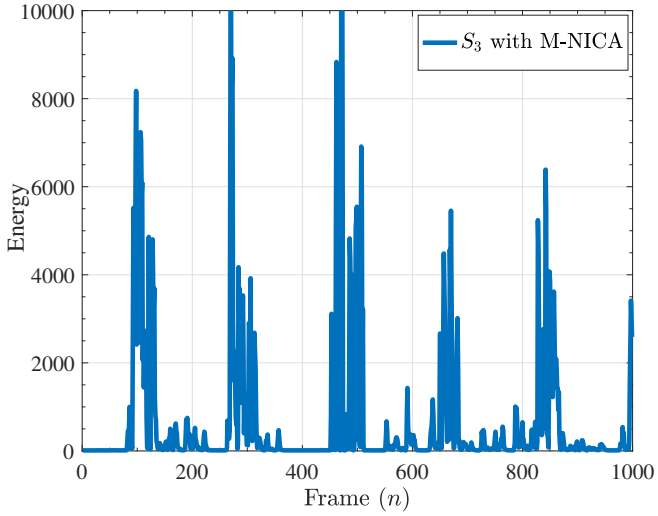


(c)

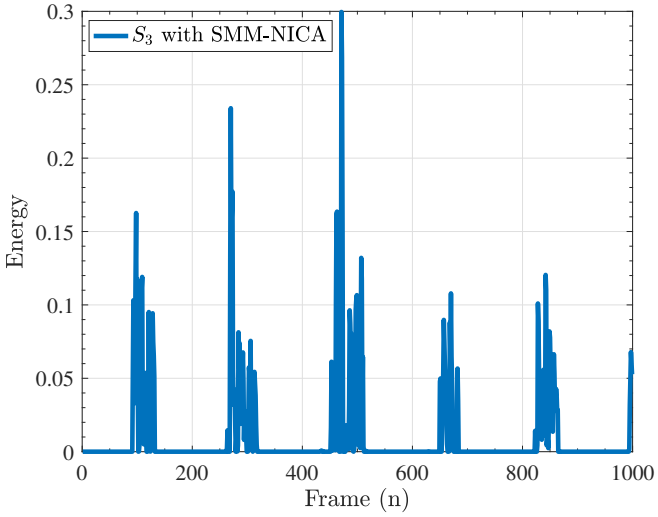
Figure 3.2: (a) Energy ground truth for the speech source S_2 of Fig. 1.2, (b) the corresponding energy estimates using the M-NICA algorithm, and (c) the energy estimates using the proposed sparse and median based multiplicative non-negative component analysis (SMM-NICA) approach, under additive white Gaussian noise with variance $\sigma_{\omega}^2 = 0.5$.



(a)



(b)



(c)

Figure 3.3: (a) Energy ground truth for the speech source S_3 of Fig. 1.2, (b) the corresponding energy estimates using the M-NICA algorithm, and (c) the energy estimates using the proposed sparse and median based multiplicative non-negative component analysis (SMM-NICA) approach, under additive white Gaussian noise with variance $\sigma_{\omega}^2 = 0.5$.

VAD results Case 2: Background babble noise				
Variance	**Source**	**VAD-based Methods**		
		M-NICA VAD (%)	SMeM-NICA VAD (%)	SMM-NICA VAD (%)
$\sigma^2_{\boldsymbol{\omega}} = 0.1$	S_2	63.8	92.9	**98.1**
	S_3	26.1	84.7	**99.3**
$\sigma^2_{\boldsymbol{\omega}} = 0.5$	S_2	62.7	88.2	**99.3**
	S_3	26.1	85.7	**99.4**

Table 3.5: Comparison of VAD performance of the original M-NICA algorithm and the proposed approaches: the sparse mean-based M-NICA (SMeM-NICA), and the sparse median-based M-NICA (SMM-NICA) for two sources (S_2 and S_3). Case 2: Additive background babble noise of variance $\sigma^2_{\boldsymbol{\omega}} = 0.1$ and $\sigma^2_{\boldsymbol{\omega}} = 0.5$.

speech enhancement is proposed that does not require any voiced/unvoiced modelling. However, multiple speech mixtures are well separated based on sparse modeling and robust decorrelation functions. Non-empirical thresholding functions for speech periods detection are introduced to improve the node-specific speech enhancement performance in a distributed WASN. Hence, a new robust VAD is designed to update the noise statistics in the subsequent speech enhancement system when faced to non-stationary noise. Extensive evaluation of the method on a multi-speaker setup under noisy conditions is carried out. Results show that the new method greatly improves the performance of the subsequent speech enhancement algorithm.

3.4.1 Proposed Robust Centralized VAD-Based Energy Source Separation Using a $t_\nu M$-SMM-NICA

In this section, the developed robust VAD technique is introduced and explained. We first look at the centralized VAD scheme. For this, we utilize the data model described by Eq. (2.3) presented in Section 2.1.2. In this case, every node $k = 1, \ldots, K$ is equipped with $M_k = 3$ microphones. The number of microphones at every node is the same and the overall number of microphones around the WASN is $M = \sum_{k=1}^{K} M_k$. The proposed $t_\nu M$-SMM-NICA algorithm starts with collecting the energy data matrix $\mathbf{Y}$ as shown in Algorithm 8, Step 1. The rows of this short-term energy matrix are associated to the observations recorded at the different microphones. We next define the empty data matrix $\mathbf{V}^{\mathcal{S}}$ where we collect iteratively the computed sparse rotations. The pre-processing of the proposed algorithm is accomplished using an iterative procedure. In other words, for every participating source q, an SVD layer extraction is first

calculated as given in Algorithm 8. Step 4. Following to that, a minimization of the optimization problem defined in Eq. (3.3) with respect to the variable $\mathbf{v}$ is performed, see Algorithm 8, Step 5. An update of the vector $\mathbf{v}$ related to source q with its sparse components is then performed and the matrix $\mathbf{V}^{\mathcal{S}}$ is revised to include the new sparse vector $\mathbf{v}$ as shown in Algorithm 8, Step 6. Then, an update of the scalar σ is also done based on the sparse vector $\mathbf{v}$ of source q. For a fixed $\mathbf{u}$, a sparse matrix $\mathbf{Y}_{\text{SSVD}}$ of lower-rank is calculated using Eq. (3.18) of Algorithm 8. This step is relevant to complete a matrix subtraction and deduce the new lower-rank matrix $\mathbf{Y}$ using Eq. (3.19). The initialization steps are iterated for every source q until reaching Q. Next, the features that we utilize for the remaining robust decorrelation steps are the ones collected in the sparse matrix $\mathbf{V}^{\mathcal{S}}$. An initial $[\mathbf{S}]_{q,n}$ then takes the absolute values of $\mathbf{V}^{\mathcal{S}}$ as illustrated in Eq. (3.20) of Algorithm 8. The covariance matrix based on the median central tendency is computed in Eq. (3.22) of Algorithm. 8. We use it for calculating the coefficients $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ that improve both the numerator and denominator of the decorrelation function in Eq. (3.25) of Algorithm 8. Moreover, we apply M-estimation in order to achieve a robust estimation of the covariance matrix. M-estimators are common methods of robust regression, which can be regarded as generalizations of the maximum-likelihood estimation. We propose using the $t_\nu M$-estimator to compute a robust $\mathbf{C}_*$ in Eq. (3.26) to minimize the correlation between the rows of $\mathbf{S} \in \mathbb{R}^{Q \times N}$. In Eq. (3.26), the vector $\mathbf{x}_n \in \mathbb{R}^{Q \times 1}$ is a Q-dimensional vector of elements $[\mathbf{S}]_{q,n}, q = [1, \ldots, Q]$ at a frame n. Moreover, $u_\nu(t)$ is a Q variate weight function.

3.4.1.1 Proposed Robust VAD Based on a Regularized $t_\nu M$-SMM-NICA for Energy Source Separation

Analogous to the $t_\nu M$-SMM-NICA presented in Algorithm 8, a regularized $t_\nu M$-SMM-NICA algorithm (R$t_\nu M$-SMM-NICA) is also proposed. The initialization step is identical to what we have introduced in Algorithm 8. However, the covariance matrix utilized in the computation of the decorrelation function is now being regularized. In this case, instead of $\mathbf{C}_*$, the matrix $\mathbf{C}_*^{\alpha,\beta}$ is computed using Eq. (3.38) of Algorithm 9. The regularization depends on the parameters α and β, which are usually chosen by cross-validation. The steps of the suggested R$t_\nu M$-SMM-NICA method are then described in Algorithm 9.

Algorithm 8 Proposed $t_\nu M$-SMM-NICA algorithm for energy separation

Input

$\mathbf{Y} = (\mathbf{y}[1], \cdots, \mathbf{y}[N]) \in \mathbb{R}_+^{M \times N}$ based on Eq. (2.3) of Section 2.1.2

$\mathbf{V}^{\mathcal{S}} \triangleq \emptyset$

Initialization

for $q = 1, \ldots, Q$ **do**

Extract a layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ that solves Eq. (3.2) of Section 3.3.1

Minimize Eq. (3.3) and update $\mathbf{v}$ using Eq. (3.7) of Section 3.3.1

Form the sparse matrix $\mathbf{V}^{\mathcal{S}} \triangleq \mathbf{V}^{\mathcal{S}} \parallel \mathbf{v}^\top$, with $\parallel$ being the concatenation symbol.

Update the singular value σ

$$\sigma = \mathbf{u}^\top \mathbf{Y} \mathbf{v} \tag{3.17}$$

Compose a sparse lower-rank matrix

$$\mathbf{Y}_{\text{SSVD}} = \sigma \mathbf{u} \mathbf{v}^\top \tag{3.18}$$

Matrix subtraction

$$\mathbf{Y} = \mathbf{Y} - \mathbf{Y}_{\text{SSVD}} \tag{3.19}$$

end for

Define

$$[\mathbf{S}]_{q,n} \leftarrow |[\mathbf{V}^{\mathcal{S}}]_{q,n}|, \forall q = [1, \ldots, Q], \forall n = [1, \ldots, N]. \tag{3.20}$$

repeat

$$\ddot{\mathbf{S}} = \underset{(q) \in Q}{\text{median}} \{\mathbf{S}_q\} \mathbf{1}_N^\top, \forall q = [1, \ldots, Q] \tag{3.21}$$

$$\mathbf{C}_S = (\mathbf{S} - \ddot{\mathbf{S}})(\mathbf{S} - \ddot{\mathbf{S}})^\top \tag{3.22}$$

$$\mathbf{\Lambda}_1 = D\{\mathbf{C}_S\} \tag{3.23}$$

$$\mathbf{\Lambda}_2 = D\{(\mathbf{\Lambda}_1^{-1} \mathbf{C}_S)^2\} \tag{3.24}$$

Minimize the correlation in $[\mathbf{S}]_{q,n}$

$$[\mathbf{S}]_{q,n} \leftarrow [\mathbf{S}]_{q,n} \left[\frac{\ddot{\mathbf{S}} \mathbf{S}^\top \mathbf{\Lambda}_1^{-1} \mathbf{S} + \mathbf{C}_* \mathbf{\Lambda}_1^{-1} \ddot{\mathbf{S}} + \mathbf{\Lambda}_2 \mathbf{S}}{\ddot{\mathbf{S}} \mathbf{S}^\top \mathbf{\Lambda}_1^{-1} \ddot{\mathbf{S}} + \mathbf{C}_* \mathbf{\Lambda}_1^{-1} \mathbf{S} + \mathbf{\Lambda}_2 \ddot{\mathbf{S}}} \right]_{q,n} \tag{3.25}$$

$$\mathbf{C}_* = \frac{1}{N} \sum_{n=1}^{N} u_\nu (\mathbf{x}_n^\top \hat{\mathbf{C}}_S^{-1} \mathbf{x}_n) \mathbf{x}_n \mathbf{x}_n^\top \tag{3.26}$$

$$u_\nu(t) = \frac{Q + \nu}{\nu + t} \tag{3.27}$$

$$t = \mathbf{x}_n^\top \hat{\mathbf{C}}_S^{-1} \mathbf{x}_n \tag{3.28}$$

until reaching a fixed-point of Eqs. (3.21)-(3.25)

Algorithm 9 Proposed source energy separation based on the R$t_\nu M$-SMM-NICA algorithm

Input

$\mathbf{Y} = (\mathbf{y}[1], \cdots, \mathbf{y}[N]) \in \mathbb{R}_+^{M \times N}$ based on Eq. (2.3) of Section 2.1.2

$\mathbf{V}^{\mathcal{S}} \triangleq \emptyset$

Initialization

for $q = 1, \ldots, Q$ **do**

Extract a layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ that solves Eq. (3.2)

Minimize Eq. (3.3) and update $\mathbf{v}$ using Eq. (3.7)

Construct the sparse matrix $\mathbf{V}^{\mathcal{S}} \triangleq \mathbf{V}^{\mathcal{S}} \parallel \mathbf{v}^\top$, with $\parallel$ being the concatenation symbol.

Set

$$\sigma = \mathbf{u}^\top \mathbf{Y} \mathbf{v} \tag{3.29}$$

Compose a sparse lower-rank matrix

$$\mathbf{Y}_{\text{SSVD}} = \sigma \mathbf{u} \mathbf{v}^\top \tag{3.30}$$

Matrix subtraction

$$\mathbf{Y} = \mathbf{Y} - \mathbf{Y}_{\text{SSVD}} \tag{3.31}$$

end for

Define

$$[\mathbf{S}]_{q,n} \leftarrow |[\mathbf{V}^{\mathcal{S}}]_{q,n}|, \forall q = [1, \ldots, Q], \forall n = [1, \ldots, N]. \tag{3.32}$$

repeat

$$\ddot{\mathbf{S}} = \underset{(q) \in Q}{\text{median}} \{\mathbf{S}_q\} \mathbf{1}_N^\top, \forall q = [1, \ldots, Q] \tag{3.33}$$

$$\mathbf{C}_S = (\mathbf{S} - \ddot{\mathbf{S}})(\mathbf{S} - \ddot{\mathbf{S}})^\top \tag{3.34}$$

$$\mathbf{\Lambda}_1 = D\{\mathbf{C}_S\} \tag{3.35}$$

$$\mathbf{\Lambda}_2 = D\{(\mathbf{\Lambda}_1^{-1} \mathbf{C}_S)^2\} \tag{3.36}$$

Minimize the correlation in $[\mathbf{S}]_{q,n}$

$$[\mathbf{S}]_{q,n} \leftarrow [\mathbf{S}]_{q,n} \left[\frac{\ddot{\mathbf{S}} \mathbf{S}^\top \mathbf{\Lambda}_1^{-1} \mathbf{S} + \mathbf{C}_*^{\alpha,\beta} \mathbf{\Lambda}_1^{-1} \ddot{\mathbf{S}} + \mathbf{\Lambda}_2 \mathbf{S}}{\ddot{\mathbf{S}} \mathbf{S}^\top \mathbf{\Lambda}_1^{-1} \ddot{\mathbf{S}} + \mathbf{C}_*^{\alpha,\beta} \mathbf{\Lambda}_1^{-1} \mathbf{S} + \mathbf{\Lambda}_2 \ddot{\mathbf{S}}} \right]_{q,n} \tag{3.37}$$

$$\mathbf{C}_*^{\alpha,\beta} = \beta \frac{1}{N} \sum_{n=1}^{N} u_\nu(\mathbf{x}_n^\top \hat{\mathbf{C}}_S^{-1} \mathbf{x}_n) \mathbf{x}_n \mathbf{x}_n^\top + \alpha \mathbf{I} \tag{3.38}$$

$$u_\nu(t) = \frac{Q + \nu}{\nu + t} \tag{3.39}$$

$$t = \mathbf{x}_n^\top \hat{\mathbf{C}}_S^{-1} \mathbf{x}_n \tag{3.40}$$

until reaching a fixed-point of Eqs. (3.33)-(3.37)

3.4.2 Proposed Distributed $t_\nu M$-SMM-NICA Algorithm for Energy Source Separation

In this section, we introduce a distributed $t_\nu M$-SMM-NICA (D$t_\nu M$-SMM-NICA) solution for separating increasing number of energy sources in a WASN. A D$t_\nu M$-SMM-NICA is achieved via applying the $t_\nu M$-SMM-NICA technique at distributed node clusters. The node clusters are formed using the LONAS technique [1]. Consequently, the steps of the D$t_\nu M$-SMM-NICA are illustrated in Algorithm 10. In Algorithm 10, since Eq. 3.49 reduces to one, the decorrelation function in Eq. 3.50 can be expressed by

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \mathbf{s}_q + c_* \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + \mathbf{s}_q}{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + c_* \grave{\lambda}_1^{-1} \mathbf{s}_q + \grave{\mathbf{s}}_q} \right]_n \tag{3.41}$$

3.4.2.1 Proposed Distributed Regularized $t_\nu M$-SMM-NICA

With regards to the distributed regularized $t_\nu M$-SMM-NICA (DR$t_\nu M$-SMM-NICA), the steps of Algorithm 9 are used at the formed sub-networks of nodes sharing the same unique source-of-interest. The resulting steps are summarized in Algorithm 11. In the same manner, since Eq. 3.63 of Algorithm 11 reduces to the value one, the proposed decorrelation function in Eq. 3.64 can be simplified to

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \mathbf{s}_q + c_*^{\alpha,\beta} \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + \mathbf{s}_q}{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + c_*^{\alpha,\beta} \grave{\lambda}_1^{-1} \mathbf{s}_q + \grave{\mathbf{s}}_q} \right]_n \tag{3.55}$$

3.4.3 Experimental Results

To assess the performance of the proposed robust VAD algorithms introduced in Section 3.4, the acoustic scenario depicted in Fig. 1.2 is simulated. Two different speech scenarios are investigated:

1. *Centralized speech use-case:* which consists of two interfering speech sources S_2 and S_3. For this less challenging setup, a centralized robust VAD algorithm,

Algorithm 10 The proposed distributed $t_\nu M$-SMM-NICA algorithm (D$t_\nu M$-SMM-NICA)

1: **for** $q = 1, \ldots, Q$ **do**

2: $\mathbf{Y}_{\mathcal{B}_q} = (\mathbf{y}_{\mathcal{B}_q}[1], \ldots, \mathbf{y}_{\mathcal{B}_q}[N]) \in \mathbb{R}_+^{(M_k \# (\mathcal{B}_q)) \times N}$ using Eq. (2.16).

3: $\mathbf{s}_q \triangleq \emptyset$

4: Extract a unique layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ that solves

$$\arg\min_{\sigma, \mathbf{u}, \mathbf{v}} \|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2, \tag{3.42}$$

5: Minimize the ℓ_1 penalized sum-of-squares regression for the source dominant model at node cluster $\mathcal{B}_q$ using

$$\|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2 + \lambda_{\mathbf{v}} \Phi(\sigma \mathbf{v}), \tag{3.43}$$

6: Update every indicator n of vector $\mathbf{v}$ at cluster $\mathcal{B}_q$ relative to source q using

$$v[n] = \frac{1}{\sigma} \left[\text{sgn} \left\{ [\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n \right\} \left(|[\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n| - \frac{\lambda_{\mathbf{v}}}{2} \right) \right], \tag{3.44}$$

7: Update the vector $\mathbf{s}_q$ with the values of the sparse vector $\mathbf{v}$ at source q using

$$\mathbf{s}_q \leftarrow \mathbf{v}. \tag{3.45}$$

8: Based on $|\mathbf{s}_q|$ compute

$$\grave{\mathbf{s}}_q = \underset{(q) \in Q}{\text{median}} \{\mathbf{s}_q\} \mathbf{1}_N^\top, \tag{3.46}$$

$$c_{\mathbf{s}_q} = (\mathbf{s}_q - \grave{\mathbf{s}}_q)(\mathbf{s}_q - \grave{\mathbf{s}}_q)^\top \tag{3.47}$$

$$\grave{\lambda}_1 = c_{\mathbf{s}_q} \tag{3.48}$$

$$\grave{\lambda}_2 = (\grave{\lambda}_1^{-1} c_{\mathbf{s}_q})^2 \tag{3.49}$$

9: Minimize the correlation function using

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \mathbf{s}_q + c_* \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + \grave{\lambda}_2 \mathbf{s}_q}{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + c_* \grave{\lambda}_1^{-1} \mathbf{s}_q + \grave{\lambda}_2 \grave{\mathbf{s}}_q} \right]_n \tag{3.50}$$

$$c_* = \frac{1}{N} \sum_{n=1}^{N} u_\nu(x_n \hat{c}_{\mathbf{s}_q}^{-1} x_n) x_n^2 \tag{3.51}$$

$$u_\nu(t) = \frac{1 + \nu}{\nu + t} \tag{3.52}$$

$$t = x_n^\top \hat{c}_{\mathbf{s}_q}^{-1} x_n \tag{3.53}$$

10: Update the sparse vector $\mathbf{s}_q$ using

$$[\mathbf{s}_q]_n \leftarrow [\mathbf{s}_q^*]_n, \quad \forall n \in N \tag{3.54}$$

11: Extract the speaker-specific zero-threshold VAD patterns for the current observations $\mathbf{Y}_{\mathcal{B}_q}$ based on the computed sparse vectors $\mathbf{s}_q$.

12: **end for**

Algorithm 11 Proposed Distributed Regularized $t_\nu M$-SMM-NICA (DR$t_\nu M$-SMM-NICA)

1: **for** $q = 1, \ldots, Q$ **do**

2: $\mathbf{Y}_{\mathcal{B}_q} = (\mathbf{y}_{\mathcal{B}_q}[1], \ldots, \mathbf{y}_{\mathcal{B}_q}[N]) \in \mathbb{R}_+^{(M_k \#(\mathcal{B}_q)) \times N}$ using Eq. (2.16).

3: $\mathbf{s}_q \triangleq \emptyset$

4: Extract a unique layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ that solves

$$\underset{\sigma,\mathbf{u},\mathbf{v}}{\arg\min} \|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2, \tag{3.56}$$

5: Minimize the ℓ_1 penalized sum-of-squares regression for the source dominant model at node cluster $\mathcal{B}_q$ using

$$\|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2 + \lambda_{\mathbf{v}} \Phi(\sigma \mathbf{v}), \tag{3.57}$$

6: Update every indicator n of vector $\mathbf{v}$ at cluster $\mathcal{B}_q$ relative to source q using

$$v[n] = \frac{1}{\sigma} \left[\text{sgn} \left\{ [\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n \right\} \left(|[\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n| - \frac{\lambda_{\mathbf{v}}}{2} \right) \right], \tag{3.58}$$

7: Update the vector $\mathbf{s}_q$ with the values of the sparse vector $\mathbf{v}$ at source q using

$$\mathbf{s}_q \leftarrow \mathbf{v}. \tag{3.59}$$

8: Based on $|\mathbf{s}_q|$ compute

$$\grave{\mathbf{s}}_q = \underset{(q) \in Q}{\text{median}} \{\mathbf{s}_q\} \mathbf{1}_N^\top, \tag{3.60}$$

$$c_{\mathbf{s}_q} = (\mathbf{s}_q - \grave{\mathbf{s}}_q)(\mathbf{s}_q - \grave{\mathbf{s}}_q)^\top \tag{3.61}$$

$$\grave{\lambda}_1 = c_{\mathbf{s}_q} \tag{3.62}$$

$$\grave{\lambda}_2 = (\grave{\lambda}_1^{-1} c_{\mathbf{s}_q})^2 \tag{3.63}$$

9: Minimize the correlation using

$$[\mathbf{s}_q^*]_n \leftarrow [\mathbf{s}_q]_n \left[\frac{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \mathbf{s}_q + c_*^{\alpha,\beta} \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + \grave{\lambda}_2 \mathbf{s}_q}{\grave{\mathbf{s}}_q \mathbf{s}_q^\top \grave{\lambda}_1^{-1} \grave{\mathbf{s}}_q + c_*^{\alpha,\beta} \grave{\lambda}_1^{-1} \mathbf{s}_q + \grave{\lambda}_2 \grave{\mathbf{s}}_q} \right]_n \tag{3.64}$$

$$c_*^{\alpha,\beta} = \frac{1}{N} \sum_{n=1}^{N} u_\nu (x_n \hat{c}_{\mathbf{s}_q}^{-1} x_n) x_n^2 \tag{3.65}$$

$$u_\nu(t) = \frac{1+\nu}{\nu+t} \tag{3.66}$$

$$t = x_n^\top \hat{c}_{\mathbf{s}_q}^{-1} x_n \tag{3.67}$$

10: Update the sparse vector $\mathbf{s}_q$ using

$$[\mathbf{s}_q]_n \leftarrow [\mathbf{s}_q^*]_n, \quad \forall n \in N \tag{3.68}$$

11: Extract the speaker-specific zero-threshold VAD patterns for the current observations $\mathbf{Y}_{\mathcal{B}_q}$ based on the computed sparse vectors $\mathbf{s}_q$.

12: **end for**

namely $t_\nu M$-SMM-NICA, is considered. The loudspeaker S_2 produces the desired speech signal, consisting of a female speaking English announcements at an airport with regions of silence between every two subsequent sentences. Source S_3 consists of a male speaking English sentences.

2. *Distributed speech-use-case:* In the second case, an increased number of speech sources is considered. A multi-speaker speech scenario consisting of six speech sources is used. The considered sources are S_2, S_4, S_5, S_6, S_7, and S_3. Since we look at a language independent VAD application, the studied speech sources are in different languages. In addition, the delivered messages were spoken by different genders.

In both cases, the mixture of speech is corrupted with an interfering additive Gaussian noise. In the studied setup, we consider 20 nodes, each having $M_k = 3$ microphones that are placed 1.5 cm apart. The microphone signals are sampled at a sampling frequency of $f_s = 16$kHz.

To validate the proposed VAD algorithms, we further consider applying the DANSE$_1$ algorithm, see [18], for local node speech enhancement. The outcome of DANSE$_1$ highly depends on the VAD input. The validation of the proposed VAD techniques uses the following parameters for the DANSE$_1$ algorithm: a weighted overlap-add (WOLA) with a DFT size of $L_{\text{DFT}} = 512$, a forgetting factor $\lambda_{\text{DANSE}} = 0.997$, and a step size parameter $\mu = 5$ to improve noise reduction. Usually, an ideal VAD is exploited in order to isolate the influence of VAD errors. In our experiments, the VAD input is the one estimated using the proposed robust VAD algorithms. It is to mention that the performed simulations, related to speech enhancement based on robust VAD inputs, are completed in a static scenario where the speech sources do not move.

3.4.3.1 Centralized Use-Case

In a first experiment, we consider a speech scenario consisting of two active sources S_2 and S_3 from Fig. 1.2. The proposed $t_\nu M$-SMM-NICA algorithm is applied to the mixture of these two sources corrupted with additive white Gaussian noise of variance $\sigma_\omega^2 = 0.01$. Table 3.6 depicts the VAD results in terms of correct detection (CD), false alarm rate (FA), and misdetection percentage (MD). It is demonstrated that, for this realistic centralized situation, the robustness parameter ν plays a crucial role to obtain good results. Hence, the robustness parameter ν is object of analysis. Figure 3.4 illustrates the speech detection performance that varies with distinct robustness degree

of freedom ν. Meanwhile, we notice that stability in speech detection is reached when ν is 249 regarding S_2. However, a robust speech activity decision for S_3 is accomplished when $\nu = 55$.

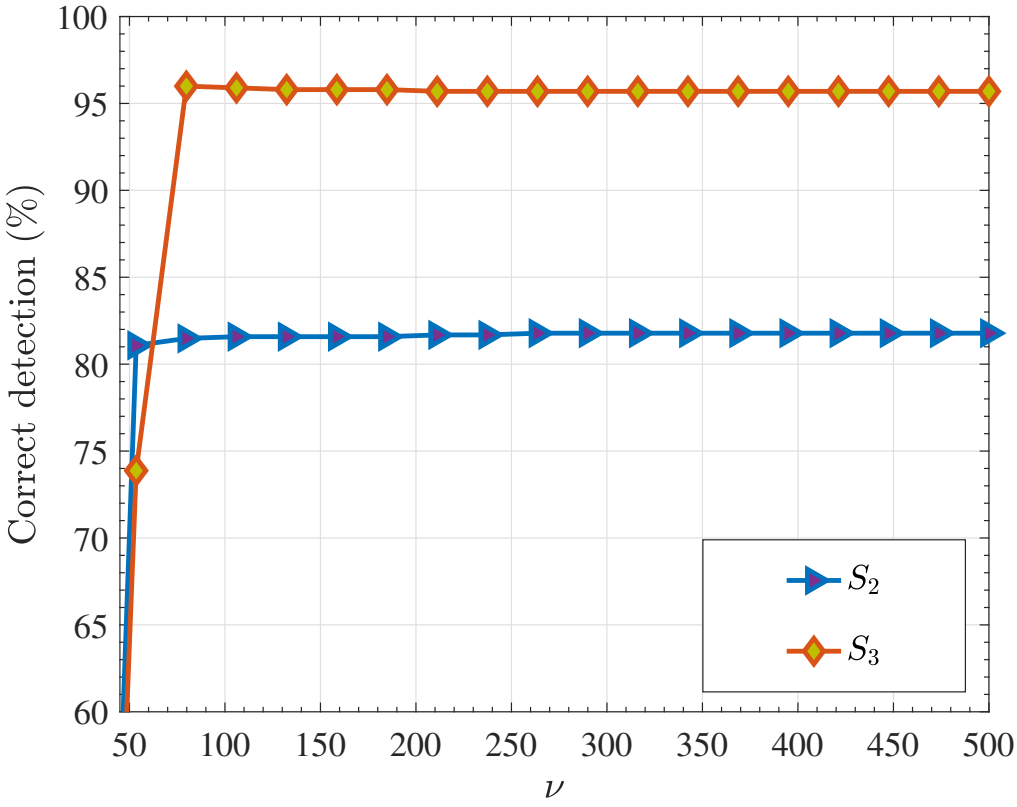


Figure 3.4: Correct detection achievement with varying degree of freedom ν for the robustness parameter applied in the $t_\nu M$-SMM-NICA speech separation and activity detection algorithm.

In Tab. 3.6, we summarize the detection results for the best selected ν values from Fig. 3.4. The developed centralized $t_\nu M$-SMM-NICA-based VAD attains 81.8% of correct decision with a minimum midetection rate of 0.4%. Minimizing the misdetection rate in the VAD procedure is valuable for a subsequent speech enhancement phase. Higher MD levels certainly affects the speech quality. A significant likelihood of misdetection of weak energy speech components as noise will cause distortion in the filtered noisy speech signal. As presented by Tab. 3.6, our proposed robust detector based on the suggested $t_\nu M$-SMM-NICA algorithm for speech separation secures a high energy detection rate for S_3 with 96.1% of maximal correct decision and a minimum MD of 1.7% when $\nu = 55$. Since the proposed $t_\nu M$-SMM-NICA is meant primarily for multiple energy speech separation based on a sparsity promoting model, the source to interference ratio (SIR) and the signal to distortion ratio (SDR) are two measures used for quantifying the quality of the separated energy speech signals as given in Tab. 3.6.

Method	Source	ν	Measures				
			CD (%)	MD (%)	FA (%)	SIR	SDR
$t_\nu M$-SMM-NICA	S_2	249	81.8	17.8	0.4	277.2	3.75
	S_3	55	96.1	2.2	1.7	280.34	5.68
R$t_\nu M$-SMM-NICA	S_2	249	83.3	14.9	1.8	279.9	6.7
	S_3	55	95.9	2.2	1.9	280.4	5.75
M-NICA	S_2	-	62.4	1.9	35.73	349.6	-1.4
	S_3	-	54.65	1.1	44.24	374.5	5.5

Table 3.6: Energy separation results and detection performance in the centralized use-case of a two-energy mixture using the $t_\nu M$-SMM-NICA, the R$t_\nu M$-SMM-NICA, and the standard M-NICA algorithm.

For the same energy use-case scenario, we measure the achievement of the suggested R$t_\nu M$-SMM-NICA in Tab. 3.6. For a fixed regularization parameter $\beta = 1$ and a value of $\alpha > 1$, higher VAD is observed for S_2, see Tab. 3.6. We compare the results of the proposed $t_\nu M$-SMM-NICA and the R$t_\nu M$-SMM-NICA to the M-NICA-based energy separation and VAD, summarized in Tab. 3.6. Clearly, the proposed $t_\nu M$-SMM-NICA and R$t_\nu M$-SMM-NICA algorithms obtain higher CD as compared to the M-NICA based detector in a centralized setup. It can be also viewed that improved SDR values are reached when separating the mixture of energy signatures using the proposed approaches as compared the the standard M-NICA. The latter is more pronounced when S_2 is estimated using M-NICA with a low negative valued SDR of -1.38. In contrast, a finer SDR for the source S_2 is realized using the robust $t_\nu M$-SMM-NICA and R$t_\nu M$-SMM-NICA approaches for energy mixtures separation. The same interpretation applies for the results achieved for S_3. The CD values obtained using the original M-NICA algorithm for S_3 are substantially lower as compared to those given by the robust and sparse $t_\nu M$-SMM-NICA and R$t_\nu M$-SMM-NICA techniques. In the following, Fig. 3.5 and Fig. 3.6 show the quality of the estimated unmixed energies of the noisy energy mixture of the two speech sources S_2 and S_3. The sparsity-integrated model causes low-valued energies representing noise to converge explicitly to null. This fact makes the ensuing VAD task easier to tackle and reduces only to counting non-zero valued energies to be labeled as active signatures.

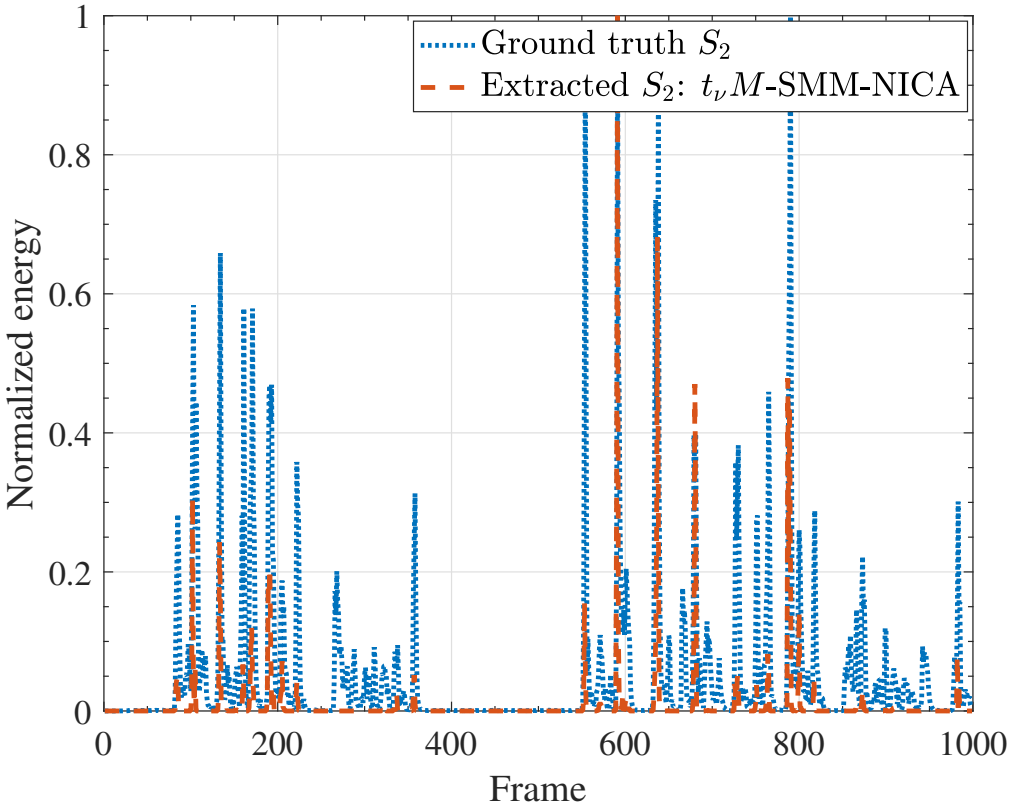


Figure 3.5: Unmixed sparse energy of source S_2 using the proposed $t_\nu M$-SMM-NICA in a centralized setup.

3.4.3.2 Validation of the Centralized $t_\nu M$-SMM-NICA with DANSE$_1$

In this section, we evaluate the assessment of the proposed robust VAD algorithms based on a node-specific signal estimation method, namely DANSE$_1$ [18]. For this, we aim at enhancing the signal at a single node when the VAD information is estimated using our detection techniques in a centralized setup. We calculate the SDR and SIR of the output enhanced signals of the sources S_2 and S_3 compared to their ground truth states. The cases when the VAD input relies on M-NICA is compared to that which relies on the proposed $t_\nu M$-SMM-NICA method. Results are summarized in Tab. 3.7 and Tab. 3.8. In Tab. 3.7, the on-off VAD regions are generated using the M-NICA algorithm. The on-off speech regions based on the proposed robust and sparse $t_\nu M$-SMM-NICA are used to generate Tab. 3.8. Clearly, in the examined centralized two-source mixture scenario contaminated with additive Gaussian noise of variance $\sigma_\omega^2 = 0.01$, high quality signals are always achieved when DANSE$_1$ uses the VAD input generated with our proposed $t_\nu M$-SMM-NICA algorithm. This is mainly noticed for the speech signal S_2 that is estimated with SDR of -9.62 for the M-NICA input, while the signal quality improves to an SDR of 8.23 when the robust VAD based on the

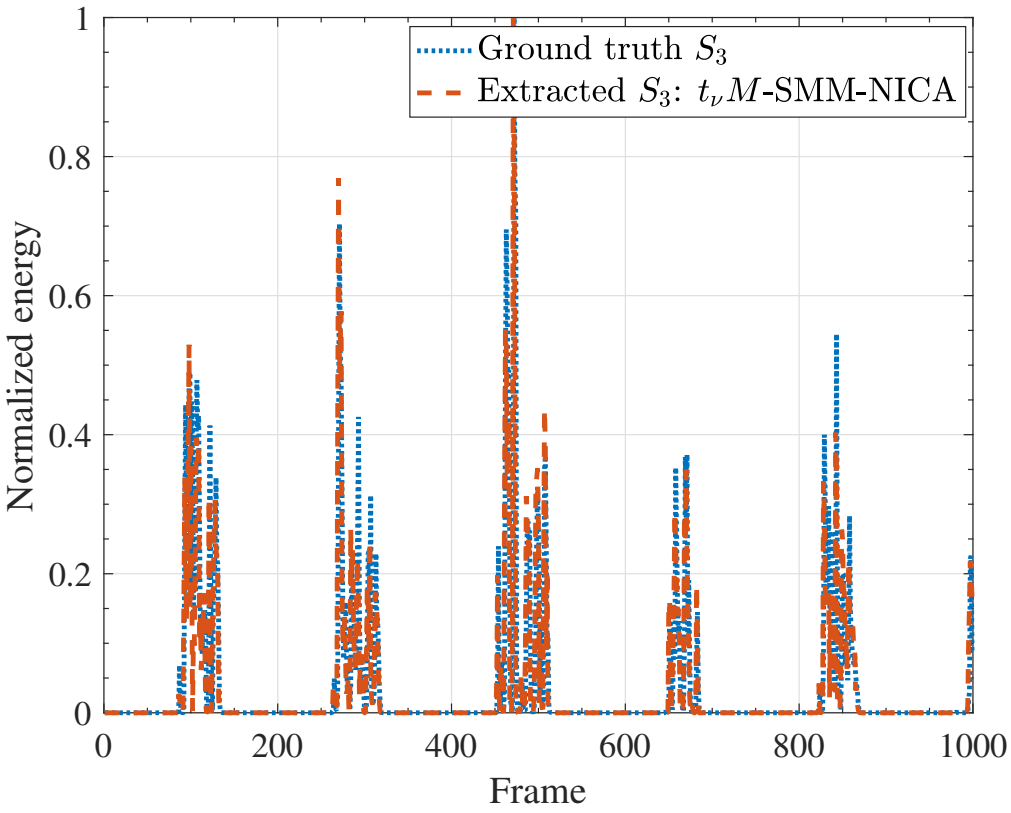


Figure 3.6: Unmixed sparse energy of source S_3 using the proposed $t_\nu M$-SMM-NICA in a centralized setup.

suggested $t_\nu M$-SMM-NICA algorithm is used.

Noise Variance	Source	Measures	
		SIR	SDR
0.01	S_2	287.1063	-9.6207
	S_3	312.737	6.9277

Table 3.7: Node-specific time-domain speech signals estimation in the centralized use-case of two sources S_2 and S_3 using the M-NICA-based VAD algorithm.

3.4.3.3 Distributed Use-Case

In this second part of the experiments, we consider evaluating the accomplishment of the proposed VAD based on a robust distributed $t_\nu M$-SMM-NICA (D$t_\nu M$-SMM-NICA). For this reason, we compute the detection performance, the energy separation results and apply the resulting VAD patterns to speech enhancement based on DANSE$_1$

Noise Variance	Source	Measures	
		SIR	SDR
0.01	S_2	287.8674	8.2389
	S_3	313.6354	26.2240

Table 3.8: Node-specific time-domain DANSE_1-based speech enhancement in the centralized use-case of two sources using the robust $t_\nu M$-SMM-NICA-based VAD algorithm.

algorithm. The D$t_\nu M$-SMM-NICA approach is based on node clustering achieved by LONAS [1] where we utilize the distributed well-clustered nodes sharing the same unique source-of-interest to produce a cluster (relative to a source) based VAD pattern. This means that the D$t_\nu M$-SMM-NICA computes the VAD output for a specific source based on a cluster of nodes rather than the whole set of nodes existing in the WASN. This framework is scalable to a large number of sources, see Chapter 2. Table 3.9 outlines the VAD results and the energy separation quality for the distributed speech use-case consisting of six speech sources. We observe that the values of the degree of freedom ν in Tab. 3.9 are smaller compared to the centralized use-case analyzed earlier.

Method	Source	ν	Measures				
			CD (%)	MD (%)	FA (%)	SIR	SDR
D$t_\nu M$-SMM-NICA	S_2	5	74.6	19.72	5.7	280.53	7.23
	S_4	10	83.5	16.5	0	279.37	4.38
	S_5	10	79.9	20.7	0	279.36	4.47
	S_6	15	88.9	2.2	8.9	280.19	6.93
	S_7	10	87.19	2.5	10.3	279.26	2.28
	S_3	15	89.8	10.21	0	279.77	4.52
M-NICA	S_2	-	60.76	6	33.23	358.1533	-55.7288
	S_4	-	46.84	3	50.15	447.1549	-9.4796
	S_5	-	56.96	3.9	39.14	381.0957	-34.5684
	S_6	-	55.85	6.4	37.73	397.0846	-14.5258
	S_7	-	45.74	5	49.25	422.5848	-34.5201
	S_3	-	46.5	2.8	50.6	375.6884	-20.2579

Table 3.9: Detection results for the distributed use-case of six source mixture corrupted with additive Gaussian noise of variance 0.01 using the D$t_\nu M$-SMM-NICA algorithm.

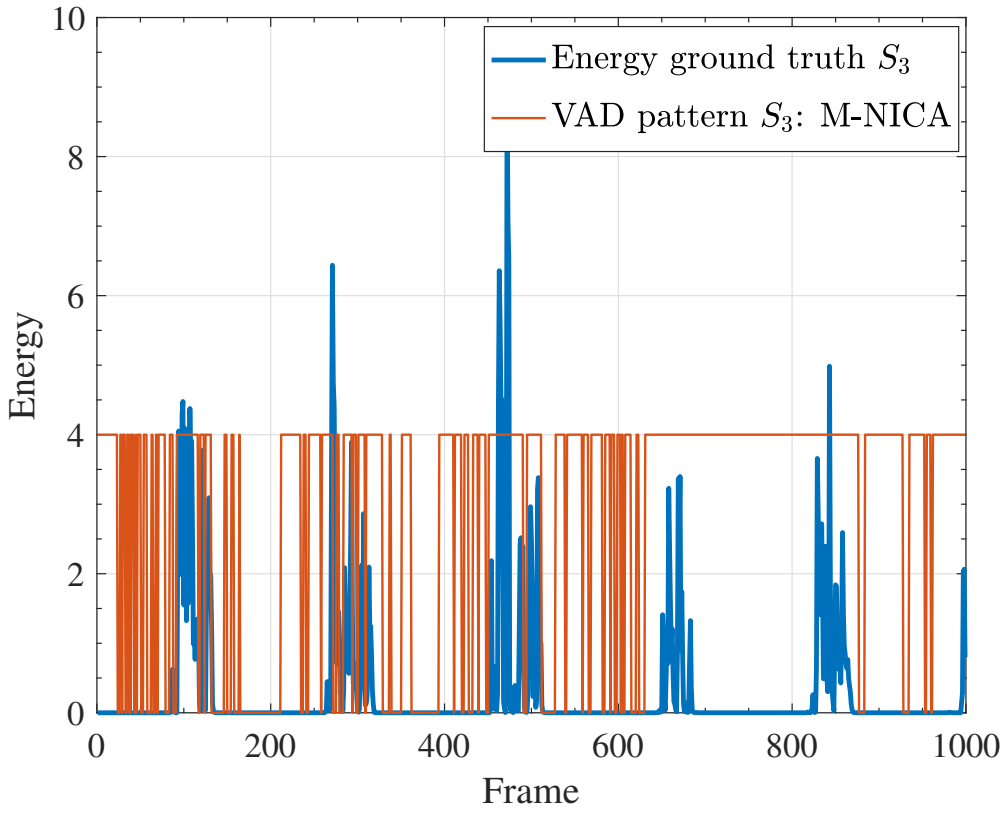
Compared to the D$t_\nu M$-SMM-NICA in Tab. 3.9, the energy separation capability and the detection results are significantly low when the M-NICA algorithm is used for this challenging scenario of six simultaneous speakers. An example of the VAD pattern (in red) estimated using M-NICA and the D$t_\nu M$-SMM-NICA algorithms are shown in Fig. 3.7. (a) and Fig. 3.7. (b), respectively. Moreover, we give an idea of the generated VAD patterns with the proposed robust VAD technique D$t_\nu M$-SMM-NICA in time-domain for the speech source S_3 in Fig. 3.8. (a). The latter can be compared to the time-domain VAD output for source S_3 produced using the original M-NICA in Fig. 3.8. (b).

3.4.3.4 Validation of the D$t_\nu M$-SMM-NICA Based on Speech Enhancement Results Using DANSE$_1$

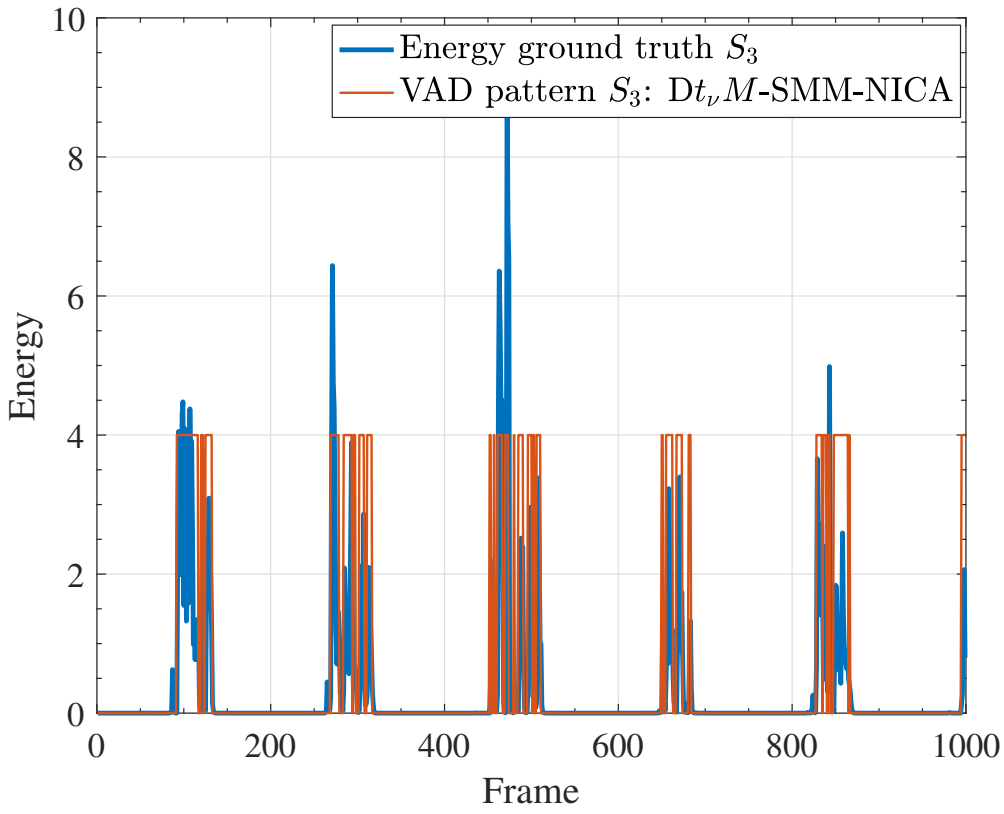
In this section, we investigate the performance of the proposed robust distributed VAD algorithm, namely D$t_\nu M$-SMM-NICA, based on the speech enhancement performance. To this end, the obtained VAD results communicated in the forth-mentioned section are subject to be used as an input for a subsequent DANSE$_1$-based speech enhancement. In Tab. 3.10, we report on the quality of the produced signals using the SIR and SDR measures when the D$t_\nu M$-SMM-NICA is employed. This is compared to the Tab. 3.11 that collects the achievement in terms of speech enhancement when the standalone M-NICA-based VAD is utilized. Again, the speech enhancement for a challenging number of sources is well obtained by the DANSE$_1$ algorithm that uses the D$t_\nu M$-SMM-NICA for estimating the multiple-speaker on-off regions.

Noise Variance	Source	Measures	
		SIR	SDR
0.01	S_2	289.1388	5.9283
	S_4	306.1935	-1.0096
	S_5	310.5850	25.0671
	S_6	313.9891	15.7383
	S_7	275.1634	8.0664
	S_3	313.2935	20.2881

Table 3.10: Node-specific speech enhancement results based on the robust distributed VAD input, or D$t_\nu M$-SMM-NICA, applied on a noisy mixture scenario of 6 sources, namely $\{S_2, \ldots, S_7\}$ of Fig. 1.2.

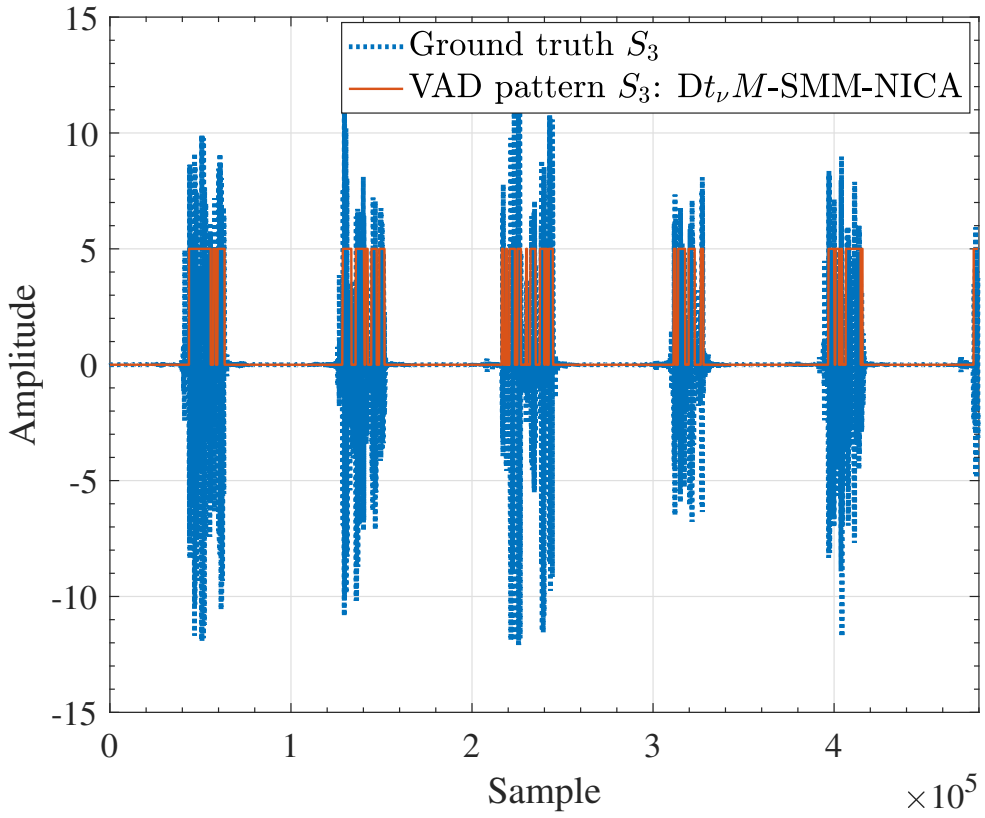


(a)

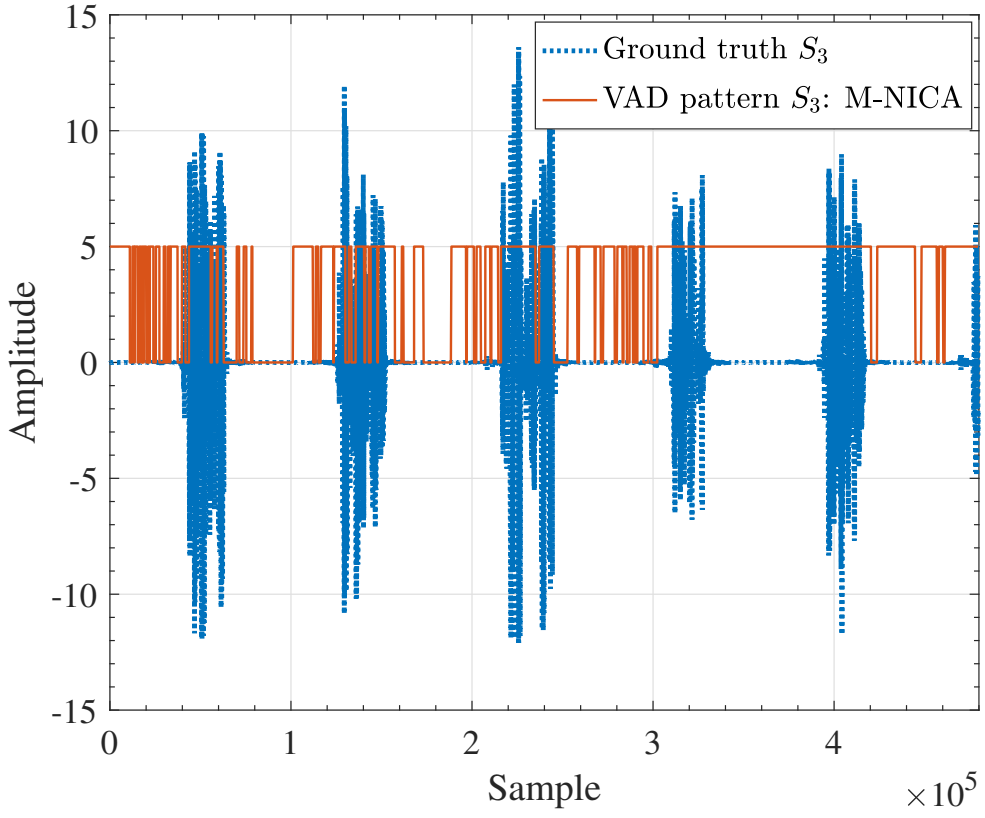


(b)

Figure 3.7: Estimated VAD pattern for the energy signature S_3 in the 6 source scenario use-case, using M-NICA in (a) and D$t_\nu M$-SMM-NICA in (b).



(a)



(b)

Figure 3.8: Time-domain VAD pattern of source S_3 in the 6 source scenario using: (a) the proposed robust D$t_\nu M$-SMM-NICA algorithm, and (b) the original M-NICA-based VAD.

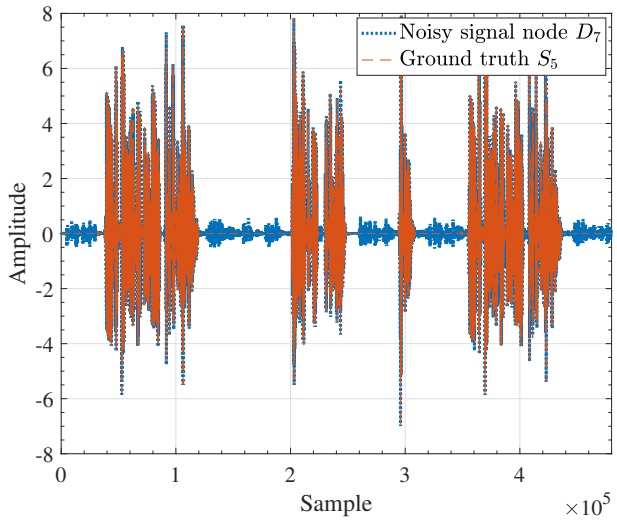
Noise Variance	Source	Measures	
		SIR	SDR
0.01	S_2	269.9669	-12.0037
	S_4	305.8562	2.0281
	S_5	305.6178	0.3762
	S_6	312.5591	-3.7797
	S_7	278.3810	3.9074
	S_3	310.0795	-5.4394

Table 3.11: Node-specific speech enhancement results with a M-NICA-based VAD input applied in a mixed scenario of 6 sources, namely $\{S_2, \ldots, S_7\}$ of Fig. 1.2.

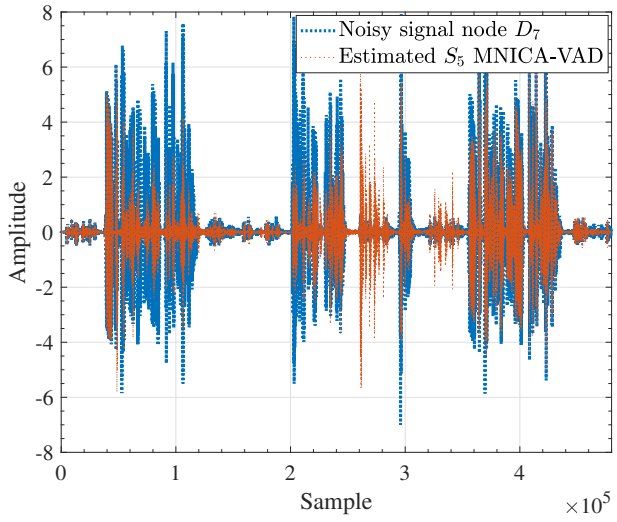
In the following, we show the improvement in speech enhancement using DANSE_1 when the proposed D$t_\nu M$-SMM-NICA is used for VAD as compared to M-NICA. For instance, Fig. 3.9. (a) displays the noisy received signal at Node D_7 (in blue) while the ground truth signal corresponding to S_5 is in red. DANSE_1 considers estimating a unique desired source-of-interest. In Fig. 3.9, the signal of interest is S_5. The remaining sources are interfering sources. We utilize this strategy of estimating single source-of-interest for every different source in the distributed use-case. Figure 3.9. (b) illustrates the speech estimation results of Source S_5 at Node D_7 when the M-NICA-based VAD is favored for the DANSE_1-based speech enhancement task. However, Fig. 3.9. (c) shows the accuracy of the speech enhancement for source S_5 when the VAD input for DANSE_1 is the one generated with our proposed D$t_\nu M$-SMM-NICA. In the same way, we show the different enhancement results for S_6, S_7, and S_3 in Fig. 3.10, Fig. 3.11, and Fig. 3.12, respectively. It is clear that the best achievement in terms of speech enhancement is the one observed at nodes utilizing the proposed D$t_\nu M$-SMM-NICA for estimating the VAD patterns.

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

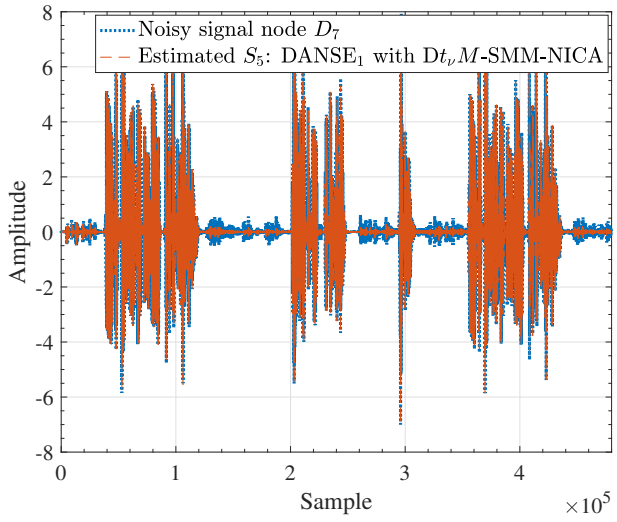
As discussed previously, the non-negativity constraint, naturally, leads to sparse signal representation in the underdetermined speech separation. However, we have shown in Section 3.3 and Section 3.4 that using an explicit Lasso-based sparse model gen-



(a)

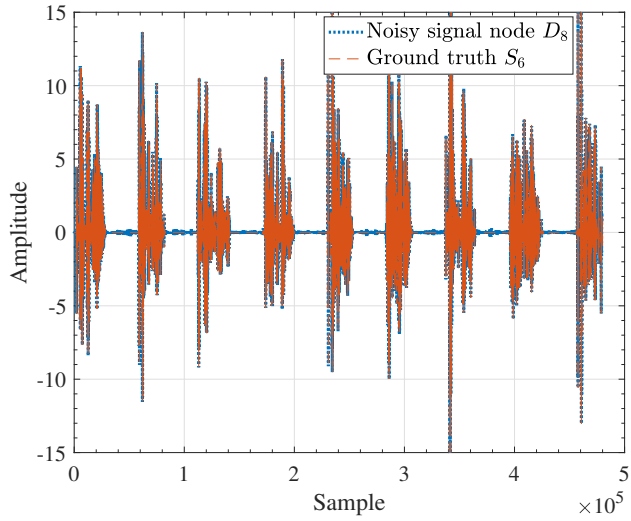


(b)

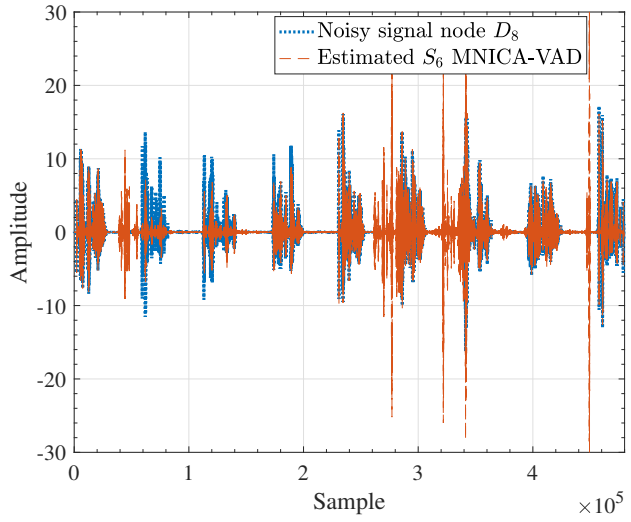


(c)

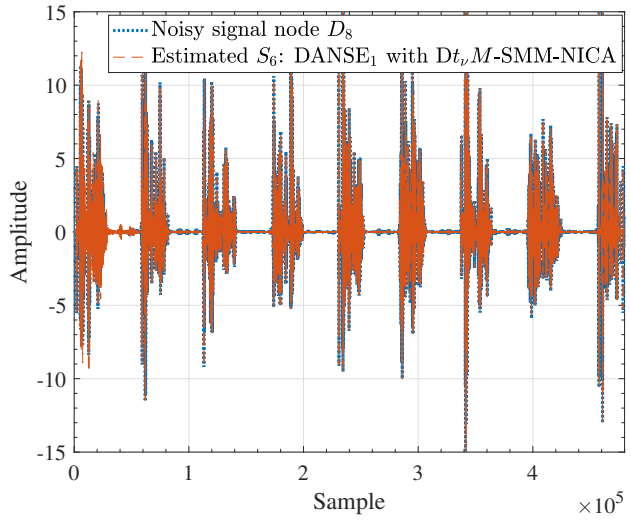
Figure 3.9: Comparison between the received speech signal at Node D_7 and (a) the ground truth source signal S_5 in red, (b) the estimated speech signal using DANSE$_1$ based on the VAD output of M-NICA, and (c) the estimated signal S_5 using DANSE$_1$ with D$t_\nu M$-SMM-NICA for VAD.



(a)

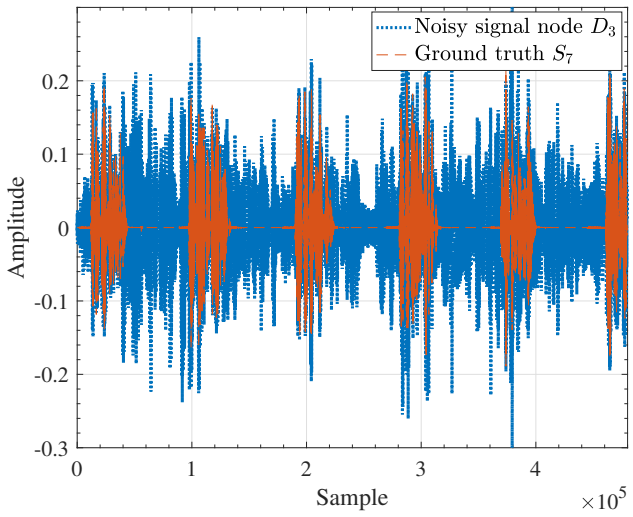


(b)

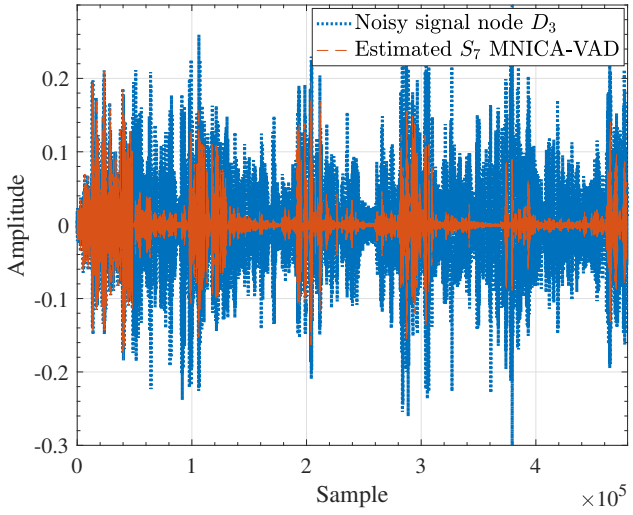


(c)

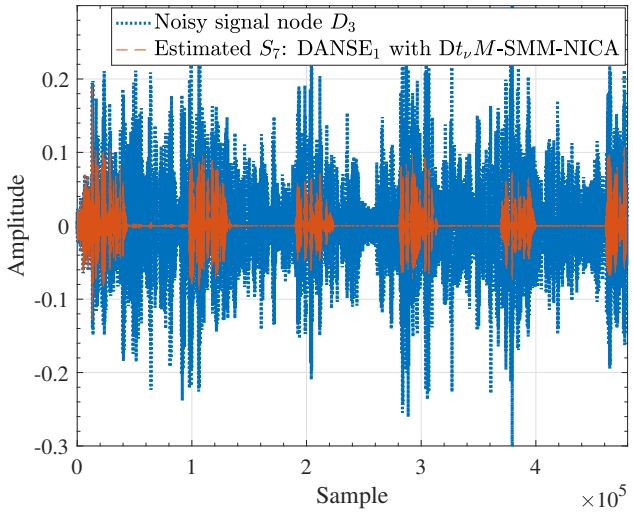
Figure 3.10: Comparison between the received speech signal at Node D_8 and (a) the ground truth source signal S_6 in red, (b) the estimated speech signal using DANSE$_1$ based on the VAD output of M-NICA, and (c) the estimated signal S_6 using DANSE$_1$ with D$t_\nu M$-SMM-NICA for VAD.



(a)

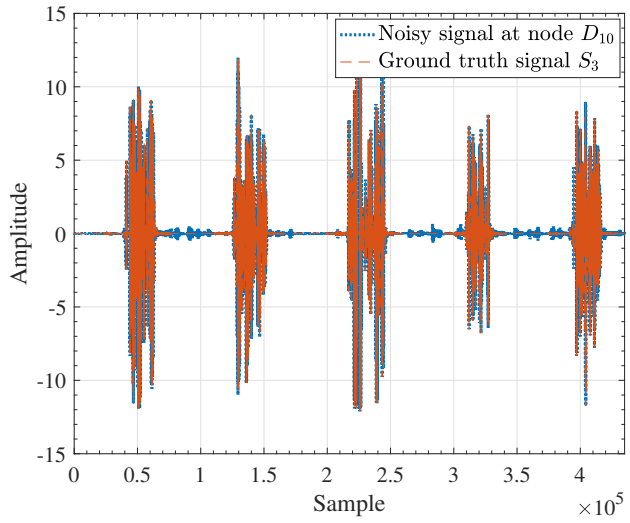


(b)

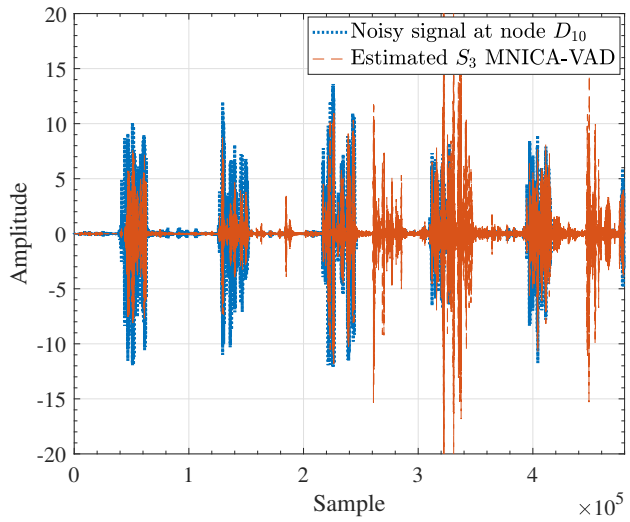


(c)

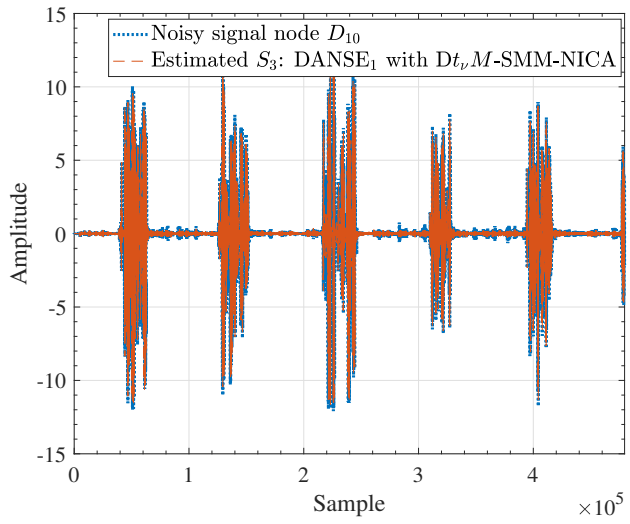
Figure 3.11: Comparison between the received speech signal at Node D_3 and (a) the ground truth source signal S_7 in red, (b) the estimated speech signal using DANSE$_1$ based on the VAD output of M-NICA, and (c) the estimated signal S_7 using DANSE$_1$ with D$t_\nu M$-SMM-NICA for VAD.



(a)



(b)



(c)

Figure 3.12: Comparison between the received speech signal at Node D_{10} and (a) the ground truth source signal S_3 in red, (b) the estimated speech signal using DANSE$_1$ based on the VAD output of M-NICA, and (c) the estimated signal S_3 using DANSE$_1$ with Dt$_\nu$M-SMM-NICA for VAD.

erates advantageous features for computationally efficient multiplicative algorithms. There has been a flourishing interest in exploring the study of sparse approximation of signals. Ubiquitous applications in speech and audio processing areas utilize sparse representations of signals including feature extraction, source separation, compressive sensing, speech enhancement, model regularization, to name a few. In general, sparsity is achieved by using an over-complete set of prototype signals, from which recovered signals are described by sparse linear combinations of these prototypes. Sparsity can be viewed as effective dimensionality reduction that leads to an efficient data sample representation. The signal, in this case, is linearly represented with a few useful parameters. With high probability, such representations can contain most of its relevant information and often yield superior signal processing algorithms. However, a question that can arise is: "How sparse should the feature of interest be?". This section[4] deals with the estimation of an optimal maximal sparsity degree convenient to highlight the on-off regions of the speech signal features.

Thus, the main focus of this section sheds light on meaningful feature extraction based on sparse modeling with an integrated automatic degree of freedom selection with an application to the field of speech processing for the multi-speaker VAD purpose. Hence, the proposed algorithms in this section do not require a complete blind energy source separation task prior to VAD.

3.5.1 Robust and Sparse Energy Feature Extraction-Based Stability Selection

Let $\mathbf{Y} \in \mathbb{R}_{+}^{M \times N}$ denote the matrix composed of entries $\mathbf{y}[n],\ n = [1, \ldots, N]$, where $\mathbf{y}[n]$ is defined in Eqs. (2.3)–(2.5). We use an SVD decomposition that projects $\mathbf{Y}$ onto

$$\mathrm{SVD}(\mathbf{Y}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\top}, \tag{3.69}$$

where $\mathbf{U} \in \mathbb{R}^{M \times M}$ and $\mathbf{V}^{\top} \in \mathbb{R}^{N \times N}$ describe the left and right orthogonal rotations of singular vectors, respectively. $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times N}$ contains the singular values on its diagonal. In essence, we target a robust derivation of sparse right-singular vectors. Thus, we suggest as in [5] to impose sparsity-inducing penalties solely on $\mathbf{V}$ within the iterative rank-one SVD layer extraction. Sparse right rotation components grant a parsimonious

[4]This section is based on our work presented in the conference article entitled: " Robust Distributed Multi-Speaker Voice Activity Detection Using Stability Selection for Sparse Non-Negative Feature Extraction", in Proc. 25$^{\text{th}}$ IEEE Eur. Signal Process. Conf. (EUSIPCO).

speech representation, which clearly emphasizes features for the posterior VAD phase. Consequently, a lower rank representation of $\mathbf{Y}$ is undertaken with the particular requirement that the right singular vectors $\mathbf{v}$, for different sources $q = [1, \ldots, Q]$, are sparse. Sparse right rotation components serve as features for the subsequent VAD phase. Accordingly, we consider a robust ℓ_1-regularized term that minimizes the penalized sum-of-squares criterion introduced previously in Eq. (3.3). From Eq. (3.3), $\mathbf{u}$ and $\mathbf{v}$ are unit vectors of length M and N, 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 as shown in Eq. (3.4).

Based on the Lasso penalized regression in Eq. (3.3), 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 [5], we use the BIC based penalty parameter selection proposed in [118]. However, the resulting sparse vectors $\mathbf{v}$ require a subsequent unmixing step, see [5]. In this section, we favor the use of stability selection [126,127] to accurately deduce the sparseness level of the right singular vectors $\mathbf{v}$ and thus robustly determine the minimal penalization value of the regularization parameter $\lambda_{\mathbf{v}}$. This approach is promising as it surmounts the imperative use of an ensuing unmixing procedure, such as M-NICA. Stability selection is utilized to improve the estimation of the right sparse singular vectors presented in Section 3.3.1, or the work presented in [5]. Stability selection is a sub-sampling based variable selection that allows the control of false alarm rates. In this work, the aim is to infer the true set of non-zero coefficients in the right singular vector using the stability selection approach. Let $\mathcal{L}_{\mathbf{v}}$ be the set of possible $\lambda_{\mathbf{v}}$ parameters that we adapt to Eq. (3.3). Every $\lambda_{\mathbf{v}} \in \mathcal{L}_{\mathbf{v}}$ points to a distinct subspace of non-zero indicators $n \in N$ of $\mathbf{v}$ denoted $\hat{\mathcal{Z}}_{\mathbf{v}}^{\lambda_{\mathbf{v}}}(M)$. The probability of selecting a non-zero coefficient is obtained via estimating the relative selection frequency of n pertaining to sub-samples $M^{\circ} \subset M$ for an arbitrary threshold τ. This combined approach allows to control the expected number of falsely selected non-zero coefficients in the right singular vectors and therefore the degree of sparsity of the resulting right singular vector layers. Based on the selection probability, the subset of non-zero coefficients n of a vector $\mathbf{v}$ given a value $\lambda_{\mathbf{v}}$ from the set of values $\mathcal{L}_{\mathbf{v}}$ is defined as

$$\hat{\mathcal{Z}}_{\mathbf{v}} = \Big\{ n : \max_{\lambda_{\mathbf{v}} \in \Lambda_{\mathbf{v}}} P(n \in \hat{\mathcal{Z}}_{\mathbf{v}}^{\lambda_{\mathbf{v}}}(M^{\circ})) \geq \tau \Big\}. \tag{3.70}$$

In Eq. (3.70), M° is a sub-sample of M drawn without replacement from the set $\{1, \ldots, M\}$. An efficient iterative algorithm that incorporates a component-wise thresholding rule to solve the penalized regression in Eq. (3.3) with respect to $\sigma\mathbf{v}$ is given

in [118]. An approximate solution to the Lasso-based minimization problem in Eq. (3.3) is detailed in Appendix (A.1). Here, $\hat{\mathcal{Z}}_{\mathbf{v}}$ encloses the stable selection set of non-attenuated candidates $n \in N$. The minimal penalization value $\lambda_{\mathbf{v}}^{\min}$ that verifies Eq. (3.70) is used to adjust the components of the vector $\mathbf{v}$. Hence

$$v[n] = \frac{1}{\sigma}\left[\operatorname{sgn}\left\{[\mathbf{Y}^{\top}\mathbf{u}]_n\right\}\left(|[\mathbf{Y}^{\top}\mathbf{u}]_n| - \frac{\lambda_{\mathbf{v}}^{\min}}{2}\right)\right]. \tag{3.71}$$

Algorithm 12 Centralized stability selection based sparse feature extraction and robust Mahalanobis classifier for VAD (SRM-VAD)

Input: Form $\mathbf{Y} = (\mathbf{y}[1], \cdots, \mathbf{y}[N]) \in \mathbb{R}_{+}^{M\times N}$ using Eq. (2.3).
VAD procedure
for $q = 1, \ldots, Q$ **do**
  Minimize Eq. (3.3) subject to the ℓ_1-norm constraints imposed on the right-singular vector $\mathbf{v}$.
  Deduce $\lambda_{\mathbf{v}}^{\min}$ through a stability approach that selects the best set of non-zero indicators n guaranteeing sparsity in $\mathbf{v}$, based on Eq. (3.70).
  Adjust $\mathbf{v}$ with its new elements using Eq. (3.71).
  Update the singular value $\sigma = \mathbf{u}^{\top}\mathbf{Y}\mathbf{v}$.
  Construct a sparse lower-rank matrix $\mathbf{Y}^{\star} = \sigma\mathbf{u}\mathbf{v}^{\top}$.
  Collect the matrix of residue $\mathbf{Y} = \mathbf{Y} - \mathbf{Y}^{\star}$.
  Based on $|\mathbf{v}|$, extract $\boldsymbol{v}_q^{(n)} \triangleq [v_{q,1}^{(n)}, v_{q,2}^{(n)}, v_{q,3}^{(n)}]^{\top}, \forall n \in N$.
  Initial speech/silence segregation $\mathbf{C}_j$ based on $\mathbf{c}_j^{\top}, j = \{1, 2\}$.
  Compute $\hat{\mathbf{R}}_{q,j}, \forall j$ using the p-variate $t_\nu M$-estimator from Eq. (3.72).
  Evaluate robust Mahalanobis distance given in Eq. (3.73).
  Decide upon speech activity for source q using Eq. (3.74).
end for
Output: VAD patterns $\mathbf{d}_1^{\top}, \cdots, \mathbf{d}_Q^{\top}$

3.5.2 Robust Mahalanobis Classifier for Multi-Speaker VAD

In this section, we explain how the Mahalanobis distance is used to model a linear classifier of the extracted speech energy features $|\mathbf{v}|$ at every source $q \in Q$. A robustification of the covariance matrix utilized in the computation of the Mahalanobis distance is also described. The output of the designed robust classifier is a binary decision pattern where the non-activity of speech energies is labeled zero and the active speech energy regions are labeled one.

3.5.2.1 K-medians Based Speech/Silence Prior Partitioning

Our focus is to first estimate a pair of centroids $\mathbf{c}_j, j = \{1, 2\}$ associated to two separate classes, namely the active and non-active speech data points $\mathbf{C}_j, q = \{1, 2\}$, respectively. For this, we collect three statistical short-term feature series $\boldsymbol{v}_q^{(n)} \triangleq [v_{q,1}^{(n)}, v_{q,2}^{(n)}, v_{q,3}^{(n)}]^\top$ analogous [1], see Chapter 2, that well characterize the sparse vector $\mathbf{v}$ for a given source q. 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 $\mathbf{c}_j, j = \{1, 2\}$, respectively, while utilizing the features $\boldsymbol{v}_q^{(n)}$. A centroid $\mathbf{c}_j^\top$ 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 $\mathbf{C}_j, j = \{1, 2\}$, of speech/silence by assigning the realizations of $\boldsymbol{v}_q^{(n)}$ to the closest class $\mathbf{C}_j$ depending on their corresponding distances to the estimated centroids $\mathbf{c}_j$.

3.5.2.2 Robust Mahalanobis-Based Speech Detection

In this subsection, we design a Mahalanobis-based similarity measure using the robust p-variate $t_\nu M$-estimator of ν degrees of freedom, see [128], for the estimation of the covariance matrix $\hat{\mathbf{R}}_{q,j}, j = \{1, 2\}$, of the speech/non-speech feature's distributions, respectively. The latter can be formulated as

$$\hat{\mathbf{R}}_{q,j} = \frac{1}{\#(\mathbf{C}_j)} \sum_{n=1}^{\#(\mathbf{C}_j)} u_\nu(\mathbf{C}_{j,n}^\top \hat{\mathbf{R}}_{q,j}^{-1} \mathbf{C}_{j,n}) \mathbf{C}_{j,n} \mathbf{C}_{j,n}^\top, \tag{3.72}$$

with $u_\nu(t) = \frac{p+\nu}{\nu+t}$ being the weight function, p the dimension of $\boldsymbol{v}_q^{(n)}$, $t = \mathbf{C}_{j,n}^\top \hat{\mathbf{R}}_{q,j}^{-1} \mathbf{C}_{j,n}$, and $\hat{\mathbf{R}}_{q,j}^{-1}$ corresponding to the inverse covariance matrix. The robust Mahalanobis distance for the speech/silence classes then becomes

$$M_j(\boldsymbol{v}_q^{(n)}) = \sqrt{(\boldsymbol{v}_q^{(n)} - \hat{\mathbf{c}}_j)^\top \hat{\mathbf{R}}_{q,j}^{-1} (\boldsymbol{v}_q^{(n)} - \hat{\mathbf{c}}_j)}, \tag{3.73}$$

Next, speech activity is determined following the decision rule

$$d_q^n = \begin{cases} 1 & \text{if } M_1(\boldsymbol{v}_q^{(n)}) < M_2(\boldsymbol{v}_q^{(n)}) \\ 0 & \text{otherwise.} \end{cases} \tag{3.74}$$

The values 0 and 1 correspond to speech absent and speech present, respectively. Moreover, $M_1(\boldsymbol{v}_q^{(n)})$ represents the measured distance to the active speech region, while, on the other hand, $M_2(\boldsymbol{v}_q^{(n)})$ describe the calculated distance to the non-active speech region. Eq. (3.74) illustrates that we recognize an active speech when the computed distance to the active speech, i.e., $M_1(\boldsymbol{v}_q^{(n)})$, is lower than that to the non-active speech, represented by $M_2(\boldsymbol{v}_q^{(n)})$. Nonetheless, we identify a silent speech when the measured distance $M_2(\boldsymbol{v}_q^{(n)})$ is smaller than the distance to the active speech, i.e., $M_2(\boldsymbol{v}_q^{(n)}) < M_1(\boldsymbol{v}_q^{(n)})$ for a specific speaker q and a frame n. The proposed multi-speaker stability selection based sparseness combined with the robust Mahalanobis classifier for VAD (SRM-VAD) is summarized in Algorithm 12.

3.5.3 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. 2.3. A preliminary divide-and-conquer-based approach is performed. To do this, we apply the LONAS algorithm, see [1], which partitions the network into Q clusters by grouping devices around a unique dominant source based on adaptive distributed eigenvalue decomposition. Figure 2.3 illustrates the resulting device clusters (dashed red), each observing a specific source-of-interest q. We define $\mathcal{B}_q$ as the set of devices k sharing a common interest in q. Based on this distributed device structure, we construct the $(M_k \#(\mathcal{B}_q))$-dimensional vector $\mathbf{y}_{\mathcal{B}_q}[n]$ by stacking the non-negative $y_{k,m}[n]$ for every device k present in $\mathcal{B}_q$. $\#(\mathcal{B}_q)$ is the device cardinality for a given source q. Based on Eq. (2.3), the distributed signal model becomes

$$\mathbf{y}_{\mathcal{B}_q}[n] \approx \mathbf{a}_{\mathcal{B}_q}\mathbf{s}[n] + \boldsymbol{\omega}_{\mathcal{B}_q}[n], \quad n = [1, \ldots, N]. \tag{3.75}$$

Here, $\mathbf{a}_{\mathcal{B}_q}$, $\boldsymbol{\omega}_{\mathcal{B}_q}[n] \in \mathbb{R}^{M_k \#(\mathcal{B}_q) \times 1}$ reduce to the mixing vector and noise for the ensemble of devices in $\mathcal{B}_q$. In such a distributed setup, our goal is to provide a sparse estimate $\hat{\mathbf{v}}_{\mathcal{B}_q}$ by observing only the linear mixture $\mathbf{y}_{\mathcal{B}_q}[n]$. The vectors $\hat{\mathbf{v}}_{\mathcal{B}_q}, \forall q \in Q$, are features used to decide upon speaker-specific activity as outlined in Algorithm 13.

Algorithm 13 Distributed stability selection based sparseness and robust Mahalanobis classifier for VAD (DSRM-VAD)

1: **for** $q = 1, \ldots, Q$ **do**
2: $\mathbf{Y}_{\mathcal{B}_q} = (\mathbf{y}_{\mathcal{B}_q}[1], \ldots, \mathbf{y}_{\mathcal{B}_q}[N]) \in \mathbb{R}_{+}^{(M_k \#(\mathcal{B}_q)) \times N}$ using Eq. (2.16).
3: $\mathbf{s}_q \triangleq \emptyset$
4: Extract a unique layer $(\sigma, \mathbf{u}, \mathbf{v})^\top$ that solves

$$\arg\min_{\sigma, \mathbf{u}, \mathbf{v}} \|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2, \tag{3.76}$$

5: Minimize the ℓ_1 penalized sum-of-squares regression at node cluster $\mathcal{B}_q$ based on

$$\|\mathbf{Y}_{\mathcal{B}_q} - \sigma \mathbf{u}\mathbf{v}^\top\|^2 + \lambda_{\mathbf{v}} \Phi(\sigma \mathbf{v}), \tag{3.77}$$

6: At every node cluster $\mathcal{B}_q$, deduce $\lambda_{\mathbf{v}}^{\min}$ through a stability approach that selects the best set of non-zero indicators n guaranteeing sparsity in $\mathbf{v}$, based on Eq. (3.70).
7: Component-wise update of the elements of the vector $\mathbf{v}$ at cluster $\mathcal{B}_q$ relative to source q such that

$$v[n] = \frac{1}{\sigma} \left[\text{sgn} \left\{ [\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n \right\} \left(|[\mathbf{Y}_{\mathcal{B}_q}^\top \mathbf{u}]_n| - \frac{\lambda_{\mathbf{v}}^{\min}}{2} \right) \right]. \tag{3.78}$$

8: Update the vector $\mathbf{s}_q$ with the values of the sparse vector $\mathbf{v}$ calculated at cluster $\mathcal{B}_q$ for a dominant source q using

$$\mathbf{s}_q \leftarrow \mathbf{v}. \tag{3.79}$$

9: Based on $|\mathbf{s}_q|$ compute, extract $\boldsymbol{v}_q^{(n)} \triangleq [v_{q,1}^{(n)}, v_{q,2}^{(n)}, v_{q,3}^{(n)}]^\top, \forall n \in N$.
10: Initial speech/silence segregation $\mathbf{C}_j$ based on $\mathbf{c}_j^\top, j = \{1, 2\}$ at node cluster $\mathcal{B}_q$.
11: Compute $\hat{\mathbf{R}}_{q,j}, \forall j$ using the p-variate $t_\nu M$-estimator from Eq. (3.72) at node cluster $\mathcal{B}_q$.
12: Evaluate robust Mahalanobis distance given in Eq. (3.73) at node cluster $\mathcal{B}_q$.
13: Extract the speaker-specific VAD pattern $\mathbf{d}_q$ for the current observations in $\mathbf{Y}_{\mathcal{B}_q}$.
14: **end for**

3.5.4 Simulation Results for VAD and Discussion

In this section, we illustrate the performance of the proposed VAD technique using two different settings. We assess the achievement of our method given in Algorithm 12 with a centralized scenario composed of $Q = 2$ speech sources. Then, we evaluate the performance of our approach in the case of higher number of competing speech sources

Q. In this case, for a scenario composed of $Q = 6$ speech sources, we favor a distributed solution sketched in Algorithm 13. We discuss the conduct of our algorithms in both centralized and distributed setups and report on the VAD results.

3.5.4.1 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. 2.3 with two simultaneously active speech sources S_2 and S_3 and an additive white Gaussian noise (AWGN) of variance $\sigma^2_{\boldsymbol{\omega}} = 0.01$. In this case, the speech mixture is recorded at every device as shown in Eq. (2.3). We apply the centralized SRM-VAD method summarized in Algorithm 12 on the collected noisy speech mixture $\mathbf{Y}$. The degree of freedom for the robust Mahalanobis is empirically chosen as $\nu = 49$. Figure 3.13 shows the impact of choosing ν on the correct detection (CD), misdetection (MD), and false alarm (FA) rates. From Tab. 3.12, 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. 3.12. 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 S_2. Meanwhile, marginally decreased performance is obtained for S_3. This is explained by the concern of stability selection in reducing the false alarm 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. 3.12. 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 summarized in Tab. 3.12. 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 algorithm, explained in Section 3.3.

3.5.4.2 Distributed Multi-Source Scenario Use-Case

As a second experiment, we consider a WASN observing six speech sources, see Fig. 2.3, affected with AWGN of variance $\sigma^2_{\boldsymbol{\omega}} = 0.01$ variance. We deal with grouped devices following their unique dominant source [1]. Devices hearing a source with higher power are more likely to cluster together in order to cooperate for an accurate VAD. Eq. (2.16)

Centralized Use-Case						
Variance	**Source**	**Method**	CD (%)	MD (%)	FA (%)	SDR
$\sigma^2_{\boldsymbol{\omega}} = 0.01$	S_2	M-NICA [4]	62.4	1.9	35.7	-3.23
		SMM-NICA [5]	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
	S_3	M-NICA [4]	54.7	1.1	44.2	5.75
		SMM-NICA [5]	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 3.12: Comparative results of the original M-NICA [4], SMM-NICA [5], and the proposed S-VAD, SM-VAD, and the SRM-VAD (with $\nu = 49$), in a centralized scenario of two sources (S_2 and S_3) with AWGN of variance $\sigma^2_{\boldsymbol{\omega}} = 0.01$.

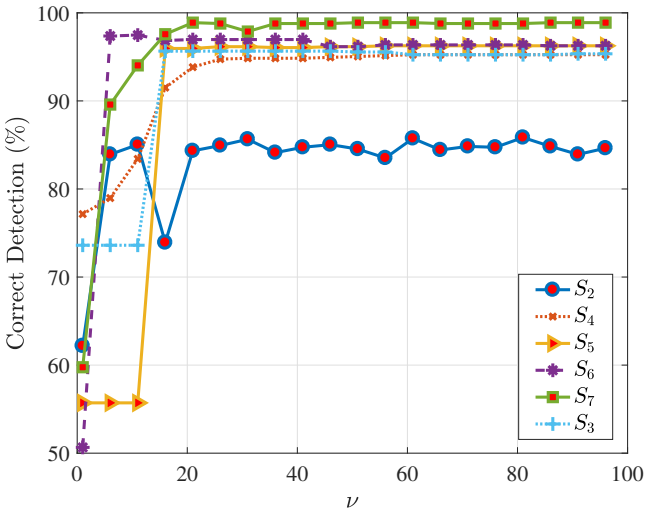
accumulates mixtures from clustered devices per primary dominant source. For the scenario sketched in Fig. 2.3, we apply Algorithm 13. The input is a sub-matrix $\mathbf{Y}_{\mathcal{B}_q}$ assembled from the $\#(\mathcal{B}_q)$ devices for source q. Table 3.13 outlines the higher decision results for the proposed distributed VAD algorithms compared to M-NICA and DM-VAD. Figure 3.14 depicts the estimated VAD patterns with high precision layered on the energy ground truth in the distributed scenario for three different speech sources S_5, S_6, and S_7.

3.6 Conclusions

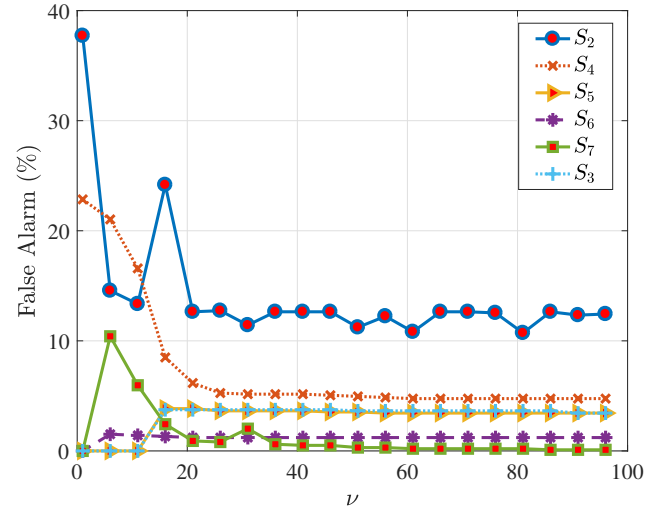
We realized a sparse constrained blind energy separation of non-negative well-grounded independent instantaneous source signals, which arises in many practical applications but is hardly ever explored for the noise-embedded model case. In contrast to the additive gradient descent updates used for instance in the non-negative principal component analysis (NPCA), we use the multiplicative weights for the update rule in M-NICA, which maintains the assumed physical condition of non-negativity. The proposed technique improves the M-NICA algorithm by integrating sparse SVD features. Then, a derivation of the SMM-NICA algorithm is introduced by applying the multiplicative

Distributed Use-Case						
Variance	**Source**	**Method**	CD (%)	MD (%)	FA (%)	SDR
$\sigma^2_{\boldsymbol{\omega}} = 0.01$	S_2	M-NICA [4]	60.8	6	33.2	-55.73
		DMVAD [1]	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
	S_4	M-NICA [4]	46.85	3	50.15	-9.5
		DMVAD [1]	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
	S_5	M-NICA [4]	56.96	3.90	39.14	-34.6
		DMVAD [1]	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
	S_6	M-NICA [4]	55.85	6.41	37.74	-14.52
		DMVAD [1]	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
	S_7	M-NICA [4]	45.75	5	49.25	-34.52
		DMVAD [1]	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
	S_3	M-NICA [4]	46.55	2.8	50.65	-20.3
		DMVAD [1]	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

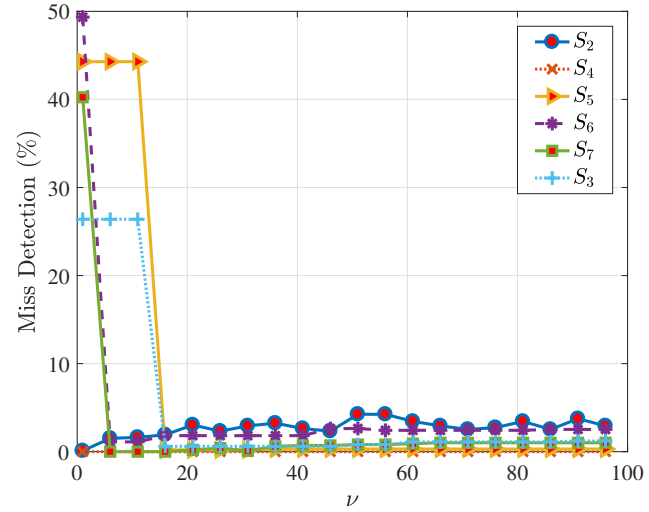
Table 3.13: Detection comparision of the original M-NICA algorithm [4], the DM-VAD approach [1], and the proposed methods: the S-VAD, the SM-VAD and the DSRM-VAD (with a degree of freedom robustness parameter $\nu = 49$), for the speech use-case scenario presented in Fig. 2.3, with AWGN of variance $\sigma^2_{\boldsymbol{\omega}} = 0.01$.



(a)

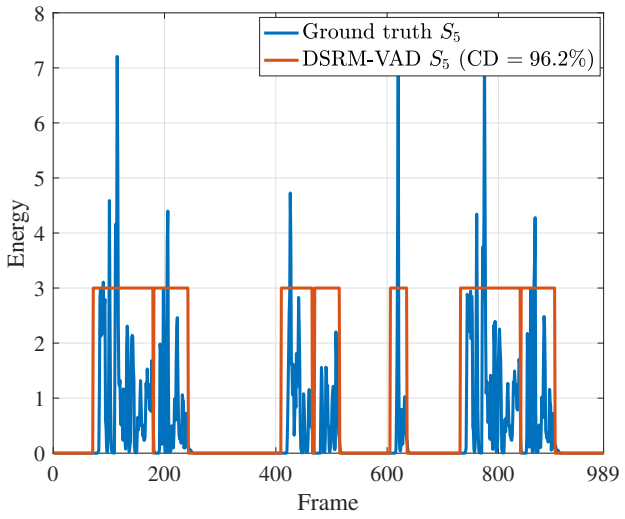


(b)

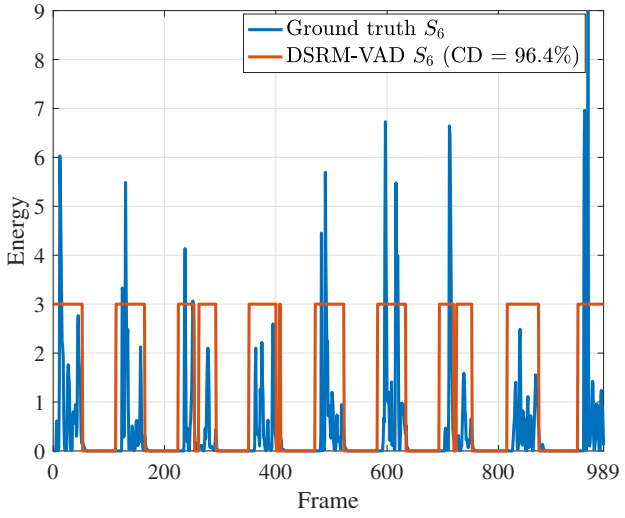


(c)

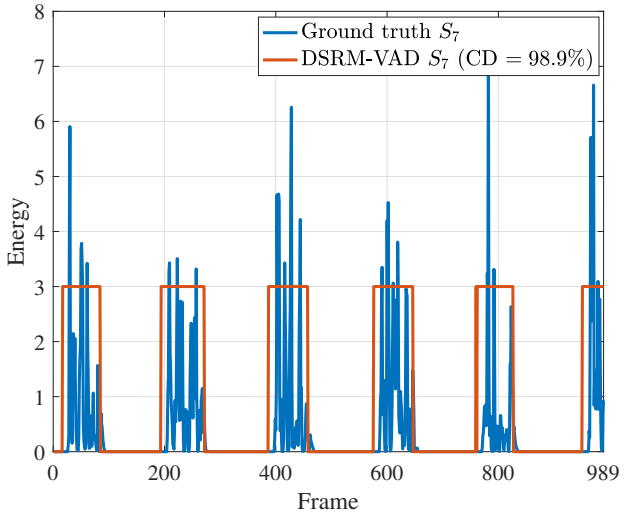
Figure 3.13: The impact of varying the degree of freedom ν on the outcome of the proposed distributed SRM-VAD in terms of (a) correct detection level, (b) false alarm rate, and (c) misdetection percentage.



(a)



(b)



(c)

Figure 3.14: The acquired VAD patterns (red) using our SRM-VAD approach in the distributed setup for (a) S_5, (b) S_6, and (c) S_7.

update rule to the robust cost function, proposed based on the mutual correlation minimization principle. The decorrelation of the sparse feature mixture is maximized with a more robust median-based multiplicative update that retains non-negativity. Since the subspace spanned by the rows of the well separated energies does not change after the initialization, our technique does not require a subsequent subspace projection correction step. The learning rate, upon which the convergence of the proposed SMM-NICA depends, is not user-defined due to the mechanism of the multiplicative update. Consequently, the multi-speaker VAD is examined as a non-negative energy separation problem for a mixture of speech signals. The VAD in the proposed scheme, reduces to determining the non-zero energies, which mitigates an empirical thresholding of the energy signals.

Furthermore, two new robust VAD algorithms are proposed to improve the speech detection robustness in non-stationary noisy environments. The presented techniques distinguish speech from noise utterances in a multi-speaker noisy scenario. The novel speech activity detectors are based on a robust and sparsely modeled multiple speech separation method. Solving the separation problem entails to inspecting a generalization of a mixture model where the underlying independent latent variables, which control the mixture components to be selected for each observation, are related to the observed signal through a linear transformation. Based on our analysis, we find that improving the speech separation of the mixture helps identifying patterns that can be used to discriminate noise from noisy speech signal and, hence, can be used as a feature for VAD. At this stage, a distributed Wiener filtering-based speech enhancement method, namely DANSE$_1$ [18], is applied to demonstrate the efficiency of the proposed robust VAD in improving the node-specific estimation of desired signals. Experimental results show that the proposed robust VADs can estimate precise speech activity under different SNR conditions, which improves the subsequent speech enhancement procedure that is sensitive to variable VAD input.

In addition, a new method for solving the multi-speaker VAD problem for WASNs in a distributed reverberant environment that does not rely on a complete energy separation step is suggested. As a matter of fact, instead of fulfilling a full source unmixing task, in the proposed approach the multi-source VAD problem is solved via a "semi" separation methodology. For this reason, energy features are extracted based on an iterative stability-based sparse SVD (SSVD). This approach involves an automatic adjustment and selection of the degree of freedom parameter in the sparse model. Consequently, our proposed method relies on a stability selection assisted technique to promote a 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 a complete energy unmixing algorithm, such as the standard M-NICA, or the proposed

SMM-NICA in Section 3.3 and Section 3.4, is no longer required. The design of a robust Mahalanobis classifier applied to reveal speaker-specific activity patterns is also proposed. At this stage, the covariance matrix of the Mahalanobis distance is determined with a robust $t_\nu M$-estimator. The robust Mahalanobis classifier aims to group similar speech features together and provide an activity pattern. The suggested approach is able to detect the presence of multiple active speech sources from a given mixture. Our algorithm is convenient for speech and non-speech discrimination with/without a prevailing blind energy source separation step. Throughout this chapter, simulation results are presented that demonstrate the high VAD performance under noisy and reverberant conditions for a challenging speech scenario in a WASN.

Chapter 4

Distributed Robust Labeling of Audio Sources in Heterogeneous Wireless Sensor Networks

'I am a slow walker, but I never walk back.'

Abraham Lincoln

The aim of MDMT in wireless microphone networks is to attain higher performance in the signal processing tasks through cooperation. In this PhD project, many source signals are assumed to be active in the WASN simultaneously. The detection and labeling of the sources is required prior to the nodes cooperation. So far, we have introduced many voice activity detection techniques for multi-source multi-device WASNs. The proposed detection algorithms rely on the available information of well-labeled sources. In this chapter, we investigate a framework for source labeling when many active targets are present in the network. Source labeling is relevant in the sense that nodes can efficiently inform each other of the specific sources-of-interest in their own task, by transmitting the corresponding labels. Source-specific features are proposed that together with an ensuing distributed clustering step allow for a unique and uniform labeling of the multiple sources throughout the WASN. With both detection and labeling tasks accomplished, cooperation amongst the devices on the basis knowledge of activity and identity of sources is enabled, where under the MDMT paradigm each node contributes to other nodes tasks following their interests.

4.1 Introduction

In the MDMT paradigm, different nodes from the network cooperate with each other to carry out different node-specific tasks. Many applications require MDMT, for instance, speech enhancement or VAD in a multi-source WASNs where devices receive speech signal mixtures. Accurate speech detection for the multiple participant speech sources should be performed in order to achieve a better subsequent speaker-related signal enhancement. In the latter, the multi-sensor devices such as smart-phones or hearing aids, are interested in enhancing their node-specific audio source-of-interest, given a

received mixture of interfering sound sources. Each node is interested in enhancing its own node-specific signal-of-interest, which may be considered to be a disturbance at a different node and vice-versa. In this case, nodes may benefit from a cooperation, even though their source-of-interest are different.

Cooperation between nodes in such an MDMT-based WASN is related to the labeling task. Pertaining to the multi-source activity detection task, the labeling information allows the devices to cluster their detected activity patterns according to the corresponding sources. Source labeling ensures that all nodes identify each source with the same label. In a WASN setting, each node observes mixtures of interfering signals transmitted by different sources, while labeling the sources requires source-specific information at each node. Hence, the labeling information must be extracted locally from the mixtures of received speech signals.

Motivated by the MDMT scheme, new techniques to solve the distributed labeling problem are needed. In this study, we deal with a realistic scenario where the features are not available a priori and have to be extracted from non-labeled sources. This is a rather challenging task due to the fact that the various speech signals, to be labeled, are mixed. Moreover, exchanging the raw sensor signals is often prohibited in practical signal processing scenarios due to communication (bandwidth/energy) constraints. Thus, the distributed multi-source labeling is performed using new simple but informative short-term features.

In this chapter[1], appropriate source-specific features are extracted at each node. In particular, we introduce a new direction-of-arrival (DoA)-based feature extraction technique that is used for the common unique labeling of all relevant speech sources that are observed by the distributed WASN. We consider non-hierarchical networks for the feature estimation. We further explain how the distributed clustering for a labeling purpose is used based on the DoA-related features. The nodes in the network perform the labeling via a distributed/cooperative unsupervised learning technique based on a similarity measure applied to the feature vectors.

[1]This chapter is based on our work presented in the conference article entitled: " Distributed Robust Labeling of Audio Sources in Heterogeneous Wireless Sensor Networks", in Proc. 40$^{\text{th}}$ IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP).

4.2 Contributions to the Distributed Multi-Source Labeling

In this chapter, we present a new framework of distributed labeling of speech sources in a WASN. A DoA-based feature extraction approach is suggested, which estimates the source-specific features from a received mixture of sound sources impinging onto the microphone array of each device in the network. For this, a subspace based algorithm is considered which provides a high resolution sub-optimal technique that exploits the Eigen-structure of the input matrix. We then consider a feature extraction step that operates in non-hierarchical networks by exploiting similarities in the frequency bands of the subspace decompositions at each node, that produce reliable DoA estimates of the speech sources. Finally, a distributed centroid clustering scheme for the DoA-related features is used for the distributed multi-source labeling purpose.

4.3 Signal Model

In this section for a better readability, we introduce a convenient notation for the basic signal model that is used in the labeling task. Consider Q narrowband far-field sources emitting waveforms impinging on K arrays of sensors. Every device $k = 1, \ldots, K$, holds M_k sensing microphones. The devices are willing and capable to cooperate in order to perform a device-specific signal processing task, e.g., speech enhancement. We deal with a sub-scenario of the speech use-case given in Fig. 1.2 of Chapter 1. The considered sub-scenario consists of $Q = 3$ speech sources, $K = 20$ devices, each equipped with $M_k = 3$ microphones in a vertically oriented ULA configuration with an inter-sensor spacing of 1.5 centimeters. From Fig. 1.2, the active speech sources are S_1, S_6, and S_3, respectively. In the current study, a FC for centralized processing is not present and the devices form a fully distributed network.

For the classical subspace DoA estimation methods, it is required that the number of sources is less than the number of sensors, that is, $Q < M_k$. Sensors collect T snapshots of the incident Q signals. The signal $y_m[\eta]$ received by the mth sensor, $m = [1, \ldots, M_k]$, at time instant η is expressed as

$$y_m[\eta] = \mathbf{a}_m^{\star}\mathbf{s}[\eta] + \omega_m[\eta], \tag{4.1}$$

where $\mathbf{a}_m^{\star}$ is a steering vector defined as

$$\mathbf{a}_m^\star = [a_m^\star(\theta_1), \ldots, a_m^\star(\theta_Q)] \in \mathbb{C}^{1\times Q}. \tag{4.2}$$

Here, θ_q denotes the DoA of the qth source. An element $a_m^\star(\theta_q)$ of the vector $\mathbf{a}_m^\star$ can be computed using

$$a_m^\star(\theta_q) = e^{\frac{-j2\pi d_{\text{mic}}}{\lambda_{\text{w}}}(m-1)sin(\theta_q)}, \tag{4.3}$$

where $m = [1, \ldots, M_k]$, refers to the reference of the microphone in the ULA, d_{mic} is the inter-sensor spacing. The wavelength is $\lambda_{\text{w}} = \frac{c_{\text{w}}}{f_{\text{w}}}$ with c_{w} being the propagation speed and f_{w} the frequency of the wave. The vector $\mathbf{s}[\eta] = [s_1[\eta], \ldots, s_Q[\eta]]^\top \in \mathbb{R}^{Q\times 1}$ represents the impinging Q signals at snapshot η. Based on Eq. (4.1), arranging the output in a node level gives

$$\mathbf{y}_k[\eta] = \mathbf{A}_k^\star \mathbf{s}[\eta] + \omega_k[\eta], . \tag{4.4}$$

In Eq. (4.4), the vector $\mathbf{y}_k[\eta]$ is an M_k-dimensional vector of elements $[y_1[\eta], \ldots, y_{M_k}[\eta]]^\top$. The matrix $\mathbf{A}_k^\star = [\mathbf{a}_k^\star(\theta_1)^\top, \ldots, \mathbf{a}_k^\star(\theta_Q)^\top] \in \mathbb{C}^{M_k\times Q}$ is the steering matrix of the angles of arrival collected from the Q impinging sources on the M_k microphones of node k. Every vector $\mathbf{a}_k^\star(\theta_q)$ is composed of the elements $\mathbf{a}_k^\star(\theta_q) = [a_1^\star(\theta_q), \ldots, a_m^\star(\theta_q), \ldots, a_{M_k}^\star(\theta_q)] \in \mathbb{C}^{1\times M_k}$. It is possible to calculate the steering vector $\mathbf{a}_k^\star(\theta_q)$ related to source q and node k using

$$\mathbf{a}_k^\star(\theta_q) = [1, e^{\frac{-j2\pi d_{\text{mic}}}{\lambda_{\text{w}}}sin(\theta_q)}, \ldots, e^{\frac{-j2\pi d_{\text{mic}}}{\lambda_{\text{w}}}(M_k-1)sin(\theta_q)}] \in \mathbb{C}^{1\times M_k}. \tag{4.5}$$

The elements of the matrix $\mathbf{A}_k^\star$ can be written as

$$\mathbf{A}_k^\star = \begin{bmatrix} 1 & \cdots & 1 & \ldots & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_m^\star(\theta_1) & \cdots & a_m^\star(\theta_q) & \cdots & a_m^\star(\theta_Q) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{M_k}^\star(\theta_1) & \cdots & a_{M_k}^\star(\theta_q) & \cdots & a_{M_k}^\star(\theta_Q) \end{bmatrix} \in \mathbb{C}^{M_k\times Q}, \tag{4.6}$$

where each column of the matrix $\mathbf{A}_k^\star$ is represented by the elements of the M_k-dimensional column vector $\mathbf{a}_k^\star(\theta_q)^\top$ in Eq. (4.5). Moreover, a row of the matrix $\mathbf{A}_k^\star$ is described by the Q-dimensional row steering vector $\mathbf{a}_m^\star$ of Eq. (4.2) at microphone m.

The network-wide narrowband signal model at time instant η can be formed by arranging the output of the $M = \sum_1^K M_k$ sensors in a linear equation, such as

$$\mathbf{y}[\eta] = \mathbf{A}^\star \mathbf{s}[\eta] + \boldsymbol{\omega}[\eta], \quad \forall \eta \in \mathbb{N}. \tag{4.7}$$

Here, $\mathbf{A}^\star \in \mathbb{C}^{M \times Q}$ is the array response matrix of the Q emitted speech signals on the M microphones of all the K nodes.

The goal of DoA is to estimate the angles of arrivals of all targets $q = [1, \ldots, Q]$. Necessary assumptions are made to ensure the stability of most of the subspace-based DoA methods:

- The number of sources Q is assumed to be known.
- The number of sources Q is less than that of the sensors M_k, $Q < M_k$
- Spatially white noise
- Independent sources and sources are also independent from noise.

Much research has been devoted to the development of alternative sub-optimal but computationally feasible DoA methods, many of which are variations of the multiple signal classification (MUSIC) algorithm [129–132]. In the sequel, we present a short overview of direction finding literature and summarize the basic steps of the Khatri-Rao MUSIC (KR-MUSIC) approach, which we use for extracting DoA related features relevant for the labeling task.

4.4 Fundamentals on Direction-of-Arrival Estimation

4.4.1 Direction-of-Arrival Estimation: State-of-the-Art

High resolution DoA estimation has received substantial attention. Many real-world applications require DoA estimation including wireless communications, radar, sonar,

tracking, and localization [132]. The signal subspace-based array processing algorithms represent an important class of techniques for DoA estimation. Subspace-based methods exploit the underlying data model via separating the space spanned by the recorded signal into noise and signal subspaces. Numerous signal processing techniques have been established to cover this area. Among others, the MUSIC algorithm and its root-MUSIC variation are the most popular [129–131]. MUSIC is a DoA estimation method applicable to arrays with arbitrary geometry. In Fig. 4.1, for instance, we show the MUSIC-based DoA estimation of three targets using a ULA. Moreover, Fig. 4.2 illustrates the estimation of the same angles using a uniform circular array (UCA) geometry for the MUSIC-based DoA estimation. An interesting MUSIC based method that uses the Khatri-Rao (KR) product is developed in [133–135] and will be briefly introduced in Section 4.4.2. In [133], the quasi-stationary signal characteristic is explored to determine the direction of a transmitting source. Moreover, unlike the standard narrowband array processing model, wideband approaches tend to operate in the frequency-domain. Different high-resolution subspace approaches to wideband DoA estimation employing the frequency-domain methods are elaborated in the literature [135–138].

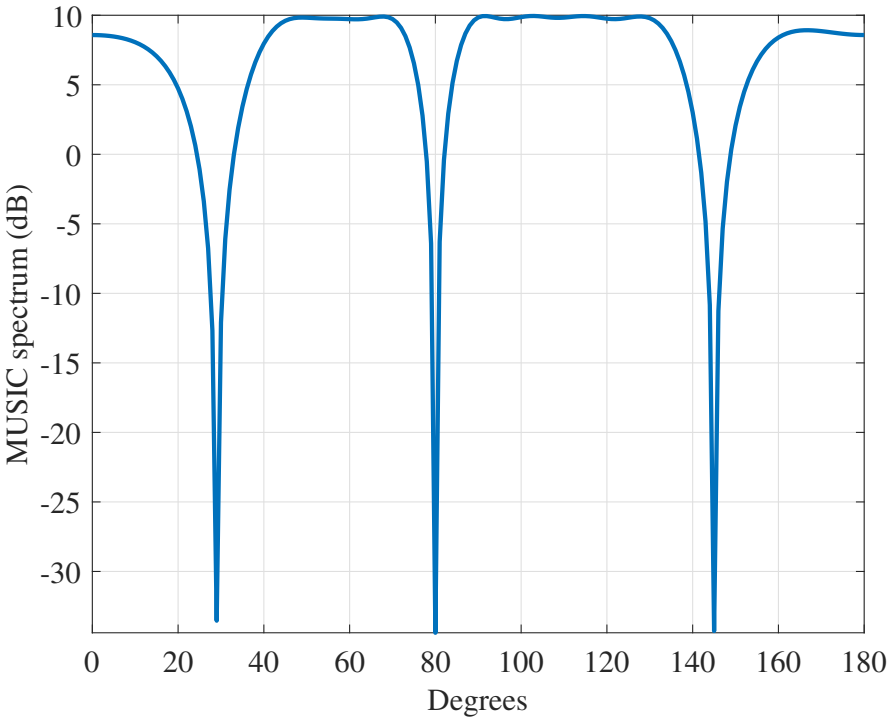


Figure 4.1: Example of DoA estimation based on the MUSIC algorithm with a ULA configuration.

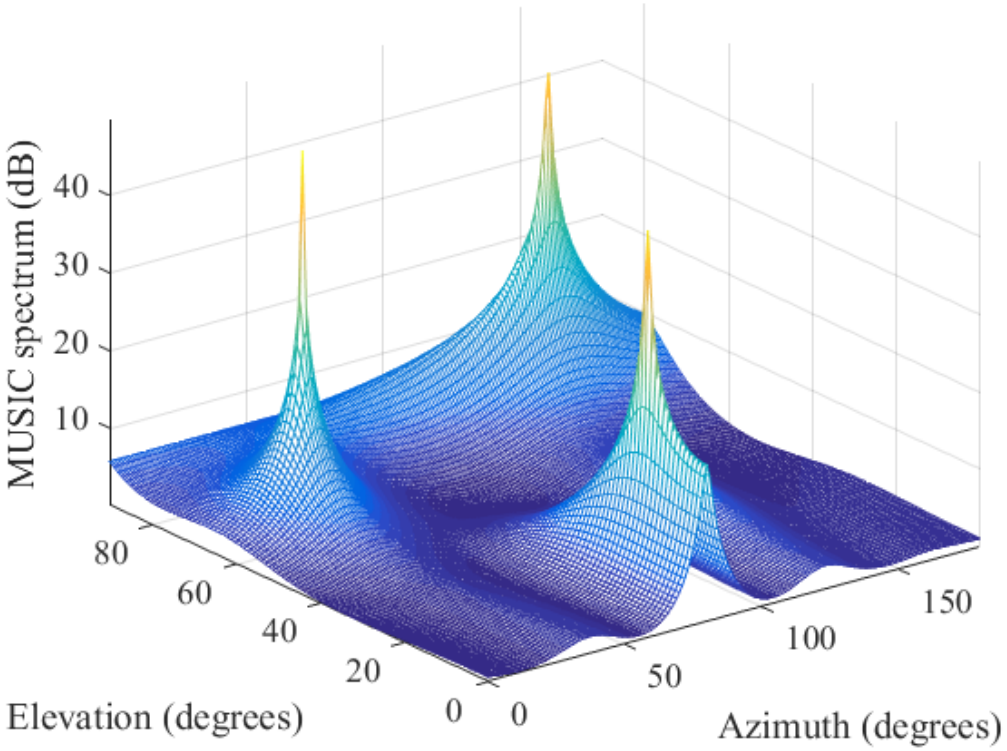


Figure 4.2: Example of DoA estimation based on the MUSIC algorithm with a UCA configuration.

4.4.2 The Khatri-Rao-MUSIC Approach

Given the received signal sequence $\mathbf{y}_k[\eta]$ at node k of Eq. (4.4) and by assuming local stationarity at intervals of length L, the KR-MUSIC approach models the local covariance matrix under the quasi-stationary assumption of the speech signal. The sensor local covariance matrix at a stationary frame n is of dimension $\mathbb{R}^{M_k \times M_k}$ and is estimated by local averaging using

$$\hat{\mathbf{R}}_n^{\text{loc}} = (1/L) \sum_{\eta=(n-1)L}^{nL-1} \mathbf{y}_k[\eta]\mathbf{y}_k^{\mathsf{H}}[\eta], \tag{4.8}$$

where n represents the frame index, $n = 1, ..., N$, and $N = T/L$. The superscript $^{\mathsf{H}}$ is the Hermitian transpose. The Khatri-Rao product properties play an important role in formulating a new array signal model where a virtual array response matrix is formed which is of dimension greater than the physical array dimension, see [139]. This

intrinsically provides the possibility to treat situations where the number of sensors is less than that of sources i.e. $M_k < Q$. The computed local covariance matrices $\hat{\mathbf{R}}_n^{\text{loc}}$ are then stacked in a way to form the matrix $\hat{\mathbf{Y}}$ as described in the ensuing formula.

$$\hat{\mathbf{Y}} = [\text{vec}(\hat{\mathbf{R}}_1^{\text{loc}})^\top, \ldots, \text{vec}(\hat{\mathbf{R}}_N^{\text{loc}})^\top], \tag{4.9}$$

where $\hat{\mathbf{Y}} \in \mathbb{R}^{(M_k \times M_k) \times N}$ and $\text{vec}(\cdot)$ is the vectorization function that transforms a matrix $\hat{\mathbf{R}}_n^{\text{loc}} \in \mathbb{R}^{M_k \times M_k}$ to a $(M_k \times M_k)$-dimensional vector. Smoothing of the covariance matrix is performed with a noise covariance elimination by applying

$$\bar{\mathbf{Y}} = \hat{\mathbf{Y}}\mathbf{P}_N, \tag{4.10}$$

with

$$\mathbf{P}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \in \mathbb{R}^{N \times N}. \tag{4.11}$$

Then, a dimension reduction step is completed using

$$\tilde{\mathbf{Y}} = \mathbf{W}_{\text{KR}}^{\frac{1}{2}}\mathbf{G}_{\text{KR}}^\top\bar{\mathbf{Y}}, \tag{4.12}$$

subject to

$$\mathbf{G}_{\mathrm{KR}} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \dots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ \\ 0 & \cdots & 1 & 0 & 0 & \dots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 & 0 \\ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \\ 1 & 0 & \cdots & 0 & 0 & \dots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{M_k^2 \times 2(M_k-1)} \tag{4.13}$$

and

$$\mathbf{W}_{\mathrm{KR}} = \mathbf{G}_{\mathrm{KR}}^{\top}\mathbf{G}_{\mathrm{KR}} = \mathrm{Diag}(1, \dots, M_k-1, M_k, M_k-1, \dots, 1) \in \mathbb{C}^{(2M_k-1)\times(2M_k-1)}, \tag{4.14}$$

with $\tilde{\mathbf{Y}} \in \mathbb{C}^{(2M_k-1)\times N}$. Next, the SVD projects $\tilde{\mathbf{Y}}$ onto

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^{\mathsf{H}}, \tag{4.15}$$

with $\tilde{\mathbf{\Sigma}} \in \mathbb{C}^{Q\times Q}$, $\tilde{\mathbf{V}} \in \mathbb{C}^{N\times Q}$, and $\tilde{\mathbf{U}} \in \mathbb{C}^{(2M_k-1)\times Q}$. The left singular matrix $\tilde{\mathbf{U}}$ yields the noise subspace matrix, such that

$$\tilde{\mathbf{U}}_{\mathrm{noise}} = [\tilde{\mathbf{u}}_{Q+1}, \dots, \tilde{\mathbf{u}}_{2M_k-1}] \in \mathbb{C}^{(2M_k-1)\times(2M_k-1-Q)} \tag{4.16}$$

Next, the DoA spatial spectrum computed based on the Khatri-Rao subspace framework takes the form

$$P_{\text{KR-MUSIC}}(\theta) = \frac{1}{\|\tilde{\mathbf{U}}_{\mathrm{noise}}^{\mathsf{H}}\mathbf{W}_{\mathrm{KR}}^{1/2}\mathbf{b}(\theta)\|^2}, \quad \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]. \tag{4.17}$$

In the KR-based spectrum formulation of Eq. (4.17), the vector $\mathbf{b}(\theta)$ represents a dimension reduced virtual array response vector and can be computed using

$$\mathbf{b}(\theta) = [e^{(M_k-1)\frac{j2\pi d_{\text{mic}}}{\lambda_{\text{w}}} sin(\theta)}, \ldots, e^{\frac{j2\pi d_{\text{mic}}}{\lambda_{\text{w}}} sin(\theta)}, 1, \ldots, e^{-(M_k-1)\frac{j2\pi d_{\text{mic}}}{\lambda_{\text{w}}} sin(\theta)}]^\top \tag{4.18}$$

The Q largest peaks of $P_{\text{KR-MUSIC}}(\theta)$ in Eq. (4.17) represent the DoA estimates for the narrowband signal model described in Eq. (4.7).

For speech scenarios, signals are non-stationary but can be modeled as stationary within local time frames. In [139], DoA estimation is performed for locally stationary signals. For this an extended DoA spectrum formulation based on Eq. (4.17) is derived. The frequency-domain representation decouples the wideband signals into many narrowband signals. The short-time Fourier transform (STFT) is used in [139] to transform quasi-stationary signals into a multitude of narrowband models. We apply the STFT on the observed noisy signals received at the microphones and denote $\check{\mathbf{y}}[n, f]$ the short-term time-frequency representation of $\mathbf{y}[\eta]$ at frame index n and frequency bin f. According to [139], a wideband signal model approximation can be formulated as

$$\check{\mathbf{y}}[n, f] \approx \mathbf{A}^\star(f)\check{\mathbf{s}}[n, f] + \check{\boldsymbol{\omega}}[n, f], \quad n = 1, \cdots, N. \tag{4.19}$$

Here, $\mathbf{A}^\star(f)$ represents a frequency dependent array response matrix. In order to estimate DoAs, combination of the subspaces at various frequencies is performed in the frequency-domain KR-MUSIC approach to obtain a spectrum fusion. The DoA spectrum of wideband models can then be computed using [139]

$$P_{\text{KR-MUSIC}}(\theta) = \frac{1}{\sum_f \|\tilde{\mathbf{U}}^{\mathsf{H}}_{\text{noise}}(f)\mathbf{W}^{1/2}_{\text{KR}}\mathbf{b}(\theta, f)\|^2}, \quad \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}], \tag{4.20}$$

where $\tilde{\mathbf{U}}^{\mathsf{H}}_{\text{noise}}(f)$ denotes the KR noise subspace at a frequency f and $\mathbf{b}(\theta, f)$ is an extended steering vector at frequency f. The Q largest peaks of $P_{\text{KR-MUSIC}}(\theta)$ in Eq. (4.20) represent the DoA estimates for the wideband signal model described in Eq. (4.19).

4.5 Distributed Labeling Based on Clustered Khatri-Rao-MUSIC Direction-of-Arrival Features

In this section, the distributed labeling of different sources in a WASN is explained. The proposed labeling approach of different speech sources throughout the network consists of two steps, compiled in

1. A DoA-based feature extraction at each device for each speech source of the WASN.
2. A distributed clustering of the computed DoA-based features.

We next detail the above steps to solve the labeling task.

4.5.1 Non-Hierarchical Feature Extraction: Exploiting Similarities in the Frequency Bands Which Produce Reliable Direction-of-Arrival Estimates

In the feature extraction phase, we ideally extract features for each speech source, which are similar from node to node. This is a challenging task, since the various speech signals are mixed and the signal powers in the mixtures differ significantly.

Non-hierarchically organized networks are able to extract features without forming subnetworks. Our study focuses on extracting features based on promising high resolution DoA estimation. However, DoA information cannot be applied directly to labeling, since, in general, the devices in a WSN do not know their positions and array orientations. Furthermore, in the considered setup, due to the ULA configuration, and the use of omnidirectional microphones, an ambiguity in the DoA estimates along the symmetry axis of the array orientation cannot be resolved.

For this reason, we propose a novel feature which exploits the similarity across devices in the particular frequency bins that produce more accurate DoA estimates for each source. The DoA is estimated with the KR subspace approach for locally stationary wide band signals. The idea of the KR method is to form a new array signal model by use of the KR-product, which generates a virtual array response matrix that is of

greater dimension than the original physical array [139]. In this way, the KR method can identify up to $Q = 2M_k - 2$ unknown sources in underdetermined mixing systems of M_k sensors. In our work, we estimate the DoA based on the frequency dependent spatial KR-MUSIC spectrum defined as

$$P_{\text{KR-MUSIC}}(\theta, f) = \frac{1}{\|\tilde{\mathbf{U}}_{\text{noise}}^{\text{H}}(f)\mathbf{W}_{\text{KR}}^{1/2}\mathbf{b}(\theta, f)\|^2}, \quad \theta \in [-\frac{\pi}{2}, \frac{\pi}{2}], \tag{4.21}$$

The Q largest peaks of $P_{\text{KR-MUSIC}}(\theta, f)$ in Eq. (4.21) at a specific frequency bin f represent the DoA estimates for the wideband signal model described in Eq. (4.19). Based on the calculated DoA at every frequency f in Eq. (4.21), an overall DoA estimate can be obtained, e.g. by taking the geometric or the arithmetic average or their robust estimates over the estimates of all frequencies. However, due to the noisy environment and multiple source interference, the DoA estimates are inappropriate at some frequency bands. For this reason, we propose a source-specific frequency bands selection that contribute in "good" DoA estimation based on the generated spectrums $P_{\text{KR-MUSIC}}(\theta, f)$ in Eq. (4.21) at every device.

Figure 4.3 displays the estimated DoA for S_6 and S_3 at device D_1 at different frequencies. Likewise, Fig. 4.4 illustrates the estimated DoA for S_6 and S_3 at device D_{14}. The overall DoA for each source $\hat{\theta}_q,\ q = 1, \ldots, Q$, is obtained by taking the median of $\hat{\theta}_q(f)$ with $0 < f < f_s/2$. The dashed red lines indicate the $\hat{\sigma}_{\hat{\theta}_q}$-interval around $\hat{\theta}_q$. If the DoA is estimated correctly, $\hat{\theta}_q(f)$ is centered around the median. However, due to noise in particular sub-bands, or due to interference from other sources, the distribution of the estimates may be heavy-tailed, as it contains outliers. It is therefore necessary to estimate $\sigma_{\hat{\theta}_q}$ robustly [125], e.g., with the median absolute deviations scale estimator. In this manner, the source-specific frequency bands that typically contribute to correct DoA estimates are selected. The proposed feature vector is formed for each source at each device by storing the frequency bin indexes within $\hat{\theta}_q \pm \hat{\sigma}_{\hat{\theta}_q}$. Figures 4.5 and 4.6 illustrate the binary feature vectors of the selected frequency bins that contribute in a better DoA estimation of source S_3 at devices D_1 and D_{14}, respectively. Section 4.5.2 discusses how the DoA-based features with selected frequencies are used to achieve a labeling of sources in the network.

4.5.2 Distributed Clustering of Direction-of-Arrival-Based Frequency Selected Features

After computing source-specific features, we employ a distributed clustering scheme. The goal of distributed clustering algorithms is to form the clusters in a way that

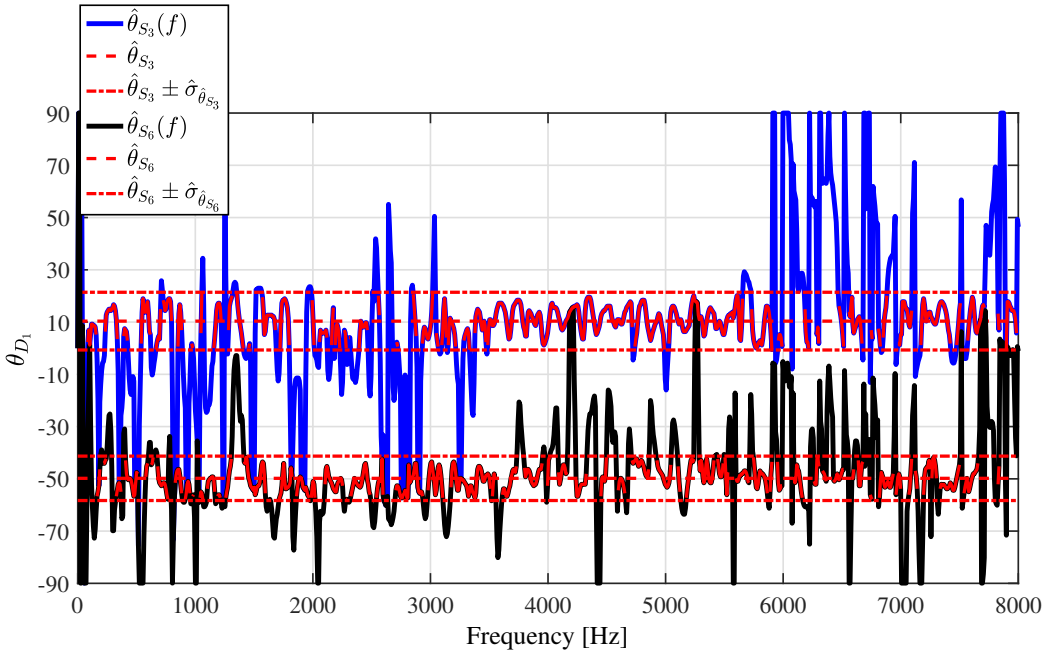


Figure 4.3: The proposed non-hierarchical feature displays which frequency bins produce reliable DoA estimates for each source at different nodes. The underlying DoA estimates from which the feature is derived are displayed for D_1, given S_6 and S_3, with positions, as depicted in Fig. 1.2.

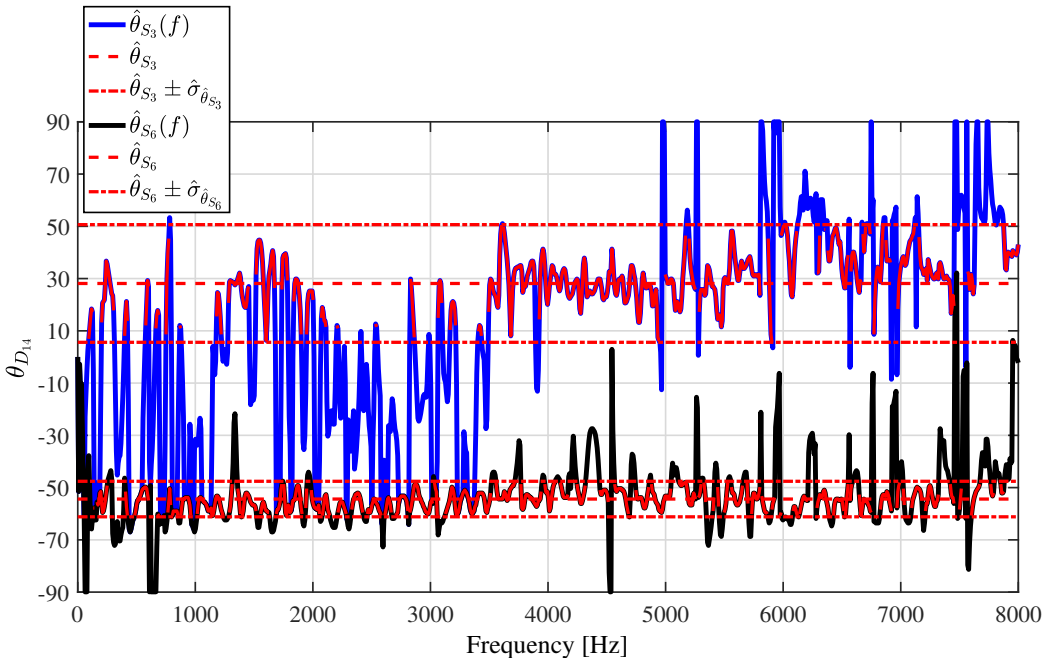


Figure 4.4: The proposed non-hierarchical feature displays which frequency bins produce reliable DoA estimates for each source at different nodes. The underlying DoA estimates from which the feature is derived are displayed for D_{14}, given S_6 and S_3, with positions, as depicted in Fig. 1.2.

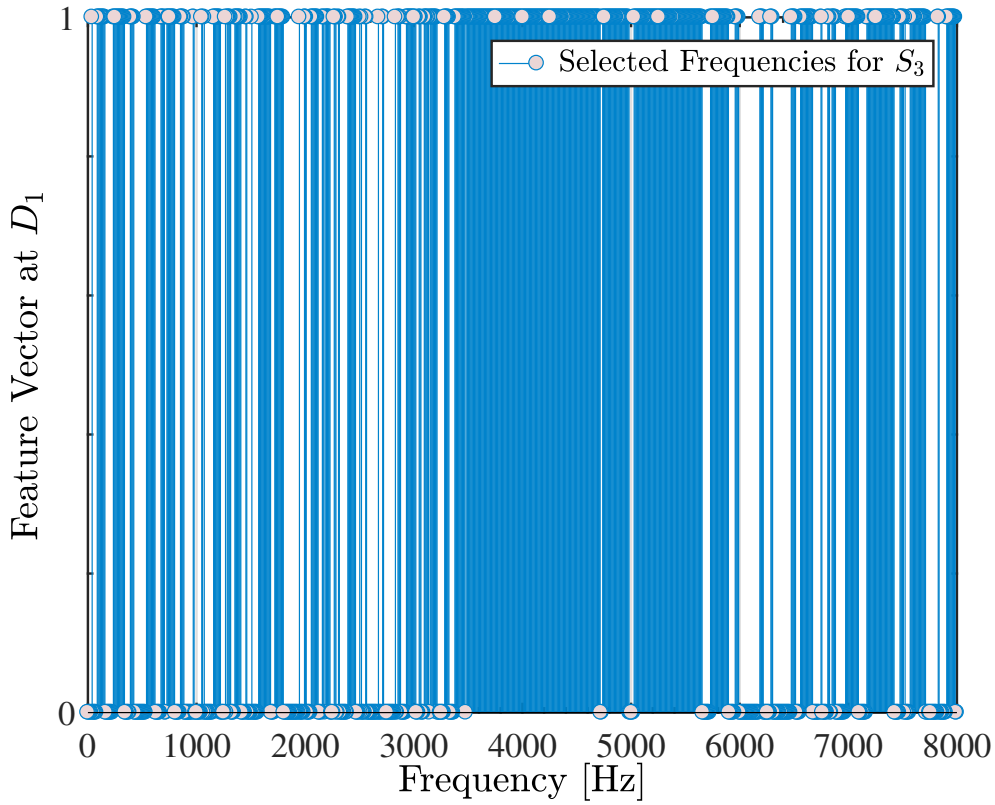


Figure 4.5: Example of a binary feature vector indicating the selected frequency bins at Node D_1 that produce a reliable DoA estimation for S_3.

they exploit all the available data of the network by relying, however, only on local processing, at each node, as well as on interactions within the node's neighborhood. One possibility to achieve this goal is by consensus averaging [80,140]. In this case, the nodes average the computed centroids of the clusters and consensus on the centroids is accomplished. In other words, the nodes compute the same centroids and consequently the same clusters. This is a crucial point for the labeling problem, since if the devices have computed the same clusters, the labeling can be readily performed.

In this Section, we will discuss how the previously extracted DoA-based features can be incorporated into a distributed clustering algorithm. Ideally, we would like each cluster to contain every feature corresponding to the same speech source. We realize this by employing a cooperative clustering scheme. This is achieved when a node cooperates with its neighbors, and these cooperate, in turn, with their neighbors. Then, the information coming from the whole network is incorporated. The distributed clustering presented in [80] is adapted so as to fit with the current context of clustering DoA-based features for source labeling. The steps to achieve this are summarized in Algorithm 14.

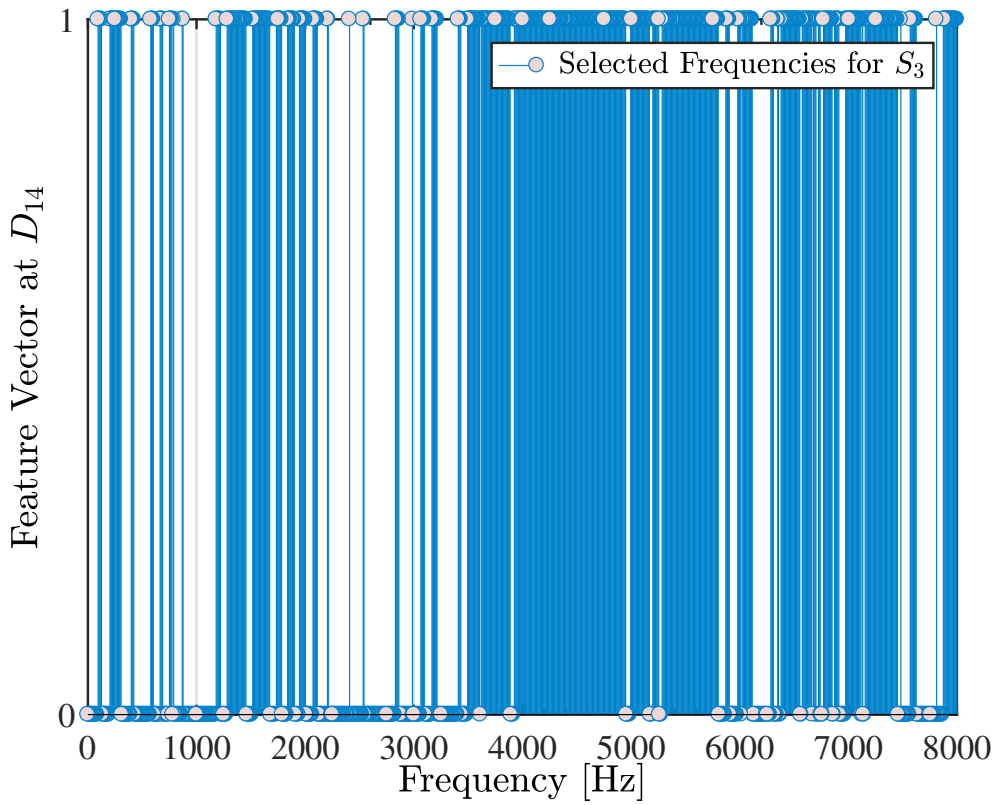


Figure 4.6: Example of a binary feature vector indicating the selected frequency bins at Node D_{14} that produce a reliable DoA estimation for S_3.

Algorithm 14 Distributed clustering for multi-source labeling based on DoA information

1: **Centroid Initialization:**
$\boldsymbol{c}_q^{(k)}(i=0) = \boldsymbol{c}_q^{(k')}(i=0)$
2: **repeat**
3: **Local Clustering Phase:**
Apply the K-means algorithm at every node k
4: **Cooperation Phase and Centroid Update:**
5:

$$\boldsymbol{c}_q^{(k'')}(i) = \frac{\widetilde{\boldsymbol{c}}_q^{(k)}(i) + \widetilde{\boldsymbol{c}}_q^{(k')}(i)}{2}, \; k'' = k, k'. \tag{4.22}$$

6: **until** reaching a centroid consensus at iteration i

As a first step for the distributed clustering, an initialization of the centroids is perfomed. The nodes (sub-networks) initialize the centroids $\boldsymbol{c}_q^{(k)}(0),\ q = 1, \ldots, Q$ as shown in Algorithm 14. 1. Methodologies for selecting the initial centroids so as to satisfy the above equality can be found in [80, Section 7.7]. Next, a local clustering is achieved. In this case, each node (sub-network) k, at iteration i, performs a local clustering scheme

by employing a distributed K-means algorithm, which uses the computed features and the previously computed centroids $\boldsymbol{c}_q^{(k)}(i-1)$. For the DoA related features each feature is assigned to the cluster, for which the Euclidean distance between the feature vector and the centroid is minimized. The centroids $\widetilde{\boldsymbol{c}}_q^{(k)}(i)$ are computed for $q = 1, \ldots, Q$. After the clustering is performed, node k' belonging to the neighborhood of k is activated with a certain probability (see also [140]). We assume that node k picks some neighbor k' with probability $1/\mathcal{N}_k$, where $\mathcal{N}_k$ is the number of neighbors of k. Nodes k, k' update their centroids $q = 1, 2, \ldots, Q$ according to Eq. (4.22) of Algorithm 14. Labeling is readily performed once the clusters are computed. The label of each speech signal q will be set equal to the number of the class, in which the respective feature belongs. The averaging that takes place in the cooperation phase of the algorithm drives the nodes of the network to a centroid consensus. This means, the nodes compute, after a sufficient number of iterations, the same centroids. This behavior is consistently observed in extensive experiments. Moreover, the centroid consensus is proved in [76] for a similar distributed K-means clustering scheme as the one we employ.

4.6 Simulation Results

In this Section, we study the performance of the proposed distributed labeling approach. We consider a sub-scenario of the network depicted in Fig. 1.2 and we validate the accuracy of the labeling, using the proposed DoA features. The achievement in terms of source labeling based on DoA is compared to the energy based features for distributed labeling, suggested in [34]. In Fig. 1.2, the speech signal S_1 corresponds to a woman making a public announcement, whereas S_6 and S_3 consist of two male speakers that are reading sentences in different languages. We use the mirror image method [141] to synthesize room impulse responses that can be used to compute the signals captured by microphones at arbitrary positions in a reverberant enclosure with multiple sound sources.

In the first experiment, we consider that two speech sources, i.e., S_6 and S_3, are active. We assume that both babble and white noise are present in the environment. The variance of the noise processes is varied, so as to validate the accuracy in different noise scenarios. The sampling frequency of the microphone signals is $f_s = 16$kHz. The DoA-based features are only computed on a single short interval of 0.5 seconds, where all sources are active. Finally, two nodes of the network are assumed to be connected if their distance is smaller than 4.5 meters.

Table 4.1 summarizes the results. It can be seen that the clustering accuracy, using the DoA estimates, drops as the variance of the noise increases. On the contrary, using

Noise Variance σ_{ω}^2	DoA Related Features	Energy-Based Features [34]
0	100%	100%
0.1	89%	100%
0.5	65%	100%

Table 4.1: Distributed source labeling: Results for the two source scenario, S_6 and S_3.

the energy-based features, proposed in [34], the distributed clustering-based labeling algorithm is able to label correctly the speech sources. This advantage comes at the cost of forming a hierarchical network. In particular, the energy-based feature, as it is apparent in Tab. 4.1, exhibits a better accuracy, compared to the DoA-based feature. However, the former requires a hierarchical network and the process takes place over the full-time signal. On the contrary, the labeling accuracy of the DoA features slightly degrades, but these features can be computed at node level and the DoAs are estimated on much shorter speech intervals of only 0.5 seconds.

In the second experiment, we consider the more challenging scenario, where all the sources, namely S_1, S_6, and S_3, are active in the network. The parameters remain the same as in the previous example and the noise variance are varied as depicted in Table 4.2.

Noise Variance σ_{ω}^2	DoA Related Features	Energy-Based Features [34]
0	80%	100%
0.1	60%	82%

Table 4.2: Distributed source labeling: Results for the three source scenario, S_1, S_6, and S_3.

As it is expected, the performance drops compared to the two-source scenario. Similarly to the previous experiment, a better accuracy is achieved by employing the energy-based features. It is worth pointing that, the performance of the labeling algorithm is degraded, due to the fact that some nodes of the network are located in positions, in which they are not able to hear all the speech sources. However, in the feature extraction phase, we force the devices to assume that 3 sources are active and to form 3 clusters. A preprocessing, through which the number of active sources in a node is computed, could potentially enhance the results.

4.7 Conclusions

In this chapter, the question of labeling multiple sources in a distributed WASN is addressed. Our proposed approach first derives features related to the DoA estimation of multiple targets using a time-frequency analysis. Source-specific frequencies that contribute to reliable DoA estimates are identified and selected for the computation of the angles of arrival. The DoA-based features are estimated for each participating source at every device of the network. Labeling of active sources in the WASN is then achieved via a distributed clustering technique. Local processing of the DoA-based feature vectors is performed, in which nodes attempt to determine locally the cluster membership of the computed features and update their centroids. Cooperation between neighboring nodes is then achieved where an averaging procedure of the centroids is applied. Extensive clustering iterations show that nodes converge to a centroid consensus over the network upon which the labeling of targets is performed at every node. Experiments show that the proposed methodology is able to accurately label speech signals in a practical speech scenario for high SNR values.

Chapter 5

Summary, Conclusions and Future Research

'Sometimes what you think is an end is only a beginning.'

Agatha Christie

5.1 Summary and Conclusions

In many current applications, low-cost sensors with high sensing capabilities are generally deployed in the WASNs to solve difficult signal processing tasks. The rapid computational improvement of the physical sensors comes with the cost of highly stringent constraints in sensor networks, which urges new wireless ad hoc networking techniques. To this end, the MDMT paradigm is of paramount importance, so as to endure the rapidly changing statistics of multimedia signals. The developed methods in this doctoral thesis are enablers for MDMT and allow for increased higher order cooperation in WASNs.

In this doctoral project, novel algorithms for distributed labeling and detection of multiple sources in a reverberant and noisy WASN have been introduced. Several speech processing applications depend on source activity detection/labeling front-end techniques. The proposed multi-source methodologies have been validated in a challenging realistic speech scenario using real speech data recorded in a reverberant noisy multi-device WASN. It is to note that the detection and labeling questions are inter-related. This means, on the one hand, solving the detection task for multiple sources requires a prior labeling task. On the other hand, labeling of multi-sources can be of higher accuracy if done on the basis of known signal activity information. The latter, indeed, helps focusing the labeling search on active signal segments that bear useful information compared to non-active regions of the signal, which are considered as noise. In contrast to long-term features obtained from longer portions of signals, in this thesis, we have utilized stationary short-term energy-based features. The centralized methods in this thesis rely on a central entity that polls the detection/labeling decisions from all the contributing nodes. For an increased number of participating speech sources,

the proposed approaches are readily applied to distributed WASN settings, where no fusion center is available and scalability is not an issue.

In the context of distributed multi-source detection, new approaches have been presented in this thesis to tackle the signal detection problem in the presence of many sources in WASNs equipped with various devices. In fact, conventional algorithms for single-source detection trade off accuracy and computational cost, while no research has yet solved the distributed VAD for multiple-concurrent-sources in a reverberant and noisy WASN environment. Essentially, we have proposed a new framework to solve the multi-source detection problem based on distributed dominant source extraction and a subsequent robust clustering approach to determine signal activity patterns. Distributed energy source unmixing has been performed via a rank-one non-negative independent component analysis that uses multiplicative updates rule, which is computed at node clusters sharing a unique source-of-interest. Following to that, a distributed partitioning-based binary source activity detection has been presented, which precisely distinguishes active from non-active utterances of the different competitor energy signals in the WASN. In a second approach, we have proposed a Lasso-based sparse blind energy source separation of noisy mixed energies recorded at different nodes in the network. At this stage, we have derived a new non-negative blind energy source separation technique that combines sparse coding and multiplicative updates rule for sparse non-negative energy extraction relative to the participating sources in the WASN. Precisely, we have assumed a sparse representation of the right rotation loads of a singular value decomposition extracted by an iterative procedure where every layer describes an energy source. The decorrelation of multiplicative nature between the rows of the matrix of sparse right vectors is maximized while preserving non-negativity of the signals. We have shown that sparse decomposition in the unsupervised learning provides high-quality blind energy source separation. Consequently, the multi-source detection problem is converted to a non-negative blind energy separation where the non-active energy frames are automatically tuned to zeros due to the enforced sparse modeling. Moreover, robust $t_\nu M$-estimator-based sparse energy separation algorithms have been suggested. The robustly unmixed sparse energy signatures of the sources readily produce a straightforward zero-threshold VAD, which detects speech activity.

In this doctoral thesis, we also aimed at surmounting the dependency on full unmixing approaches in solving the multi-source detection problem. Our consecutive logic suggests using a partial separation technique based on a sparse singular value decomposition jointly with a robust stability approach for sparseness variable selection. We have shown that the proposed robust non-negative sparse energy signal decomposition generates sufficiently separated energies for a subsequent suggested robust voice activity classifier based on the Mahalanobis distance.

In the context of labeling, a distributed multi-source labeling algorithm has been developed to uniquely label all energy source signals throughout the WASN. In this part, extraction of proper energy-based source-specific features from the mixed signals is achieved. The features are based on a STFT applied to the estimation of DoAs. We have exploited these DoA related features via a distributed unsupervised learning technique to reach an accurate source labeling.

5.2 Future Research Directions Based on the Proposed Multi-Source VAD and Labeling Techniques

In this section, we suggest some future research directions.

5.2.1 Image Unmixing

Linear unmixing of non-negative image mixtures is considered in [85]. Due to the non-negativity of the pixels, image mixtures separation can naturally be seen as a NICA problem. Pertaining to hyperspectral image processing, which finds usage, e.g., in military target recognition applications, hyperspectral unmixing is a crucial preprocessing step. Due to the non-negativity of the hyperspectral images' spectra, sparse NICA-based methods, similar to what we suggest in Chapter 3, can be applied.

5.2.2 Distributed Multi-speaker Diarization and Localization Based on Joint Robust VAD, Labeling, and DoA Estimation

Speaker diarization consists of answering the question "who speaks when?". It is considered as an extension of the speaker identification task in speaker recognition where additionally the occurrence time of a speech segment has to be determined. Speaker diarization has gained increased visibility and significance in society as speech technology continues to expand. In real-world distributed multi-source speech scenarios such as group meeting situations, the speakers should be identified and associated to their respective speech segments without any prior information about the audio nor the speakers. As a continuation of our research, distributed multi-speaker diarization

can be accomplished. In this case, we suggest using distributed processing to extract the main information related to multi-source VAD, labeling and DoA-based localization. Jointly exploiting this prior knowledge to build a diarization system seems to be a promising idea. This information makes it possible to detect "who speaks when, and where a speech segment occurs". The proposed speaker diarization system relies on a series of steps summarized in Fig. 5.1. First, a pre-processing of the raw input data is performed consisting of a distributed node clustering, for example using the LONAS algorithm [1], and a subsequent distributed sparse energy feature extraction. In order to avoid the adverse effect of impulsive noise and the interfering neighboring sources, a node-specific speech enhancement technique based on multi-source VAD information can be employed. Improving speech quality allows for the design of sophisticated algorithms. Text-independent VAD information is essential to determine when a specific speaker is talking. The direction of an impinging speech signal relating to a speaker is resolved using a DoA-based method. Estimating DoAs of the different participating sources is done only using active speech segments, which carry more useful information. DoAs are more accurate when using exclusively active speech segments, estimated via VAD, and the computation cost in this case is less expensive. The labeling information of the multiple sources is extracted using DoA-based features as we propose in Chapter 4. Combining this information yields a diarization system of higher performance where questions on speaker identity, localization and activity are well achieved.

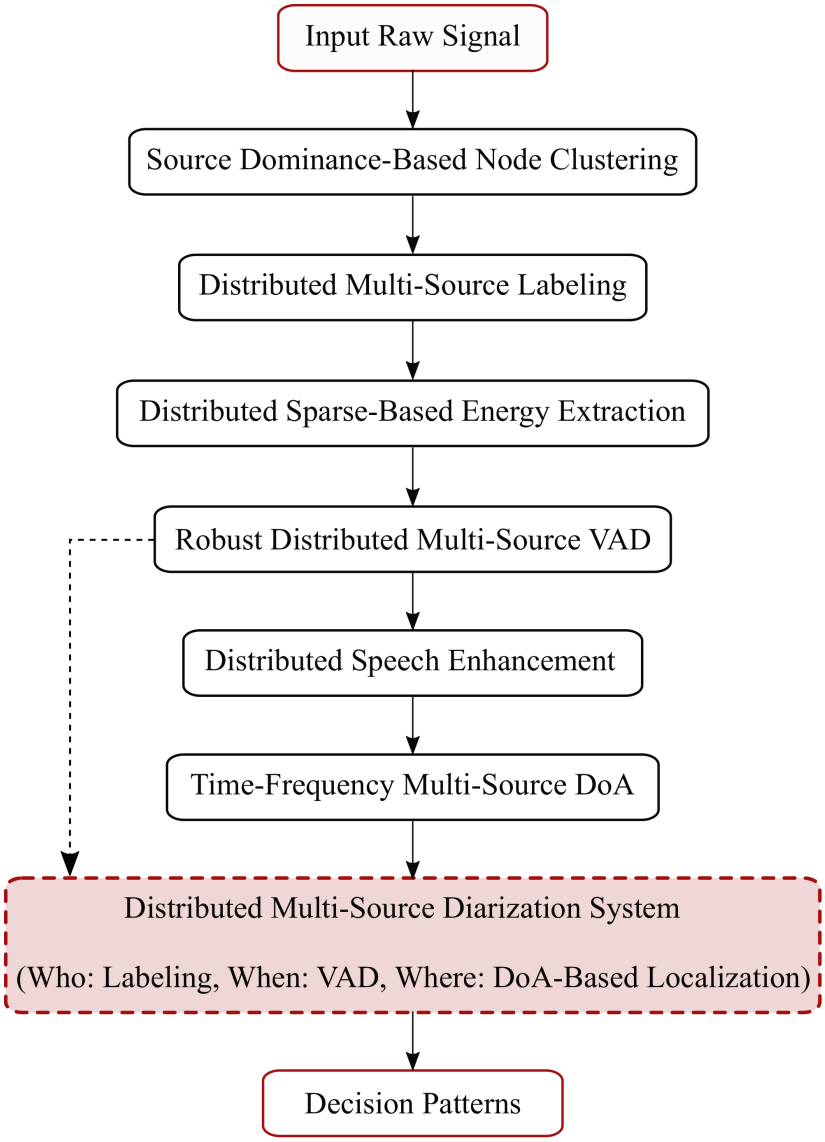


Figure 5.1: Approach for solving the distributed multi-speaker diarization problem.

Appendix

A.1 Closed-Form Solution for a Quadratic Optimization Based on Component-Wise Thresholding Rule

Extracting sparse right rotations based on a penalized SVD regression as described in Chapter 3 suggests solving the regression to obtain proper sparse right singular vectors from the SVD layers. This can be done, for instance, using the algorithm proposed in [142]. However, [118] derives an efficient algorithm that well explores the structure of SSVD. In the following, we present the necessary derivations for the update of the Lasso-based sparse right rotations simply based on a thresholding rule suggested in [118].

Let the following formulation be the penalized sum-of-squares criterion

$$\|\mathbf{Y} - \sigma\mathbf{u}\mathbf{v}^\top\|^2 + \lambda_\mathbf{v}\Phi(\sigma\mathbf{v}). \tag{A.1}$$

As described in Chapter 3, $\lambda_\mathbf{v}$ is the non-negative penalty coefficient. Tweaking this parameter controls the desired amount of sparseness, which depends on the data-driven application and the characteristics of the signal. An explicit sparseness constraint imposed in Eq. (A.1) is defined with $\Phi(\sigma\mathbf{v})$, which is a sparsity-inducing penalty term. In our study, we use a Lasso-based regularization sparse model. Based on the Lasso penalty and a fixed right rotations $\mathbf{u}$, minimizing Eq. (A.1) with respect to the right singular vectors $\mathbf{v}$ refers to minimizing

$$\|\mathbf{Y} - \sigma\mathbf{u}\mathbf{v}^\top\|^2 + \lambda_\mathbf{v}\sum_{n=1}^{N}|\sigma v[n]|. \tag{A.2}$$

σ is a non-negative scalar, $\mathbf{u}$ is a unit M-vector, and $\mathbf{v}$ is a unit N-vector. Let $\breve{v}[n]$ replace the product $\breve{v}[n] = \sigma v[n]$. Then, for a fixed $\mathbf{u}$, the minimization of Eq. (A.2) corresponds to minimizing

$$\|\mathbf{Y}\|^2 + \sum_{n=1}^{N}\{\breve{v}[n]^2 - 2\breve{v}[n](\mathbf{Y}^\top\mathbf{u}) + \lambda_\mathbf{v}|\breve{v}[n]|\}. \tag{A.3}$$

Solving the regression for a Lasso-based sparse right rotations extracted from the SVD layers can be fulfilled using component-wise thresholding rules developed in [118]. We use the lemma introduced in [118] to deduce a closed-form solution for Lasso-based minimization problem presented in our objective in Eq. (A.3).

Lemma 1. *Given Eq. (A.3), a minimiser for the formula of the form* $\gamma^2 - 2y\gamma + 2\lambda|\gamma|$ *is* $\check{\gamma} = \text{sgn}(y)(|y| - \lambda)_+$. *A simple thresholding rule can be defined as*

$$\check{\gamma} = \begin{cases} y - \lambda, & \text{if } y > \lambda \\ y + \lambda, & \text{if } y < -\lambda \\ 0, & \text{otherwise} \end{cases} \tag{A.4}$$

In Lemma 1, we set y to represent the nth component of $\mathbf{Y}^\top \mathbf{u}$. Moreover, in our case $\lambda = \frac{\lambda_{\mathbf{v}}}{2}$. Then, a minimal $v[n]$ for Eq. (A.3) is given by

$$\check{v}[n] = \text{sgn}\{(\mathbf{Y}^\top \mathbf{u})_n\}(|(\mathbf{Y}^\top \mathbf{u})_n| - \frac{\lambda_{\mathbf{v}}}{2})_+. \tag{A.5}$$

Next, an update of the scalar σ is done by $\sigma = \|\check{\mathbf{v}}\|$, with $\check{\mathbf{v}} = (\check{v}[1], \ldots, \check{v}[N])$. Similarly, $\mathbf{v}$ is then updated with $\mathbf{v} = \frac{\check{\mathbf{v}}}{\sigma}$.

List of Acronyms

ANN	artificial neural network
ASR	automatic speech recognition
BSS	blind source separation
DoA	direction-of-arrival
DM-VAD	distributed multi-speaker voice activity detection
EMD	empirical mode decomposition
EVD	eigenvalue decomposition
FC	fusion center
HANDiCAMS	heteregenous ad hoc networks for distributed, cooperative, and adaptive multimedia signal processing
HOS	higher order statistics
ICA	independent component analysis
KR-MUSIC	Khatri Rao-MUSIC
LARS	least angle regression and shrinkage
Lasso	Least absolute shrinkage and selection operator
LONAS	locating nodes around sources
MAD	mean absolute deviation
mad	median absolute deviation
M-NICA	multiplicative non-negative independent component analysis
MSA	modulation spectrum analysis
MTF	modulation transfer function
MUSIC	multiple signal classification
MWF	multi-channel Wiener filtering
NBSS	non-negative blind source separation

NICA	non-negative independent component analysis
NMF	non-negative matrix factorization
NPCA	non-negative principal component analysis
PCA	principal component analysis
pdf	probability density function
RMSE	root mean square error
SDR	signal to distortion ratio
SINR	signal-to-interference-plus-noise-ratio
SNR	signal-to-noise-ratio
STFT	short-time Fourier transform
SMM-NICA	sparse median-based multiplicative non-negative independent component analysis
SIR	source to interference ratio
SVM	support vector machine
UCA	uniform circular array
ULA	uniform linear array
VAD	voice activity detection
WASN	wireless acoustic sensor network
WCSS	within-cluster sum of squares
WOLA	weighted overlap-add
WSN	wireless sensor network

List of Symbols

$^\top$	transposition
$^\mathsf{H}$	Hermitian transposition
$^{-1}$	matrix inversion
$\emptyset$	empty vector symbol
$\varnothing$	empty matrix symbol
$\mathrm{sgn}(\cdot)$	sign function
$\log(\cdot)$	logarithm function
$\mathbf{1}_N$	N-dimensional column vector of ones
$\#(\bullet)$	cardinality symbol
$\lvert\cdot\rvert$	absolute value
$\|$	concatenation symbol
$P(\cdot)$	probability of
$\mathbf{I}_N$	identity matrix
M	number of microphones in the network
$\tilde{s}_q[\eta]$	qth signal sample at time instant η
$\tilde{\mathbf{s}}_q$	signal vector emitted from source q
L	signal block length where stationarity is assumed
$\mathbf{s}[n]$	Q-dimensional vector at frame n
$\tilde{y}_{k,m}$	observed signal at microphone m of device k
$\mathbf{y}[n]$	observed M-dimensional energy vector at all K devices
$\mathbf{A}$	mixing matrix
$\boldsymbol{\omega}[n]$	additive white noise process
$\mathcal{B}_q$	qth cluster of nodes
$\bar{\mathbf{Y}}$	matrix of received energies at every microphone m and frame n
$\bar{\mathbf{S}}$	matrix of recovered Q energy sources at all N frames
$\mathbf{U}$	left rotation matrix
$\mathbf{V}$	right rotation matrix
$\boldsymbol{\Sigma}$	scaling matrix of singular values
$\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2$	diagonal matrices of tweaking parameters for the decorrelation function
$D\{\cdot\}$	sets off-diagonal elements of a matrix to zero
$\grave{\mathbf{S}}$	matrix of mean values
$\mathbf{C}_{\bar{\mathbf{S}}}$	sample covariance matrix

$\mathbf{y}_{\mathcal{B}_q}[n]$	$\lvert\mathcal{B}_q\rvert$-dimensional instantaneous energy vector at frame n
$\mathbf{a}_{\mathcal{B}_q}$	power attenuation between source q and $k \in \mathcal{B}_q$
$\bar{\mathbf{Y}}_{\mathcal{B}_q}$	collected energy matrix at nodes $k \in \mathcal{B}_q$ relative to source q
$\mathbf{s}_q$	qth source dominant vector of length N
$c_{\mathbf{s}_q}$	variance of $\mathbf{s}_q$
$\grave{\lambda}_1, \grave{\lambda}_2$	weighting scalars
$\grave{\mathbf{s}}_q$	vector of mean values
$\boldsymbol{v}_q^{(n)}$	qth energy-based feature vector at frame n
$\hat{\mathbf{c}}_j^{(q)}$	centroid vector at class j and source q
$\mathbf{C}_j$	feature cluster j
$f^\star$	index of the energy-based speech features
$\boldsymbol{L}_{f^\star}$	matrix of measured distances to the active/non-active centroids
$\mu_{\boldsymbol{L}_{f^\star}}$	average distance for each feature $f^\star$
$w_{f^\star}$	feature related weights
$t_j(\cdot)$	feature cluster membership determination function
$\delta_q^{(n)}$	binary voice activity decision rule
$\mathbf{v}$	right singular vector
σ	singular value
$\mathbf{u}$	left singular vector
$\Phi(\cdot)$	ℓ_1 regularization function
$\lambda_{\mathbf{v}}$	non-negative penalty parameter
$g(\cdot)$	degree of sparsity function
$\mathbf{V}^{\mathcal{S}}$	matrix of sparse singular vectors
$\mathbf{Y}_{\text{SSVD}}$	sparse lower-rank matrix
$\ddot{\mathbf{S}}$	matrix of median values
$\sigma_{\boldsymbol{\omega}}^2$	additive noise power
$\mathbf{C}_*$	robust covariance matrix based on $t_\nu M$-estimator
$u_\nu(\cdot)$	weight function of the t_νM-estimator
$\mathbf{C}_*^{\alpha,\beta}$	regularized robust covariance matrix
$\mu = 5$	step size
L_{DFT}	DFT size
λ_{DANSE}	forgetting factor
$\mathcal{L}_{\mathbf{v}}$	set of $\lambda_{\mathbf{v}}$ parameters
$\hat{\mathcal{Z}}_{\mathbf{v}}^{\lambda_{\mathbf{v}}}(\cdot)$	subspace of non-zero indicators n of $\mathbf{v}$ for a specific $\lambda_{\mathbf{v}}$

$\lambda_{\mathbf{v}}^{\min}$	minimal penelization value that ensures a stable set $\hat{\mathcal{Z}}_{\mathbf{v}}$
$\hat{\mathbf{R}}_{q,j}$	robust covariance matrix of speech/non-speech feature's distributions
$M_j(\cdot)$	robust Mahalanobis distance for the speech/silence classes
$\mathbf{d}_q$	binary decision pattern for source q based on robust Mahalanobis classifier
$\mathbf{s}[\eta]$	impinging Q signals at snapshot η
$\mathbf{A}_k^\star$	complex-valued steering matrix
$\mathbf{a}_k^\star(\theta_q)$	steering vector of source q and node k
$\mathbf{y}[\eta]$	network-wide received signal at instant η using a narrowband model
$\check{\mathbf{y}}[n,f]$	network-wide received signal at frame n and frequency f using a wideband model
$\mathbf{A}^\star(f)$	frequency dependent array response matrix
$\mathbf{U}_{\text{noise}}^{\mathsf{H}}(f)$	noise subspace at a frequency f
$P_{\text{KR-MUSIC}}(\theta,f)$	frequency dependent spatial KR-MUSIC spectrum
$\hat{\theta}_q$	estimated DoA angle for source q based on all $\hat{\theta}_q(f)$
$\mathcal{N}_k$	number of neighbors of k
$\hat{\sigma}_{\hat{\theta}_q}$	
$\sigma_{\boldsymbol{\omega}}^2$	variance of the noise process
f_s	sampling frequency
d_{mic}	inter-sensor spacing
λ_{w}	wavelength
c_{w}	propagation speed of a wave
f_{w}	frequency of a wave
$\hat{\mathbf{R}}_n^{\text{loc}}$	sensor local covariance matrix at a stationary frame n
k', k''	nodes in the neighborhood of k
$\boldsymbol{c}_q^{(k)}(i)$	qth computed centroid at node k and iteration i

Bibliography

[1] 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.

[2] 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.

[3] E. Verteletskaya and K. Sakhnov, "Voice activity detection for speech enhancement applications," *Acta Polytech.*, vol. 50, no. 4, pp. 100–105, 2010.

[4] 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.

[5] 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)*, Mar. 2017, pp. 4611–4615.

[6] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *18th IEEE Symp. Commun. Veh. Technol. Benelux (SCVT)*. IEEE, 2011, pp. 1–6.

[7] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks – part I: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277 –5291, Oct. 2010.

[8] ——, "Distributed adaptive node-specific signal estimation in fully connected sensor networks – part II: simultaneous & asynchronous node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5292–5306, 2010.

[9] 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.

[10] 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.

[11] A. Bertrand and M. Moonen, "Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2196–2210, 2011.

[12] L. Lanbo, Z. Shengli, and C. Jun-Hong, "Prospects and problems of wireless communication for underwater sensor networks," *Wirel. Commun. Mobile Comput.*, vol. 8, no. 8, pp. 977–994, 2008.

[13] M. F. F. B. Ismail and L. W. Yie, "Acoustic monitoring system using wireless sensor networks," *Procedia Eng.*, vol. 41, pp. 68–74, 2012.

[14] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proc. IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[15] A. Bertrand and M. Moonen, "Robust distributed noise reduction in hearing aids with external acoustic sensor nodes," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 12, 2009.

[16] 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.

[17] 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.

[18] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010. [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/sista/abertran/papers_website/IWAENC10.html

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[20] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 4. IEEE, 1979, pp. 208–211.

[21] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[22] J. Szurley, A. Bertrand, I. Moerman, and M. Moonen, "Network topology selection for distributed speech enhancement in wireless acoustic sensor networks," in *IEEE Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[23] A. Hassani, J. Plata-Chaves, A. Bertrand, and M. Moonen, "Multi-task wireless acoustic sensor network for node-specific speech enhancement and DOA estimation," in *IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, 2016, pp. 1–5.

[24] A. Hassani, J. Plata-Chaves, M. H. Bahari, M. Moonen, and A. Bertrand, "Multi-task wireless sensor network for joint distributed node-specific signal enhancement, LCMV beamforming and DOA estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 518–533, 2017.

[25] 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 Process.*, vol. 107, pp. 4–20, 2015.

[26] D. Ampeliotis, N. Bogdanovic, and K. Berberidis, "Coalitional games for a distributed signal enhancement application," in *IEEE 23rd Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 1885–1889.

[27] A. Hassani, A. Bertrand, and M. Moonen, "Low-rank approximation-based distributed node-specific signal estimation in a fully-connected wireless sensor network," in *40th IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 2839–2843.

[28] ——, "Distributed signal subspace estimation based on local generalized eigenvector matrix inversion," in *23rd Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2015, pp. 1386–1390.

[29] A. Bertrand and M. Moonen, "Distributed canonical correlation analysis in wireless sensor networks with application to distributed blind source separation," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4800–4813, 2015.

[30] F. de la Hucha Arce, F. Rosas, M. Moonen, M. Verhelst, and A. Bertrand, "Energy-vs-performance trade-offs in speech enhancement in wireless acoustic sensor networks," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2015, pp. 1561–1565.

[31] J. Szurley, A. Bertrand, and M. Moonen, "Distributed adaptive node-specific signal estimation in heterogeneous and mixed-topology wireless sensor networks," *Signal Process.*, vol. 117, pp. 44–60, 2015.

[32] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2557–2572, 2016.

[33] J. Szurley, A. Bertrand, and M. Moonen, "Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Trans. Signal Inf. Process. over Networks*, vol. 3, no. 1, pp. 130–144, 2017.

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

[35] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. Zoubir, "Heterogeneous and multi-task wireless sensor networks-algorithms, applications and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 450–465, 2017.

[36] L. K. Hamaidi, M. Muma, and A. M. Zoubir, "Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction." in *Proc. 25th IEEE Eur. Signal Process. Conf. (EUSIPCO'17)*.

[37] M. H. Bahari, J. Plata-Chaves, A. Bertrand, and M. Moonen, "Distributed labelling of audio sources in wireless acoustic sensor networks using consensus and matching," in *Proc. 24th IEEE Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 2345–2349.

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

[39] L. F. Villa, C. Salazar, O. Quintero *et al.*, "A simple but efficient voice activity detection algorithm through Hilbert transform and dynamic threshold for speech pathologies," in *J. Physics: Conf. Ser.*, vol. 705, no. 1, 2016.

[40] 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.

[41] T. Fukuda, O. Ichikawa, and M. Nishimura, "Improved voice activity detection using static harmonic features," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 4482–4485.

[42] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.

[43] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," *6th Int. Conf. Signal Process.*, vol. 1, pp. 464–467, Aug. 2002.

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

[45] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," *J. Signal Process. Syst.*, vol. 82, no. 2, pp. 163–173, 2016.

[46] Y. Kanai and M. Unoki, "Robust voice activity detection using empirical mode decomposition and modulation spectrum analysis," in *8th Int. Symposium Chinese Spoken Language Process. (ISCSLP)*, Dec. 2012, pp. 400–404.

[47] T. Higuchi and H. Kameoka, "Unified approach for underdetermined BSS, VAD, dereverberation and DOA estimation with multichannel factorial HMM," in *IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, 2014, pp. 562–566.

[48] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *IEEE Int. Conf. Comput. Commun. Technol., (RIVF)*, 2009, pp. 1–8.

[49] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 252–256, 2016.

[50] J. Freudenberger and S. Stenzel, "Time-frequency dependent voice activity detection based on a simple threshold test," *IEEE Stat. Signal Process. Workshop (SSP)*, pp. 541–544, June 2011.

[51] S. Valipour, F. Razzazi, A. Fard, and N. Esfandian, "A novel voice activity detector for noisy environments using Gaussian clustering of noise in spectro-temporal domain," *2nd Int. Conf. Comput. Intel., Model. Simulation (CIMSiM)*, pp. 345–350, Sep. 2010.

[52] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Process.*, vol. 3, no. 3, pp. 205–210, 2009.

[53] Y. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP '01)*, vol. 2, pp. 737–740, 2001.

[54] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.

[55] J. Wu and X.-L. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, May 2011.

[56] ——, "An efficient voice activity detection algorithm by combining statistical model and energy detection," *EURASIP J. Adv. Signal Process.*, July.

[57] 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, Jun. 2006.

[58] I. Hwang, H.-M. Park, and J.-H. Chang, "Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection," *Comput. Speech Language*, vol. 38, pp. 1–12, 2016.

[59] X. Li, R. Horaud, L. Girin, and S. Gannot, "Voice activity detection based on statistical likelihood ratio with adaptive thresholding," in *IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[60] L. Tan, B. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," *IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, pp. 4466–4469, Mar. 2010.

[61] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance Gamma distribution," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 4, pp. 1129–1134, May 2007.

[62] J. Ramirez, J. Segura, J. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 8, pp. 2177–2189, Nov. 2007.

[63] J. M. Górriz, J. Ramírez, E. W. Lang, and C. G. Puntonet, "Hard C-means clustering for voice activity detection," *Speech Commun.*, vol. 48, no. 12, pp. 1638–1649, 2006.

[64] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3, pp. 271–287, 2004.

[65] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE Region 10 Conf. Comput. Commun. Control Power Eng. (TENCON'93)*, vol. 3, 1993, pp. 321–324.

[66] M. W. Mak and H. B. Yu, "Robust voice activity detection for interview speech in NIST speaker recognition evaluation," *Proc. APSIPA ASC*, pp. 1–4, 2010.

[67] T. Tran, W. Cowley, and A. Pollok, "Multi-speaker beamforming for voice activity classification," in *IEEE Commun. Theor. Workshop (AusCTW)*, 2013, pp. 116–121.

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

[69] 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 *Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA)*, 2011, pp. 92–97.

[70] 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.

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

[72] R. R. Brooks, P. Ramanathan, and A. M. Sayeed, "Distributed target classification and tracking in sensor networks," *Proc. IEEE*, vol. 91, no. 8, pp. 1163–1171, 2003.

[73] B. Koetz, F. Morsdorf, S. Van der Linden, T. Curt, and B. Allgöwer, "Multi-source land cover classification for forest fire management based on imaging spectrometry and LiDAR data," *Forest Ecol. Manag.*, vol. 256, no. 3, pp. 263–271, 2008.

[74] Z. J. Towfic, J. Chen, and A. H. Sayed, "On distributed online classification in the midst of concept drifts," *Neurocomputing*, vol. 112, pp. 138–152, 2013.

[75] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Research*, vol. 11, no. May, pp. 1663–1707, 2010.

[76] ——, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 707–724, 2011.

[77] Y. Lu, V. Roychowdhury, and L. Vandenberghe, "Distributed parallel support vector machines in strongly connected networks," *IEEE Trans. Neural Networks*, vol. 19, no. 7, pp. 1167–1178, 2008.

[78] G. Jagannathan and R. N. Wright, "Privacy-preserving distributed K-means clustering over arbitrarily partitioned data," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2005, pp. 593–599.

[79] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, 2003, pp. I–I.

[80] S. Datta, C. R. Giannella, and H. Kargupta, "Approximate distributed K-means clustering over a peer-to-peer network," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1372–1388, 2009.

[81] F. K. Teklehaymanot, M. Muma, B. Béjar, P. Binder, A. Zoubir, and M. Vetterli, "Robust diffusion-based unsupervised object labelling in distributed camera networks," in *IEEE AFRICON*, Sept. 2015, pp. 1–6.

[82] P. Binder, M. Muma, and A. M. Zoubir, "Robust and adaptive diffusion-based classification in distributed networks," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 34, 2016.

[83] F. Teklehaymanot, M. Muma, and A. M. Zoubir, "Adaptive diffusion-based track assisted multi-object labeling in distributed camera networks," in *Eur. Signal Process. Conf. (EUSIPCO)*, 2017, pp. 2363–2367.

[84] E. Oja and M. Plumbley, "Blind separation of positive sources using non-negative PCA," in *Proc. 4th Int. Symposium Independent Component Anal. Blind Signal Separation (ICA)*, 2003, pp. 11–16.

[85] A. Bertrand and M. Moonen, "Blind separation of non-negative source signals using multiplicative updates and subspace projection," *Signal Process.*, vol. 90, no. 10, pp. 2877–2890, 2010.

[86] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[87] L. Li, J. A. Chambers, C. G. Lopes, and A. H. Sayed, "Distributed estimation over an adaptive incremental network based on the affine projection algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 151–164, 2010.

[88] L. Li, A. Scaglione, and J. H. Manton, "Distributed principal subspace estimation in wireless sensor networks," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 725–738, 2011.

[89] Y.-A. Le Borgne, S. Raybaud, and G. Bontempi, "Distributed principal component analysis for wireless sensor networks," *Sensors*, vol. 8, no. 8, pp. 4821–4850, 2008.

[90] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen–Loeve transform," *IEEE Trans. Inf. Theor.*, vol. 52, no. 12, pp. 5177–5196, 2006.

[91] A. Bertrand and M. Moonen, "Distributed adaptive estimation of covariance matrix eigenvectors in wireless sensor networks with application to distributed PCA," *Signal Process.*, vol. 104, pp. 120–135, 2014.

[92] T.-W. Lee, "Independent component analysis," in *Independent Component Anal.* Springer, 1998, pp. 27–66.

[93] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[94] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[95] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," in *Nat. Image Stat.* Springer, 2009, pp. 151–175.

[96] A. Hyvärinen, "Independent component analysis: recent advances," *Phil. Trans. R. Soc. A*, vol. 371, no. 1984, 2013.

[97] ——, "Survey on independent component analysis," *Neural Comput. Surveys*, pp. 94–128.

[98] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[99] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, 4th ed. Academic Press, 2008.

[100] S. Theodoridis, "Machine learning: A signal and information processing perspective," 2014.

[101] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert. Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.

[102] S. Arora, P. Raghavan, and S. Rao, "Approximation schemes for Euclidean K-medians and related problems," in *Proc. 13th annual ACM symposium Theor. computing*. ACM, 1998, pp. 106–113.

[103] Z. Lu and A. M. Zoubir, "Source enumeration in array processing using a two-step test." *IEEE Trans. Signal Process.*, vol. 63, no. 10, pp. 2718–2727, 2015.

[104] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4, 2004, pp. 2529–2533.

[105] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," 2008.

[106] M. D. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 534–543, 2003.

[107] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.

[108] M. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 177–180, 2002.

[109] M. D. Plumbley and E. Oja, "A "nonnegative PCA" algorithm for independent component analysis," *IEEE Trans. Neural Networks*, vol. 15, no. 1, pp. 66–76, 2004.

[110] E. Oja, J. Karhunen, H. Valpola, J. Särelä, M. Inki, A. Honkela, A. Ilin, K. Raju, T. Ristaniemi, and E. Bingham, "Independent component analysis and blind source separation," Technical report, Helsinki University of Technology, Tech. Rep., 2003.

[111] H. Y. Li, G. L. Ren, and B. J. Xiao, "Image denoising algorithm based on independent component analysis," in *World Congr. Software Eng. (WCSE'09.)*, vol. 4, May 2009, pp. 465–469.

[112] Y. Huang, M. Li, C. Lin, and L. Tian, "Gabor-based kernel independent component analysis for face recognition," in *6th Int. Conf. Intell. Inf. Hiding and Multimedia Signal Process. (IIH-MSP)*, Oct. 2010, pp. 376–379.

[113] H. Y. Li, Q. H. Zhao, G. L. Ren, and B. J. Xiao, "Speech enhancement algorithm based on independent component analysis," in *5th Int. Conf. Nat. Comput.*, vol. 2, Aug. 2009, pp. 598–602.

[114] D. D. Lee and H. S. Seung, "Learn. the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[115] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, pp. 267–288, 1996.

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

[117] ——, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, no. Nov., pp. 1457–1469, 2004.

[118] 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.

[119] J. Xi and A. Li, "Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 4, pp. 656–668, Jul./Aug. 2015.

[120] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Adv. Neural Inf. Process. Syst.*, 2006, pp. 801–808.

[121] L. L. Scharf, *Statistical Signal Processing.* Addison-Wesley Reading, MA, 1991, vol. 98.

[122] V. D. Vrabie, J. I. Mars, and J. L. Lacoume, "Modified singular value decomposition by means of independent component analysis," *Signal Process.*, vol. 84, no. 3, pp. 645–652, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168403003475

[123] H. Zou, T. Hastie, R. Tibshirani *et al.*, "On the degrees of freedom of the Lasso," *Ann. Stat.*, vol. 35, no. 5, pp. 2173–2192, 2007.

[124] G. Schwarz *et al.*, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[125] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, Jul. 2012.

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

[127] 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.

[128] 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.

[129] A. Vesa, "Direction of arrival estimation using MUSIC and Root-MUSIC algorithm," in *18th Telecommun. Forum, Pg*, 2010, pp. 582–585.

[130] M. Cheney, "The linear sampling method and the MUSIC algorithm," *Inverse Probl.*, vol. 17, no. 4, p. 591, 2001.

[131] H. Hwang, Z. Aliyazicioglu, M. Grice, and A. Yakovlev, "Direction of arrival estimation using a Root-MUSIC algorithm," in *Proc. Int. MultiConf. Eng. Comput. Scientists*, vol. 2. Citeseer, 2008.

[132] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[133] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: a Khatri–Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, 2010.

[134] ——, "DOA estimation of quasi-stationary signals via Khatri-Rao subspace," in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2009, pp. 2165–2168.

[135] D. Feng, M. Bao, Z. Ye, L. Guan, and X. Li, "A novel wideband DOA estimator based on Khatri–Rao subspace approach," *Signal Process.*, vol. 91, no. 10, pp. 2415–2419, 2011.

[136] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 4, pp. 823–831, 1985.

[137] E. D. Di Claudio and R. Parisi, "Waves: Weighted average of signal subspaces for robust wideband direction finding," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2179–2191, 2001.

[138] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "Tops: New DOA estimator for wideband signals," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1977–1989, 2006.

[139] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: A Khatri-Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, April 2010.

[140] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[141] J. Allen and D. Berkley, "Image method for efficiently simulating small–room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 1, pp. 943–950, Apr. 1979.

[142] H. Zou, "The adaptive Lasso and its oracle properties," *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

Curriculum Vitae

Name:	Lala Khadidja Hamaidi
Date of birth:	22.11.1989
Place of birth:	Algiers (Algeria)
Family status:	Single

Education

Since 10/2013	Technische Universität Darmstadt (Darmstadt, Germany) Electrical Engineering and Information Technology, Telecommunications Institute.
07/2012	Master of Science (M.Sc): "Discovering Association Rules From Semantic WebData" Sciences and Technologies of Information and Communication (STIC), Computer Science Department, Faculty of Engineering Sciences, Annaba University of Technology UBMA-Algeria.
07/2010	Bachelor of Science (B.Sc): "Creating a Dynamic Website for the Preparatory School of Sciences and Technologies EPST Annaba, Algeria" Mathematics and Informatics (MIAS), Computer Science Department, Faculty of Engineering Sciences, Annaba University of Technology UBMA-Algeria.
07/2007	High school degree (Abitur) at Lycée El-Feth, Blida, Algeria

Work experience

since 10/2013	Research Associate at Signal Processing Group Technische Universität Darmstadt.
12/2016 - 03/2017	Researcher in an industry project with AGT, Darmstadt, Germany.
10/2013 - 09/2016	PhD Student, Scholarship holder at Graduate School of Excellence Computational Engineering (CE) Technische Universität Darmstadt.

Erklärung laut §9 der Promotionsordnung

Ich versichere hiermit, dass ich die vorliegende Dissertation allein und nur unter Verwendung der angegebenen Literatur verfasst habe. Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.