DISTRIBUTED ROBUST LABELING OF AUDIO SOURCES IN HETEROGENEOUS WIRELESS SENSOR NETWORKS

Symeon Chouvardas[1], *Michael Muma*[2], *Khadidja Hamaidi*[2,3], *Sergios Theodoridis*[1], *Abdelhak M. Zoubir*[2,3]

[1] Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis Ilissia, 157 84 Athens, Greece
Email: {schouv,stheodor}@di.uoa.gr

[2] Signal Processing Group, Technische Universität Darmstadt, Merckstraße 25, 64283 Darmstadt, Germany
Email: {muma, zoubir}@spg.tu-darmstadt.de

[3] Graduate School CE, Technische Universität Darmstadt, Dolivostr. 15, D-64293 Darmstadt Germany
Email: hamaidi@gsc.tu-darmstadt.de

ABSTRACT

A novel algorithm for distributed labeling of speech sources is proposed. We consider a wireless sensor network comprising devices that are equipped with multiple microphones, which can "hear" a number of speech signals. The labeling task is performed in a decentralized fashion with a new two–step approach. The first step corresponds to the distributed extraction of proper source–specific features from the mixed signals. In the second step, these features are exploited via a distributed unsupervised learning technique. We present approaches that can be used in hierarchically organized or in non-hierarchically organized network configurations. Numerical examples using real data display the performance of the proposed technique.

Index Terms— distributed clustering, speech labeling, wireless sensor network, cooperative signal processing, feature extraction

1. INTRODUCTION

Much research in cooperative signal processing for wireless sensor networks (WSN) has focused on solving a common signal processing task given a distributed set of devices that are equipped with sensing, computing and communication capabilities. A new and emerging research direction concerns the question how multiple devices can cooperate in multiple tasks (MDMT). Consider, for example, MDMT distributed audio signal enhancement in a public area, such as an airport, a stadium, etc. Here, the multi-sensor devices (e.g., smart-phones, 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. Clearly, in this case, nodes may benefit from a cooperation, even though their source of interest differs. A crucial step in order to achieve a benefit, e.g., a better node-specific signal enhancement, in such an MDMT application, is the common *unique labeling* of all relevant speech sources that are observed by the WSN. In this setting, the labeling information must be extracted locally from the mixtures of received signals. To the best of our knowledge, there exists no approach to solve the distributed labeling problem up to date, and novel approaches are urgently sought for to enable MDMT.

This work was supported by the project HANDiCAMS which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 323944. The work of K. Hamaidi is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

One may think of applying distributed blind source separation (BSS) techniques, e.g., [1], followed by source–specific feature extraction, given the separated signals. However, distributed BSS is a very challenging research direction of its own and constitutes a computationally demanding task, that requires the distributed computation of higher order statistical moments. Furthermore, the exchange of raw sensor signals is often prohibited in practical scenarios due to communication (bandwidth/energy) constraints. New simple but informative features, that do not require BSS, are thus considered for the distributed/cooperative unsupervised learning.

Related work: While some work has been done on distributed classification [2, 3, 4, 5], to the best of our knowledge, the distributed labeling task has not yet been addressed. A distributed algorithm for supervised learning, in the presence of a fusion center has been proposed in [5] and totally decentralized schemes in [3, 2]. Distributed algorithms for unsupervised learning have been proposed in [6, 4].

All the previously mentioned efforts assume that the features are available at each node. In the current study, however, we consider the more realistic scenario where the features are not available a–priori. This is a rather challenging task due to the fact that the various speech signals, to be labeled, are mixed.

Contributions: This paper presents a new framework of distributed labeling of speech sources in a WSN. We present two types of feature extraction approaches, which estimate the source-specific features from a received mixture of sound sources impinging onto the microphone array of each device. We consider a hierarchical network in the first proposed feature extraction approach. Here the energy signatures of each source are extracted by a computationally attractive non–negative independent component analysis in each sub-network. The second proposed feature extraction approach operates in non–hierarchical networks by exploiting similarities in the frequency bands of the subspace decompositions at each node, that produce reliable direction-of-arrival estimates of the speech sources. We next propose a distributed centroid clustering scheme for both feature extraction approaches. We also provide a realistic experiment that evaluates the performance of the proposed methods for different noise levels, in a 20 node WSN with three speech sources.

The paper is organized as follows. Section 2 provides the problem formulation. Section 3 is dedicated to introducing, in detail, two new features that are used for the labeling task. We consider both hierarchical and non-hierarchical networks for the feature estimation. Section 4 describes how the distributed clustering is performed based on the previously extracted features. Section 5 evaluates the performance of the proposed methods in a practical scenario, and Section 6 concludes the paper.

2. PROBLEM FORMULATION

Consider a public area, such as an airport, a subway station hall, a stadium etc., containing N sound sources and D devices, each equipped with J' microphones. The total number of microphones of the network will be denoted by J. The devices are willing and capable to cooperate in order to perform a device-specific signal processing task, e.g., speech enhancement. An example of such a scenario is illustrated in Fig.1, which depicts a use-case of a 20×10 meter room with a reverberation time of $T60 = 0.3$ seconds, containing $N = 3$ sound sources, $D = 20$ devices, each equipped with $J' = 3$ microphones in a vertically oriented uniform linear array configuration with an inter-sensor spacing of 1.5 centimeters.

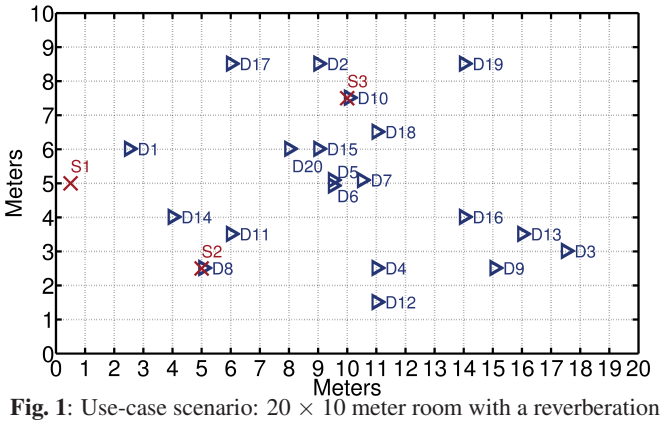


Fig. 1: Use-case scenario: 20×10 meter room with a reverberation time of $T60 = 0.3$ seconds. $D = 20$ devices, each equipped with $J' = 3$ microphones, cooperate in order to label the $N = 3$ speech sources. The microphone signals are sampled at $f_s = 16$ kHz.

Obviously, cooperation among the nodes requires that the devices exchange information related to a certain signal of interest. To that end, the various speech signals of the network need to be *labeled*. More importantly, the labels corresponding to speech signals should be *the same* throughout the whole network. In the current study we focus on the scenario, where a fusion center for centralized processing is not present and the devices form a fully decentralized network.

3. DISTRIBUTED LABELING VIA CLUSTERING

For the labeling of the different speech sources throughout the network, we propose a two–step approach. The steps are the following:

1. Feature extraction at each device for each speech source of the network.
2. Distributed clustering using the previously computed features.

We next detail the above steps, which constitute our proposed approach. Regarding the feature extraction, ideally we seek for 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. In our approach, we do not consider distributed source separation, for the reasons stated in Section 1.

After computing proper features, we employ a distributed clustering scheme. The goal of distributed clustering algorithms is to form the clusters in a way that 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 (see for example [6, 7]), where the nodes average the computed centroids of the clusters and consensus on the centroids is achieved. 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.

3.1. Hierarchical Feature Extraction: Correlating the Separated Energy–Signatures

The first feature, which will be extracted, is related to the energy–signature of the speech signals, which can be expected to be similar across the nodes of the network.

To be more specific, let us denote the speech signals by $\tilde{s}_n[i],\ n = 1, \ldots, N$, where i stands for the (sampling) time index. The instantaneous energy of signal n, given a block of length L, at sample time iL is equal to

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

In a similar fashion, we define the instantaneous energy at microphone j by

$$y_j[i] = \sum_{l=0}^{L-1} \tilde{y}_j^2[iL + l], \tag{2}$$

where $\tilde{y}_j[\cdot]$ is the signal of the j–th microphone. Stacking the energy related signals in vectors, while assuming that the signals $\tilde{s}_n$, $n = 1, \ldots, N$, are mutually independent and neglecting any reverberation effects over the block edges (see [8] for more details), it holds that

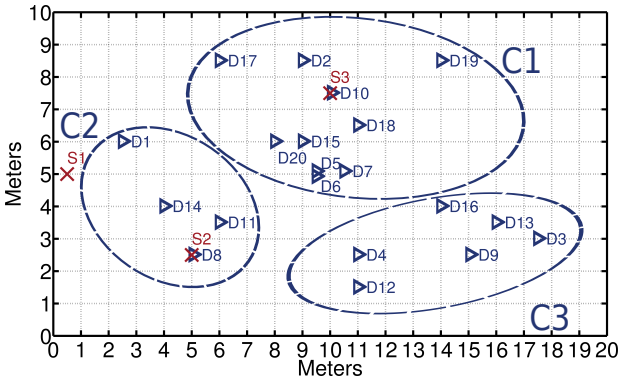
$$\boldsymbol{y}[i] \approx \boldsymbol{A}\boldsymbol{s}[i], \tag{3}$$

where $\boldsymbol{s}[i] = [s_1[i], \ldots, s_N[i]]^T$, $\boldsymbol{y}[i] = [y_1[i], \ldots, y_J[i]]^T$, and $\boldsymbol{A}$ is the $J \times N$ mixing matrix, whose j, n–th element describes the power attenuation between the respective speech sources and the microphones. Algorithms, which compute both $\boldsymbol{A}$ and $\boldsymbol{s}[i]$, have been proposed in [9], by employing the Non–Negative Principal Component Analysis (NNPCA), and in [10, 8] by using the Non–Negative Independent Component Analysis (NNICA). In the current study, we focus on the latter approach. Since it is expected that all nodes extract "similar" estimates of the sources' energies as a function of time (something that is verified by the experiments), we show that these energy signatures can serve as features in a distributed labeling task.

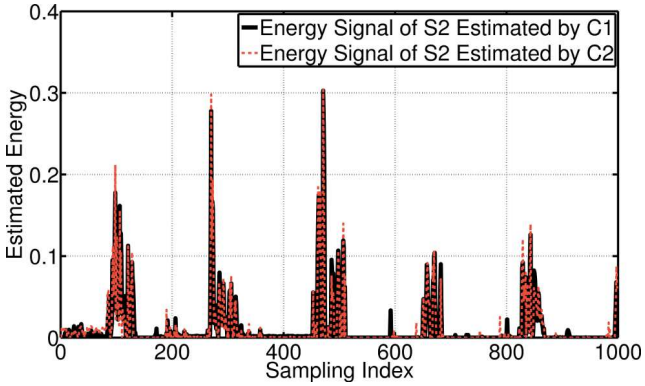
The system of (3) implies that the information is centrally available, e.g., in a fusion center, which is not the case in the considered scenario. One could think, first, to perform the decomposition at node level. However, as it is documented in [8], NNICA requires sufficient spatial diversity between the microphones, which is obviously not satisfied at a node level. We thus adopt a hierarchical approach. In particular, we consider that a network–clustering[1] algorithm, e.g., [11], is performed.[2] This formation of subnetworks is illustrated in Fig. 2a.

[1] One should not confuse the network–clustering, which divides the network in subnetworks, with the feature–clustering, which is employed for the labeling task.

[2] In the hierarchical clustering, we assume that a node of each subnetwork performs all the essential computations and sends the information to the other nodes of the subnetwork.



(a) The network of Fig. 1 is divided into subnetworks $C1, C2, C3$. The NNICA is then computed within each subnetwork for the estimation of the energy signals.



(b) Energy signals corresponding to speech source $S2$ computed by the NNICA at subnetworks $C1,\ C2$.

The model, in that case, can be recast as follows:

$$\boldsymbol{y}_k[i] \approx \boldsymbol{A}_k \boldsymbol{s}[i], \quad k = 1, \ldots, K, \tag{4}$$

where the subscript k denotes the subnetwork index and K stands for the total number of subnetworks. After this step, the NNICA is performed for $k = 1, \ldots, K$ and the nodes of the subnetwork compute the estimated energy signals of the speech sources of the network. Fig. 2b displays two energy signals estimated by subnetworks $C1$ and $C2$ via the NNICA, corresponding to the same speech source $S2$. Clearly, these signals are similar and, thus, they can potentially serve as features in the distributed clustering task, as will be discussed in Section 4.

3.2. Non-hierarchical Feature Extraction: Exploiting Similarities in the Frequency Bands which Produce Reliable Direction-of-Arrival Estimates

Non–hierarchically organized networks are able to extract features without forming subnetworks. One promising approach is to extract features based on high resolution Direction-of-Arrival (DoA) estimation. However, DoA information cannot be applied directly to labelling, since, in general, the devices in an WSN do not know their positions and array orientations. Furthermore, in the considered setup, due to the uniform linear array configuration, and the use of omnidirectional microphones, an ambiguity in the DoA estimates along the symmetry axis of the array orientation cannot be resolved.

We thus propose a novel feature which exploits the similarity across devices in the particular frequency bins that produce "good" DoA estimates for each source. The DoA is estimated with the Khatri-Rao (KR) subspace approach for locally stationary wide band signals. The idea of KR methods 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 [12]. In this way, KR methods can identify up to $N = 2J' - 2$ unknown sources in undetermined mixing systems of J' sensors. We estimate the DoA based on the spatial KR–MUSIC spectrum, which is given by:

$$P_{\mathrm{KR-MUSIC}}(\theta) = \frac{1}{\|\boldsymbol{U}_n^{\mathrm{H}} \boldsymbol{W}^{1/2} \boldsymbol{b}(\theta)\|^2}, \tag{5}$$

where the superscript $^{\mathrm{H}}$ is the hermitian operator, $\boldsymbol{U}_n$ is an estimate of the noise subspace matrix that exploits the local stationarity of the speech signals by averaging over stationary segments, and is computed as described in [12], $\boldsymbol{b}(\theta) = [e^{(J-1)\frac{j2\pi d}{\lambda}\sin(\theta)}, \ldots, 1, \ldots, e^{-(J-1)\frac{j2\pi d}{\lambda}\sin(\theta)}]^{\mathrm{T}}$ represents a dimension reduced virtual array response vector, and $\boldsymbol{W} = \mathrm{Diag}(1, 2, ..., J-1, J, J-1, ..., 2, 1)$. The N largest peaks serve as DoA estimates. Full details on KR-MUSIC are given in [12].

Fig. 2 displays the estimated DoA for $S2$ and $S3$ at devices $D1$ and $D14$. The overall DoA for each source $\hat{\theta}_n,\ n = 1, \ldots, N$, is obtained by taking the median of $\hat{\theta}_n(f)$ with $0 < f < f_s/2$. The dashed lines indicate the $\hat{\sigma}_n$-interval around $\hat{\theta}_n$. If the DoA is estimated correctly, $\hat{\theta}_n(f)$ is centered around the median. However, due to noise in particular subbands, 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 σ_n robustly [13], e.g., with the median absolute deviations scale estimator. In this way, the source specific frequency bands that typically contribute to correct DoA estimates are selected. Our proposed feature vector is formed for each source at each device by storing the frequency bin indexes within $\hat{\theta}_n \pm \hat{\sigma}$. Section 4 discusses how this feature is used in the labeling task.

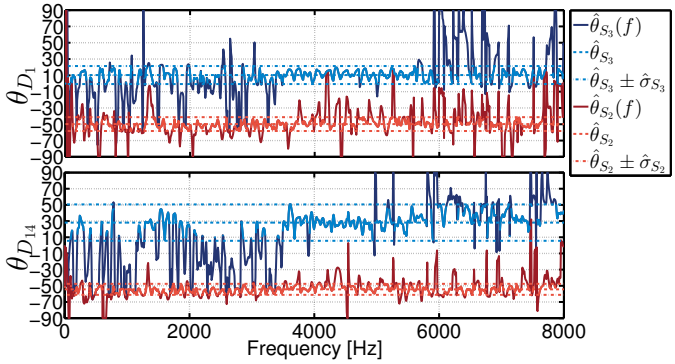


Fig. 2: The proposed non–hierarchical feature displays which frequency bins produce "good" DoA estimates for each source at different nodes. The underlying DoA estimates from which the feature is derived are displayed for $D1$ (top) and $D14$ (bottom), given $S2$ and $S3$, with positions, as depicted in Fig. 1.

4. DISTRIBUTED CLUSTERING

In this Section, we will show how the previously extracted features are used for distributed clustering. Ideally, we would like each cluster to contain every feature corresponding to the same speech source.

We achieve this by employing a cooperative clustering scheme, in the sense that, if a node cooperates with its neighbors, and these cooperate, in turn, with their neighbors, the information coming from the whole network, is incorporated. Our starting point is the distributed clustering methodology presented in [6]. We adapt this algorithm so as to fit with the current context. The steps are summarized in the sequel.

1. Initialization: The nodes (subnetworks) initialize the centroids $\boldsymbol{c}_n^{(k)}(0),\ n = 1, \ldots, N$ so that $\boldsymbol{c}_n^{(k)}(0) = \boldsymbol{c}_n^{(l)}(0)$. Methodologies for selecting the initial centroids so as to satisfy the above equality can be found in [6, Section 7.7].

2. Local Clustering Phase: Each node (subnetwork) k, at iteration i, performs a local clustering scheme by employing a K–means algorithm [14], which uses the computed features and the previously computed centroids $\boldsymbol{c}_n^{(k)}(i-1)$. 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*, whereas for the energy–based features each feature is assigned to the cluster, in which the correlation coefficient between the energy-signatures is maximized. In both cases, the local/temporal centroids $\widetilde{\boldsymbol{c}}_n^{(k)}(i)$ are computed for $n = 1, \ldots, N$.

3. Cooperation Phase Node l belonging to the neighborhood of k is activated with a certain probability (see also [7]). For simplicity, we assume that node k picks some neighbor l with probability $1/\mathcal{N}_k$, where $\mathcal{N}_k$ is the number of neighbors of k. Nodes k, l update their centroids $n = 1, 2, \ldots, N$ according to:

$$\boldsymbol{c}_n^{(m)}(i) = \frac{\widetilde{\boldsymbol{c}}_n^{(k)}(i) + \widetilde{\boldsymbol{c}}_n^{(l)}(i)}{2},\ m = k, l.$$

After the computation of the clusters, the labeling can be readily performed. The label of each speech signal will be set equal to the number of the class, in which the respective feature belongs.

Remark 1 *It is important to point out that, each of the proposed features has its own pros and cons. In particular, the energy based feature, as it will become apparent in the simulations section, 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 is slightly degraded, but they can be computed at node level and they are estimated on a single much shorter interval of only* 0.5 *seconds.*

Remark 2 *The averaging that is taking place in the cooperation phase of the algorithm drives the nodes of the network to centroid consensus; that is, the nodes compute, after a sufficient number of iterations, the same centroids. We consistently observed this behavior in extensive experiments. Furthermore, centroid consensus has been proved in [4], for a distributed K–means algorithm of similar flavor to the one employed here.*

5. SIMULATIONS

In this Section, we study the performance of the proposed distributed labeling approach. We consider the network depicted in Fig. 1 and we will validate the accuracy of the labeling, using both of the proposed features. The speech signal $S1$ corresponds to a woman making a public announcement, whereas $S2$ and $S3$ consist of two male speakers that are reading sentences in different languages. We used the mirror image method [15] 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. The proposed methodology is not compared to other techniques, since to the best of our knowledge, an algorithm suitable for distributed labeling has not been proposed, yet, in the literature.

In the first experiment, we consider that two speech sources, i.e., $S2,\ S3$, are active. We assume that both babble and white noise are present in the environment. Two nodes of the network are assumed to be connected if their distance is smaller than 4.5 meters. 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. Moreover, for the energy related feature, the network is divided in subnetworks as depicted in Fig. 2a. There, it is considered that $C2,\ C3$ can exchange information with $C1$ but they cannot communicate between each-other. The energy of the signals is computed at intervals of 30ms, corresponding to $L = 480$. The DoA based features are only computed on a single short interval of 0.5 seconds, where all sources are active. Table 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 the energy–based features, our proposed algorithm is able to label correctly the speech sources. This advantage comes at the cost of forming a hierarchical network.

Table 1: Source Labeling: Results for the Two Source Scenario

SNR	DoA related Features	Energy–Based Features
10	100%	100%
3	89%	100%
0	65%	100%

In the second experiment, we consider the more challenging scenario, where all the sources, namely $S1,\ S2,\ S3$, 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 2.

Table 2: Source Labeling: Results for the Three Source Scenario

SNR	DoA related Features	Energy–Based Features
10	80%	100%
0	60%	82%

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.

6. CONCLUSIONS

We have studied the problem of labeling speech signals in distributed environments. Our proposed methodology first derived features, related to the energy–signature of each source and the source specific frequencies which contribute to "good" DoA estimates. Then, the sources were labeled with a distributed clustering technique. Experiments showed that the proposed methodology is able to accurately label speech signals in a practical scenario.

7. REFERENCES

[1] Y. Hioka and W.B. Kleijn, "Distributed blind source separation with an application to audio signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'11)*, May 2011, pp. 233–236.

[2] 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.

[3] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 11, pp. 1663–1707, 2010.

[4] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics in Signal Process.*, vol. 5, no. 4, pp. 707–724, 2011.

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

[6] 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.

[7] 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.

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

[9] E. Oja and M. Plumbley, "Blind separation of positive sources using non-negative PCA," in *Proc. 4th Int. Symp. Indep. Comp. Anal. Blind Signal Sep. (ICA'03)*, 2003, pp. 11–16.

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

[11] M. Gerla and J. T.-C. Tsai, "Multicluster, mobile, multimedia radio network," *Wirel. Netw.*, vol. 1, no. 3, pp. 255–265, 1995.

[12] 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.

[13] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A turotial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, July 2012.

[14] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, Academic Press, 4th edition, 2009.

[15] 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.