ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 4, 331-343 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Distributed Microphone Arrays, Emerging Speech and Audio Signal Processing Platforms: A Review

Shahab Pasha*[1], Jan Lundgren[1], Christian Ritz[2], Yuexian Zou[3]

[1]*STC Research Centre, Mid-Sweden University, Sundsvall 85230, Sweden*

[2]*School of Electrical, Computer and Telecommunication Engineering, University of Wollongong, NSW 2500, Australia*

[3]*ADSPLAB/Intelligent Lab, School of ECE, Peking University, Shenzhen 518055, China*

ARTICLE INFO

Article history:
Received: 14 April, 2020
Accepted: 03 June, 2020
Online: 28 July, 2020

Keywords:
Distributed signal processing
Machine learning
Spatial selectivity
Speech enhancement

ABSTRACT

Given ubiquitous digital devices with recording capability, distributed microphone arrays are emerging recording tools for hands-free communications and spontaneous tele-conferencings. However, the analysis of signals recorded with diverse sampling rates, time delays, and qualities by distributed microphone arrays is not straightforward and entails important considerations. The crucial challenges include the unknown/changeable geometry of distributed arrays, asynchronous recording, sampling rate mismatch, and gain inconsistency. Researchers have recently proposed solutions to these problems for applications such as source localization and dereverberation, though there is less literature on real-time practical issues. This article reviews recent research on distributed signal processing techniques and applications. New applications benefitting from the wide coverage of distributed microphones are reviewed and their limitations are discussed. This survey does not cover partially or fully connected wireless acoustic sensor networks.

1. Introduction

With new portable devices, such as smartphones and tablets, conventional microphone arrays are no longer the main signal and speech acquisition platform; rather, distributed microphone arrays (also called ad hoc microphone arrays) formed by the joint analysis of randomly located independent recording devices such as laptops and cell phones are emerging recording platforms for various applications [1]. Conventional compact microphone arrays and recording devices are now processed within ad hoc arrays and are jointly analyzed with other such devices.

For this reason, distributed microphones are popular and have been used in a wide range of applications, such as speaker tracking and speech recognition systems [2]. Currently, there is substantial potential for applications that use digital recording devices collaboratively as a virtual array [3]. By positioning such devices at random locations within the acoustic scene, the array geometry is no longer limited to standard structures. Distributed microphone nodes provide wide spatial coverage and are ideal for capturing multiple speakers located meters away from one another. This advantage in the means of signal (i.e., sound and other cues derived from it) acquisition allows different audio scenes to be captured to give a more accurate picture of the user environment. To enable such ubiquitous and flexible teleconferencing and multimedia applications within the distributed signal processing context, several important technical and theoretical problems should be addressed. Some of the main challenges are the unknown/changeable array structure, inconsistent sampling frequencies, varying gains (due to varying source-to-microphone distances), and unsynchronized recordings. Although speech and signal processing applications are well studied and straightforward [4] in the context of known-geometry compact microphone arrays, existing methods are not directly applicable to distributed scenarios.

This survey article reviews the strengths and limitations of recently proposed distributed signal processing methods. The

*Corresponding Author: Shahab Pasha, Email: shahab.pasha@miun.se

https://dx.doi.org/10.25046/aj050439

focus is on scenarios in which the microphones are independent and do not communicate within the array.

2. Literature Selection Methodology

The authors have reviewed the most recent theoretical and practical literature on distributed signal processing and related, overlapping fields of study. More than 250 research papers, theses, patents, and online tools and resources were studied and compared over a five-year period. One hundred peer-reviewed published works were chosen for final comparison and investigation according to the novelty of the work and the credibility of the journal/conference. The current set of references covers the literature and online resources as of 2020, though most of the main methods were proposed between 2013 and 2017.

3. Distributed Signal Processing

Ad hoc microphone arrays [1, 3] consist of a set of recording devices (referred to as nodes [5]) randomly distributed in an acoustic environment to record an unknown acoustic scene with wide spatial coverage (Figure 1). The nodes can be identical [5] or different [6] in terms of their structure and number of elements [7, 8] (Table 1). Ad hoc arrays eliminate the restrictions on microphone and source placement at fixed locations and facilitate dynamic and flexible recording experiences.

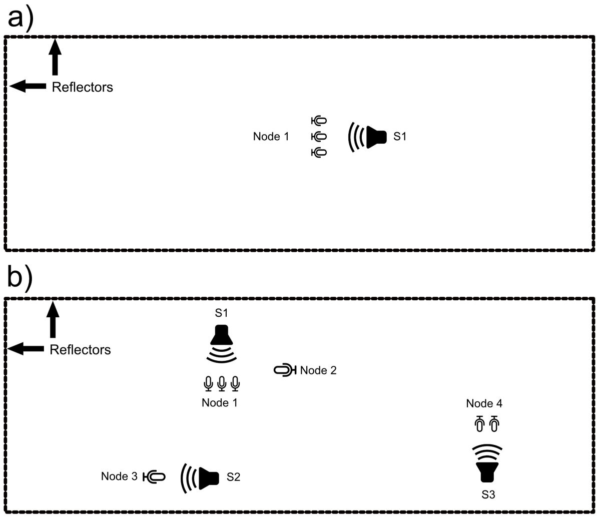


Figure 1: a) A three-element linear microphone array recording a single source; b) a distributed acoustic scene comprising three sound sources, two single-channel nodes, and two multi-channel linear arrays.

Table 1: Ad hoc microphone arrays compared with compact arrays

Distributed	***Compact***
Unknown structure [9]	Known structure
Changeable microphone locations [10]	Fixed topology
Unknown inter-channel time delays [11]	Known inter-channel time delays
Inconsistent gain within the array [12]	Consistent gain
Uncertain direction of arrival (DOA) definition [13]	Straightforward DOA definition
Large phase differences (i.e., spatial aliasing)	Negligible spatial differences
Inconsistent signal quality [5]	Consistent signal quality

3.1. Definition

M recording devices (both single- and multi-channel devices) located randomly to capture N sound sources form an ad hoc microphone array. Node m contains M_m channels and the signal picked up by the ith channel forming the mth node is modelled as

$$x_{m,i}(n) = \sum_j \sum_i s_j(n) * h_{m,i,j}(n) + v_{m,i}(n) + w_{m,i}(n), \tag{1}$$

where $s_j(n)$ is the jth sound source, $h_{m,i,j}(n)$ represents the room acoustic response between the ith channel forming the mth node and source j, and $v_{m,i}(n)$ is the additive noise; $w_{m,i}(n)$ is the non-coherent component and represents reverberation and diffuse noise. Each channel location is modelled as

$$\mathbf{r}_{m,i} = [x_{m,i}, y_{m,i}, z_{m,i}], \tag{2}$$

and the jth source location is

$$\mathbf{r}_{s,j} = [x_j, y_j, z_j]. \tag{3}$$

in the Cartesian coordinate system. $h_{m,i}(n)$, $\mathbf{r}_{m,i}$, and $\mathbf{r}_{s,j}$ are assumed unknowns for all the values of m, i, j, and n. The total number of channels in the array is

$$M_{ch} = \sum_{m=1}^{M} M_m. \tag{4}$$

For ad hoc scenarios where all the nodes are single channel

$$M_{ch} = M. \tag{5}$$

The truncated RIRs ($\mathbf{h}_{m,i,j}$) of length L are modelled as follows:

$$h_m(n) = \sum_{k=0}^{L} a_{m,k}\delta(n - \tau_{m,k}), \tag{6}$$

with time delays, $\tau_{m,k}$, and amplitudes, $a_{m,k}$. L is chosen based on the application and the reverberation time [14]. $\tau_{m,0}$ represents the time of arrival (TOA) at node m.

Assuming that the distances between the channels at each node are relatively small and that each node forms a compact microphone array [5, 7], each node can deliver a single-channel output, so there will be only M distributed recordings [15]. Having only one active source during the short frames simplifies (1) to

$$x_m(n) = s(n) * h_m(n) + v_m(n) + w_m(n). \tag{7}$$

3.2. Significance and applications

Tablets, smartphones, sound recorders, and other portable and wearable digital devices are becoming prevalent in workplaces, homes, and lecture halls, redefining how we communicate and record our communications. Consequently, these devices are becoming key tools for daily activities, including teleconferencing and hands-free speech communication [16]; ad hoc signal processing is therefore inevitable.

Advances in ad hoc signal processing technologies can also relax the highly demanding constraints on network layer design [17]. Ad hoc arrays also improve the quality of speech communication, acoustic scene analysis, and speech recognition (Figure 2) due to the multiple observations they make within a wider area [18].

3.2.1. Improved Distributed Meetings

Meetings are an important part of everyday life for many types of educational and professional workgroups. The main application of ad hoc microphone arrays is for distributed meetings (DMs) [19], as such meetings are recorded by microphones at unknown and varying environments with different signal qualities. DM systems enable the high-quality broadcasting and recording of meetings by utilizing the flexible spatial coverage of ad hoc arrays. Distributed meetings using signal processing techniques are not necessarily online meetings, but could be group meetings [20]. Ad hoc arrays are more suitable recording tools for distributed meetings than are compact microphone arrays, as the recording devices can be spread out in the meeting room. Joint analysis of the distributed microphone signals yields more accurate results for signal processing applications such as active source detection (i.e., localization) [21].

3.2.2. Hearing Aids

Reduction of interference and noise is important in hearing aids (HAs) to provide intelligible speech signals in noisy environments [22]. Using an array of microphones, it is possible to exploit the spatial characteristics of the acoustic scenario to obtain more information about the target scene. Although the scenario investigated by Bertrand and Moonen [22] was a fully connected binaural network, the unknown geometry of the array and the random recording setup (i.e., the source locations) form an ad hoc acoustic scene.

3.2.3. Hands-Free Communication

Wearable recording nodes are essential to hands-free communication systems, and the node structure can also vary. As the node and source locations in such systems are unknown and changeable, classic array-processing methods cannot be applied. Ad hoc signal processing offers a solution for the joint analysis of the nodes in hands-free voice communication systems for improved speech enhancement and source localization applications. The general scenario of hands-free audio recording and voice recognition with distributed nodes has been described and investigated by Jia et al. [23].

3.2.4. Ambient Intelligence and Smart Homes

Ambient intelligence advances are based on advances in sensors, pervasive computing, and artificial intelligence [24]. Very little can be done by an ambient intelligent system without detecting the user's presence and movements. The ambient intelligence system must be aware of the users' locations in each time period. One way to do this is by tracking the sound sources and acoustic cues. Spatially distributed microphone arrays are effective tools for monitoring user movements [25]. The insights gained provide important clues as to the type of activities the user is engaged in, and make the system responsive to the user's

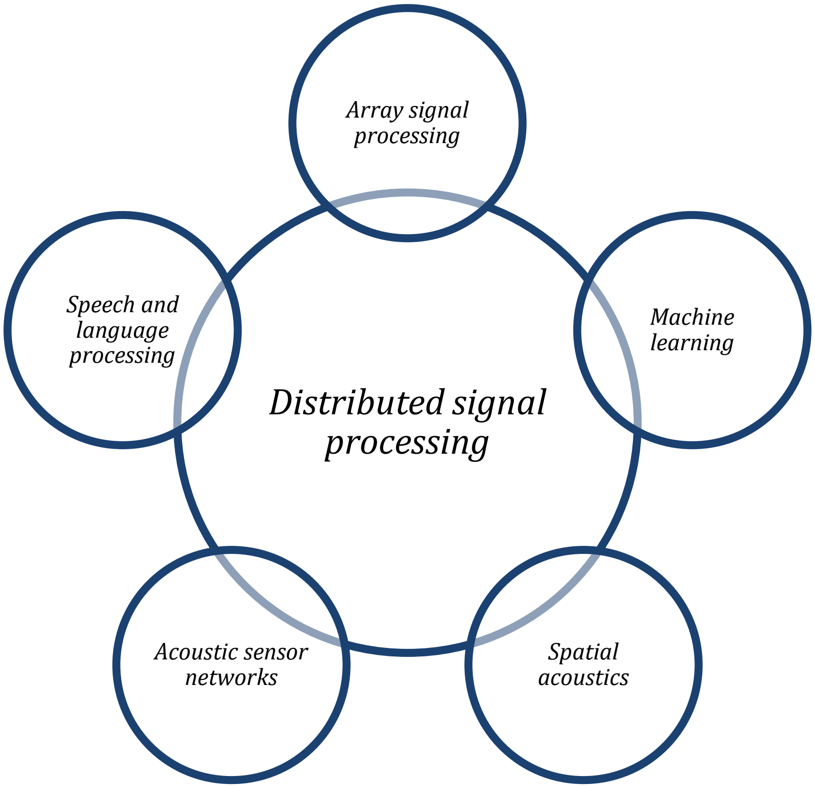


Figure 2: Distributed signal processing and overlapping fields of study

location [26]. Speaker verification and identification by ad hoc microphone arrays and speech recognition for automated human–computer interaction [18] are important fields of research that can help customize ambient intelligence applications.

3.2.5. Monitoring

Some recent research focuses on analyzing the environment by means of acoustic sensing due to its unobtrusive nature. Acoustic monitoring covers a wide range of applications, from security tasks (e.g., intrusion detection, malicious activity early detection, and traffic monitoring [27]) to whale migration tracking [28] and bird behavior studies [29].

3.2.6. Medical Signal Processing

Unconventional distributed microphone arrays have been used in medical applications for accurate and high-quality chest sound pick-up [30] and vital sound separation [31]. Compared with conventional compact arrays, ad hoc microphone arrays provide more accurate recordings of closely located organs such as the lungs and heart. The design and development of ad hoc acoustic sensors and microphones are essential for medical and E-health research.

3.3. Room Acoustics and Distributed recording

Unlike conventional compact microphone arrays, in which the noise and reverberation levels are consistent, in distributed signal processing each microphone has its own unique reverberation level and room acoustic response. Researchers have recently shown that the unique RIR and echo pattern at each ad hoc microphone location contain location and distance information even if the recording setup is unknown [32, 33]. This idea is applicable to ad hoc microphone arrays for microphone clustering [11], room geometry reconstruction [33], and microphone localization [34].

The framework for recording using distributed microphone nodes was formulated by Tavakoli et al. [16] (1), covering both reverberation and noise. The signals recorded by all M_{ch} channels (4) within an ad hoc array are modeled as

$$\mathbf{x}(n) = [x_1(n), \dots, x_M(n)]^T. \tag{8}$$

As each ad hoc microphone receives its own unique distorted version of the source signal, the RIR and reverberation at each microphone location contain important information [11, 35]. Each RIR ($h_m(n)$) (6) can be represented by sets of reverberation times and reverberation amplitudes [11], as follows:

$$[a_{m,0}, \dots, a_{m,L}], \tag{9}$$

where a_0 is the direct path impulse amplitude, and

$$[\tau_{m,0}, \dots, \tau_{m,L}]. \tag{10}$$

Researchers have divided the RIR (1) into two segments: early and late echoes [36]. This segmentation is specifically important in the context of ad hoc microphone arrays and is the basis for defining discriminative features. The clarity feature is used for sound source localization by ad hoc microphones [37]. Table 2 summarizes the acoustic features applied to signal processing applications.

Table 2: Room acoustics features applied to distributed signal processing

Method/features	***Application***	***Reference(s)***	***Year***
Energy	Localization	Liu et al. [12]	2007
Noise coherence	Clustered beamforming	Himawan et al. [49]	2011
Voice activity detection	Speech enhancement	Sakanashi et al. [50]	2013
Sparsity analysis	Localization	Asaei et al. [21]	2014
Time delay and sound level	Speech separation	Souden et al. [7]	2014
Signal power	Traffic monitoring	Toyoda et al. [27]	2014
Euclidian distance matrix (EDM)	Microphone localization	Dokmanic et al. [51]	2015
Kurtosis of the linear prediction residuals	Clustering	Pasha et al. [8]	2015
Generalized cross-correlation	Speaker tracking	Tian et al. [52]	2015
Reverberation and echoes	Clustering	Pasha et al. [11]	2015
Non-negative matrix factorization (NMF)	Calibration	Asaei et al. [53]	2015
Pseudo coherence vector	Node selection for speech enhancement	Tavakoli et al. [5]	2015
Reverberation	DOA estimation	Pasha et al. [54]	2015
Cepstral features	Interference suppression	Gregen et al. [9]	2016
C_{50} (short-time reverberation)	Multi-talk detection	Pasha et al. [37]	2016
Magnitude-squared coherence (MSC)	Crosstalk and multi-talk detection	Pasha et al. [37]	2016
Echoes	Room geometry reconstruction	Dokmanic et al. [33, 55]	2016
Coherent-to-diffuse ratio	Multi-talk and crosstalk detection	Pasha et al. [56]	2017
Power spectral density (PSD)	Spotforming	Habet et al. [57]	2017
Signal to interference and noise ratio	Noise cancellation	Tavakoli et al. [20]	2017
Time difference	Microphone localization	Woźniak et al. [58]	2019
Distributed unscented Kalman particle filter (DUKPF)	Source localization	Zhang et al. [59]	2020

3.4. Challenges and Limitations

Large scale distributed arrays are inherently asynchronous [38]. Inconsistent sampling rates [39], gain differences [12], and different signal-to-noise ratios (SNRs) at different locations [5, 12] are challenges with distributed signal processing. The main differences between ad hoc array and compact array signal processing are summarized in Table 1.

4. Distributed Signal Processing

The most recent existing distributed signal processing techniques proposed in the literature are reviewed here in terms of their target scenarios and requirements. Distributed signal processing overlaps with array processing [4, 40], wireless sensor networks [41] (not reviewed here), feature extraction and machine learning [42], hands-free speech communication [43], and guided/informed signal and speech processing [44, 45].

4.1. Microphone Calibration

Microphone calibration is especially important in the context of large arrays [46, 47] and distributed microphones, as the area covered by the microphones can be large [48]. These large distances and attendant time delays should be considered when time aligning the signals for beamforming and speech enhancement applications. Representing the gain of the array by $\boldsymbol{g} = \{g_1, \dots, g_M\}$, (1) can be rewritten as

$$x_m(n) = s(n) * g_m \bar{h}_m(n) + v_m(n), \qquad (11)$$

for one active source, where g_m and $\bar{h}_{m,i}(n)$ are the gain and normalized RIR of microphone m, respectively. Calibration often consists of estimating the distances between the pairs of microphones and reconstructing the array geometry given all the pairwise distances [53, 60]. Microphone array calibration in general suffers from reverberation, noise, and complicated mathematical computations (making calibration infeasible for real-time applications); specifically, sampling frequency mismatch, inconsistent microphone gain, and non-stationary array geometry make calibration an ongoing process throughout the recording session. Due to calibration difficulties, some methods prefer to avoid microphone calibration altogether, if possible applying methods inherently robust to microphone placement and steering error [61].

For microphone array calibration, some advanced mathematical methods use joint source and microphone localization methods [62] and incorporate matrix completion constrained by Euclidean space properties [63, 64]. Such methods require partial knowledge of pairwise microphone distances [65]. It has been shown that using sound emissions for self-calibration can result in a calibration method more robust to sampling frequency mismatch [66]; however, the method is only applicable to devices that have both recording and sound emitting capabilities.

4.2. Signal Synchronization

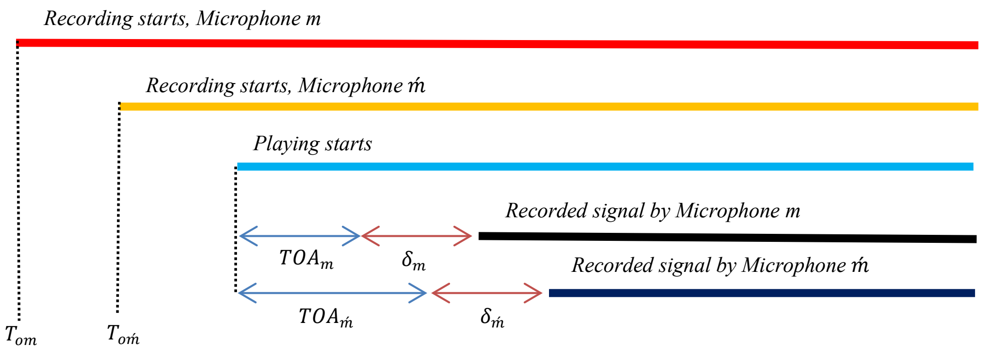
Signal synchronization is an essential task for more advanced applications such as dereverberation and speech enhancement. Improperly addressing the synchronization issue leads to poor dereverberation performance. The different types of delays, including microphone internal, TOA, and onset time delays (Figure 3), have been investigated and explained [67, 68]. The lack of a reference channel and inconsistent sampling frequencies [69] in ad hoc arrays lead to critical phase differences [70] between the channels, fundamentally challenging most speech and audio processing applications. Existing solutions to clock drift [71] require limiting assumptions, such as unmoving sources and stationary amplitudes.

If the array topology is available, the time difference of arrival (TDOA) (13) between any two microphones in the array can be calculated as

$$TOA_{s,m} = \frac{|\mathbf{r}_s - \mathbf{r}_m|}{c} + \delta_m \qquad (12)$$

$$TDOA_{m,\acute{m}} = \frac{|\mathbf{r}_s - \mathbf{r}_m|}{c} - \frac{|\mathbf{r}_s - \mathbf{r}_{\acute{m}}|}{c} + (\delta_m - \delta_{\acute{m}}) + (T_{om} - T_{o\acute{m}}), \qquad (13)$$

where δ_m and T_{om} represent the internal delay and onset time of microphone m, respectively (Figure 3) [67, 68], and $\mathbf{r}_s =$



Recording starts, Microphone m
Recording starts, Microphone ḿ
Playing starts
Recorded signal by Microphone m
TOA_m
δ_m
Recorded signal by Microphone ḿ
$TOA_{\acute{m}}$
$\delta_{\acute{m}}$
T_{om}
$T_{o\acute{m}}$

Figure 3: Time delays

$[x_s, y_s\, z_s]^T$, $\mathbf{r}_m = [x_m, y_m\, z_m]^T$, and $\mathbf{r}_{\acute{m}} = [x_{\acute{m}}, y_{\acute{m}}\, z_{\acute{m}}]^T$ are the source, microphone m, and microphone $\acute{m}$ locations in Cartesian coordinate space, respectively. The objective of synchronization is estimating the overall time delay between every pair of microphones. The time delays are given as

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \cdots & \tau_{1M} \\ \vdots & \ddots & \vdots \\ \tau_{M1} & \cdots & \tau_{MM} \end{bmatrix}, \tag{14}$$

where $\tau_{mm} = 0$ for m = 1 to M and $\tau_{m\acute{m}} = \tau_{\acute{m}m}$ for all m and $\acute{m}$ values.

Researchers have used the time alignment of ad hoc channels for source localization by means of generalized cross correlation (GCC) [17] and for defining the parametric squared errors of time differences [72].

Some advanced techniques use the least squares for the temporal offset estimation [73] and for audio fingerprinting [38]. These methods are developed based on the clustering and synchronizing methods applied to unorganized multi-camera videos [74] which are based on matching the time-frequency landmarks between two channels.

The effect of synchronization on blind source separation (BSS) by a wireless acoustic sensor networks (WSAN) was investigated by Lienhart et al. [75], who concluded that full synchronization increases the BSS cost function by an average of 4 dB.

Most of the proposed synchronization methods can time-align the signals accurately and calculate the TDOA with an error of 1–10 milliseconds. The other important factor is the computational cost. The watermark-based algorithms [38] are shown to be more efficient than GCC methods [74].

4.3. Spatial Multi-Channel Linear Prediction (LP)

Multi-channel LP was developed for compact arrays and has applications, such as dereverberation [76] and compression [77, 78].

Using x_m from (1), the autocorrelation $r_m(c)$ is obtained for channel m from

$$r_m(c) = E\big(x_m(n)x_m(n+c)\big),\ \ c = 0,1,2,\ldots \tag{15}$$

where E is the mathematical expectation.

$$\bar{r}(c) = \frac{1}{M} \times \sum_{m=1}^{M} r_m(c). \tag{16}$$

The baseline autocorrelation function (16), can be formulated in the more general form of a weighted average autocorrelation ($\bar{r}_w(c)$) (17). Assuming that the applied weights are $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$, the weighted average autocorrelation function is calculated as

$$\bar{r}_w(c) = \frac{1}{\sum_{m=1}^{M} \beta_m} \times \sum_{m=1}^{M} \beta_m r_m(c), \tag{17}$$

where β_m is the weights given to $r_m(c)$.

$$\mathbf{w}_s = \begin{bmatrix} \bar{r}_w(0) & \cdots & \bar{r}_w(P_{short}-1) \\ \vdots & \ddots & \vdots \\ \bar{r}_w(P_{short}-1) & \cdots & \bar{r}_w(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_w(1) \\ \vdots \\ \bar{r}_w(P_{short}) \end{bmatrix}, \tag{18}$$

and the pre-whitened signal is

$$\tilde{e}_m(n) = x_m(n) - \sum_{k=1}^{P_{short}} w_{s,k} x_m(n-k), \tag{19}$$

where $w_s = \{w_{s,1}, \ldots, w_{s,P_{short}}\}$ [13]. Having the source-to-microphone distances $\{q_{1,s}, \ldots, q_{M,s}\}$, the ideal distance weights are $\mathbf{q} = \{\frac{1}{q_{1,s}}, \ldots, \frac{1}{q_{M,s}}\}$. It is observed that using $\mathbf{q}$ as the weights significantly improves the autocorrelation function estimation in (18).

4.4. Beamforming

Beamforming, i.e., the process of focusing on a specific signal (based on the DOA or other characteristics), is widely used as part of multi-channel speech enhancement methods [6, 79]. Three beamforming techniques (listed below) have been applied to ad hoc microphone arrays. Generally, delay and sum beamforming (DSB) [8, 49] is more flexible and does not require limiting requirements, whereas more advanced beamforming methods assume some prior knowledge, which might not be the case for general scenarios.

4.4.1. Delay and Sum Beamforming (DSB)

This beamforming technique has been successfully applied to ad hoc microphone arrays [49]. Using $\boldsymbol{y}(n)$ (8), the DSB output is calculated as

$$\bar{x}_{DSB} = \sum_{m=1}^{M} x_m\left(n - \tau_{m,ref}\right), \tag{20}$$

where $\tau_{i,ref}$ (14) is the time delay between channel i and the reference channel. The beamformer filter coefficients are obtained by:

$$\widehat{W} = \arg\min_{w} \mathbf{w}^H \varphi_x \mathbf{w}, \tag{21}$$

where $\varphi_x = E\{xx^H\}$ is the covariance. The solution is

$$\widehat{\mathbf{W}} = \frac{\varphi_x{}^{-1}\mathbf{h}}{\mathbf{h}^H \varphi_x{}^{-1}\mathbf{h}}. \tag{22}$$

4.4.2. Minimum Variance Distortionless Response beamforming (MVDR)

An optimization method for the MVDR beamformer using the pseudo-coherence model of the array (24), based on the coherence function (20), is proposed and successfully tested by Tavakoli et al. [5]:

$$P_{X_m X_{ref}} = \frac{E\left[\mathrm{X}_m X_{ref}{}^{*}\right]}{E\left[\left|X_{ref}\right|^2\right]}, \tag{23}$$

$$X_m = \sum_{p=1}^{P} P_{x_m^p X_{ref}} X_{ref}^p, \tag{24}$$

where X_m is the frequency domain signal of $\mathbf{x_m}$ (1) and * is the complex conjugate. The MVDR beamformer requires knowledge of the source DOA and of the steering vector of the array when applied to compact microphone arrays; however, under certain assumptions, it is possible to modify the MVDR beamformer and apply it to the distributed scenarios. The assumptions include connection between the channels [80] and an ad hoc array with nodes of known geometry [5].

4.4.3. Linearly Constrained Minimum Variance (LCMV)

The LCMV was applied to distributed scenarios by Wood et al. [81] in experimental setups covering a wide range of random meeting scenarios. However, the LCMV beamformer requires knowledge of the RIR at each microphone location for each source in the meeting room. Himawan [82] proposed and successfully tested a clustered approach to blind beamforming for ad hoc microphone arrays, the applied features being the coherence between the diffuse noise and TDOA. The fact that the noise coherence between two microphones depends on the inter-microphone distance is exploited to estimate how close two microphones are. It is also well known that microphones located near each other have lower TDOAs, whereas microphones located farther away (i.e., metres) from each other have larger TDOAs.

4.5. Speech Enhancement

Speech enhancement can cover applications, such as noise cancellation [22, 83], beamforming [5], and echo cancellation [8]. These applications can be used separately or jointly as a combined speech enhancement method. The state-of-the-art speech enhancement techniques proposed for conventional arrays of known geometry are inapplicable to ad hoc microphone arrays, and existing approaches are confined to basic beamforming techniques [5, 49]. Some basic techniques apply centralized multi-channel Wiener filters and the so-called distributed adaptive node-specific signal estimation (DANSE) algorithm [22] to remove noise in distributed hearing aid systems. These methods assume that the channels can communicate and transmit time stamps. Other speech enhancement methods proposed for ad hoc arrays require limiting supervision requirements, such as user identification of the target speech [50].

A general scenario with M ad hoc microphones is rewritten as

$$\mathbf{x}(n) = \mathbf{h}(n) * s(n) + \mathbf{v}(n), \tag{25}$$

$$\mathbf{h}(n) = \begin{bmatrix} h_1(n) \\ \vdots \\ h_M(n) \end{bmatrix}, \tag{26}$$

where $\mathbf{x}(n) = [x_1(n), \dots, x_M(n)]^T$ (8) contains the multi-channel recording of all M microphones in the array, $\mathbf{h}(n)$ is the RIR matrix at each microphone's location for source j, and $\mathbf{v}(n)$ is the diffuse noise. The goal is to retrieve $s(n)$ from $\mathbf{x}(n)$(25). A clustered speech enhancement approach based on beamforming and auto-regressive (AR) modelling of the speech signal was proposed and tested by Pasha and Ritz [8]. They showed that removing the microphone located far from the active source and exclusively applying the multi-channel dereverberation method for the microphones located nearer the source improved the speech enhancement performance.

$$e_{x_m}(n) = x_m(n) - \sum_{k=1}^{p} b_k x_m(n-k). \tag{27}$$

where $x_m(n)$ (7) is the single-channel recording of an ad hoc microphone array and b_k is the LPC coefficient of order P. The kurtosis of the LP residual signals is then calculated as

$$k_{x_m}(n) = \frac{E\{e_{x_m}^4(n)\}}{E^2\{e_{x_m}^2(n)\}} - 3 \tag{28}$$

where $E\{\}$ denotes the mathematical expectation and $k_{x_m}(n)$ is the kurtosis of the LP residual signals.

4.6. Source Localization and DOA Estimation

DOA estimation refers to one-dimensional source localization (i.e., angle of arrival) [84], whereas source localization can have a more general meaning, such as pinpointing the source location in a room (i.e., two- or three-dimensional localization) [85]. In this section, DOA estimation and source localization are reviewed together as they both require location feature estimation [21].

Source localization methods proposed for microphone arrays are based on extracting location features from the recorded signals and analyzing them to localize the active source [7]. This approach has limiting assumptions when applied to ad hoc arrays; for instance, if the relative distance between the microphones is unknown, extracted features such as amplitude attenuation and time delays cannot be accurately translated to location features. Researchers have tried to address this issue for the gain feature by proposing a relative attenuation feature [12, 62] that can localize collocated sources and microphones when the microphones have different gains. It has also been shown that if some sources (i.e., three out of seven) are not collocated with any microphones, the method can still localize all the sources and microphones accurately.

It has been shown that arbitrarily arranged sensors (forming a network of acoustic sensors) can be effectively applied for source DOA estimation [13]. The results indicate that the proposed method detects the source angle of arrival in degrees with less than 2% error.

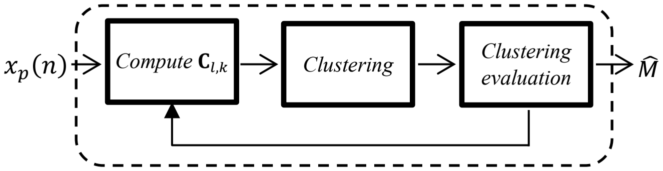


Figure 4: The proposed source counting system

Features derived from RIRs, such attenuation [54], and the clarity feature (C_{50}) [37], are also applied for the two-dimensional localization of sources.

4.7. Source Counting and crosstalk detection

Speech processing methods use voice activity detection (VAD) to detect the periods with an active speaker. In scenarios with more than one microphone, VAD can be applied in source counting (Figure 4) and multi-talk detection applications as well [9].

Inspired by VAD algorithms, researchers have proposed multi-talk detectors in which the source and microphone locations are not available (Figure 4). Moonen and Bertrand [86] suggested a multi-speaker voice activity detection method that tracks the power of multiple simultaneous speakers. Coherent-to-diffuse ratio (CDR) values (32) calculated or estimated at dual microphone node locations are also applied for source counting [56].

For two-element nodes where $\tilde{x}_{n,p}(t)$ (29) represents the signals recorded by the two channels, $p \in \{1, 2\}$, at node $n \in \{1, \dots, N\}$, $\tilde{s}(t)$ is the source signal and $\tilde{h}_{n,p}(t)$ is the RIR at the nth node as in (6). The CDR features are calculated using the following:

$$\tilde{x}_{n,p}(t) = \tilde{s}(t) * \tilde{h}_{n,p}(t) + u_{n,p}(t), \tag{29}$$

$$C_{v_n}(f) = \frac{\left|\varphi_{v_{n,1}|v_{n,2}}(f)\right|^2}{\varphi_{v_{n,1}|v_{n,1}}(f)\,\varphi_{v_{n,2}|v_{n,2}}(f)}, \tag{30}$$

$$C_{x_n}(l,f) = \frac{\left|\varphi_{x_{n,1}|x_{n,2}}(l,f)\right|^2}{\varphi_{x_{n,1}|x_{n,1}}(l,f)\,\varphi_{x_{n,2}|x_{n,2}}(l,f)}, \tag{31}$$

where $C_{v_n}(f)$ and $C_{x_n}(f)$ are the noise and the signal magnitude squared coherence (MSC) values respectively and l represents the frame index. The CDR is

$$CDR_n(l,f) = \frac{C_{u_n}(f) - C_{x_n}(l,f)}{C_{x_n}(l,f) - C_s(l,f)}, \tag{32}$$

from which the use of the average CDR over the entire frequency band and L frames is given by

$$\overline{CDR}_n = \frac{1}{L(f_B - f_0)} \int_{f=f_0}^{f_B} \sum_{l=1}^{L} CDR_n(l,f)\ \mathrm{d}f. \tag{33}$$

The main limitation of the method proposed by Pasha et al. [56] is that all the nodes must be of the same structure, which limits the method's applicability. The MSC is found using the cross-power spectral density (CPSD) as presented by Pasha et al. [87] (Figure 4):

$$c(l,f) = \frac{\left|\varphi_{x_1|x_2}(l,f)\right|^2}{\varphi_{x_1|x_1}(l,f)\,\varphi_{x_2|x_2}(l,f)}, \tag{34}$$

where $f \in \{1, \dots, F\}$ is the frequency index of F total frequencies. The CPSD function used in (34) is defined as

$$\varphi_{x_1|x_2}(l,f) \triangleq \frac{1}{AK} \sum_{a,b} (x_1 \star x_2)(Al + a, b) e^{\frac{-j2\pi kb}{K}}, \tag{35}$$

where $a \in \{1, \dots, A\}$ is a frame index and $j = \sqrt{-1}$ represents the imaginary unit. The cross-correlation is calculated by:

$$(x_1 \star x_2)(\cdot, b) \triangleq \sum_n x_1(\cdot, n) x_2(\cdot, n + b), \tag{36}$$

where b is the displacement and $x_p(n)$ framed is $x_p(Al + a, n)$.

4.8. Source Separation

Crosstalk and speaker overlap decrease the signal quality and intelligibility in scenarios such as teleconferencing and meetings [88, 89].

The problem is mathematically formulated for M microphones and N sources (1) as

$$y_m(n) = \sum_j s_j(n) * h_{m,i,j}(n) \tag{37}$$

where $y_m(n)$ is the speech mixture recorded by the mth microphone from the array, $s_i(n)$ is the jth-source speech signal, and $h_{m,i,j}(n)$ is the RIR for the mth microphone and the jth source (1). The goal is to obtain $\mathbf{s}(n) = \{s_1(n), .., s_N(n)\}$ from $\mathbf{y}(n) = \{y_1(n), .., y_M(n)\}$.

Independent component analysis (ICA) was applied to blind speech separation in the online teleconferencing applications by Dmochowsky et al. [90]. Although the proposed ICA method is formulated for a general scenario, the experimental setup does not cover various scenarios and is limited to a two-element array. The noise and reverberation levels are low in the experimental setup and challenging scenarios (e.g. reverberation times higher than 800ms and very low SNRs) are not investigated.

More advanced sound source separation methods take into account the spatial coverage of the distributed arrays [91]. The signals obtained by the sub-arrays are then filtered by a geometric filter to achieve the highest output SIR. It is concluded that the proposed method can suppress (reject) interference by up to 40 dB in a reverberant environment. The novelty of this method is its use of passing and rejecting masks in the time-frequency domain to partition the microphones based on their power spectral density (PSD). The experimental setup is confined to one scenario with three sources and three microphones located near each other in pairs.

4.9. Speech Recognition

Speech recognition as a main aspect of human–machine interaction has attracted significant attention in recent years. Distributed microphone arrays have significant advantages over compact microphone arrays as they provide unobtrusive and spatially flexible interaction between humans and personal devices spread out within a room. A series of ad hoc signal processing techniques for speech recognition applications, such as spatial directivity, beamforming, and speech feature extraction, was discussed by Himawan [18], the main focus being on beamforming for speech enhancement. Generalized side lobe cancelling techniques [92] and linear prediction (LP)-based speech enhancement [79] have proven to be successful speech recognition methods in ad hoc scenarios.

5. Machine Learning Applied to Distributed Scenarios

Machine learning techniques have been widely used in different areas of speech and audio signal processing, such as emotion recognition and source localization [93]. Machine learning and data mining techniques have been shown to be effective for learning and predicting nonlinear patterns. They have been widely used for beamforming via support vector machines (SVMs), source localization via neural networks, and other applications. As machine learning techniques are highly sensitive to the training set and parameters, using them in the flexible and uncertain distributed scenarios is very challenging. However, researchers have managed to define informative discriminative features for clustering and classifying microphones and signals, features that are independent of any specific setup [8, 91] and can discriminate among the microphones within an ad hoc array regardless of array topology. These methods flexibly exclude a subset of the microphones (nodes) from the multi-channel process (e.g., multi-channel speech enhancement [8]) and are based on certain predefined selection criteria [5].

5.1. Microphone Clustering

Clustering is an unsupervised machine learning technique the goal of which is to assign objects (e.g., microphones) to groups with small intra-group differences and large inter-group differences [94]. The problem of clustering microphones based on their spatial locations was investigated by Gregen et al. [95]. It is important to cluster the microphones based on their spatial distances to select an optimal subset of microphones. Clustered speech processing approaches are applied to take advantage of the spatial selectivity of beamforming [49], dereverberation [8], and interference suppression [9]. The mel-frequency cepstral coefficients (MFCCs) [95], the coherence feature [3], the Legendre polynomial-based cepstral modulation ratio regression (LP-CMRARE) [95] (38)–(39), and the kurtosis of the LP residual [8] are used as features in various clustering methods reported in the literature. Along with the discriminative clustering feature, the clustering technique is important as it determines the limitations of the overall clustering method. For instance, the fuzzy clustering method presented by Gregen et al. [95] requires prior knowledge of the number of sources, whereas a flexible codebook-based clustering method based on RIRs was applied by Pasha et al. [11]:

$$\hat{X}_C = |\mathrm{F}(X_c)| \tag{38}$$

where X_c is the recorded signal in the cepstral domain and F denotes the Fourier transform. The average magnitude over the window length (C_T) of the modulation spectrum is calculated using

$$\bar{X}_c = \frac{1}{C_T}\sum_{c=0}^{C_T-1}|\hat{X}_C|, \tag{39}$$

as the clustering feature. A practical clustering method in which the numbers of microphones and clusters are not known was proposed by Pasha and Ritz [8], who applied clustering to exclude highly distorted and reverberant microphones from the dereverberation process. It is concluded that the clustered approach improves the direct-to-reverberant ratio (DRR) by 5 dB. Although it is impossible to evaluate the clustering methods decisively [95] when there is no ground truth, researchers have proposed clustering performance measures such as the success rate (SR) [11] and purity [49].

5.2. Signal Classification

Supervised machine learning methods such as classification are highly sensitive to the training set and setup, and are not always applicable to uncertain changeable scenarios such as ad hoc microphone arrays and meetings [96]. However, it has been shown that under certain assumptions (e.g., the availability of a clean training set) [95], it is possible to classify the recorded signals based on certain predefined classes (e.g., speech, noise, and music). It is concluded that cepstral features such as MFCC and LP-CMRARE are reliable features with which to discriminate speech, music, and noise signals [95].The microphone clustering method uses MFCC and LP_CMRARE as the features and divides microphones into two (i.e., the number of sources, which is assumed to be known) clusters. The signals recorded by the clustered microphones are then classified based on the predefined classes. It is also assumed that clean training data for each class are available [97].

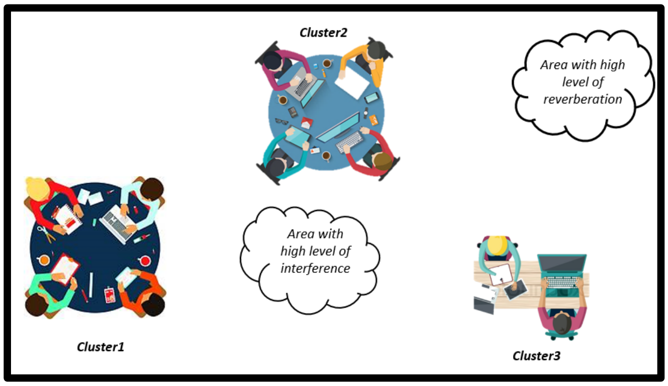

Cluster2
Area with high level of reverberation
Area with high level of interference
Cluster1
Cluster3

Figure 5: Microphone clusters [3]

6. Distributed Signal Processing Resources

An extensive database recorded using ad hoc microphones was reported and applied by Wood et al. [81] for real-world beamforming experiments. Twenty-four microphones were positioned in various locations on a central table in a reverberant room, and their outputs were recorded while four target talkers seated at the table read some text or had natural conversations. The recorded speech signals have been made publicly available [98]. Distributed beamforming tools and tutorials covering distributed speech enhancement and random microphone deployment have been made available [99]. Materials and resources associated with the distributed speech recognition (DSR) research conducted by IBM have been archived and sourced [100]. A coherence-based (31) source counting method for distributed scenarios has been made available by Donley [101]; the method counts the number of active speakers (up to eight) in a spontaneous meeting in reverberant environments.

University of Illinois have provided a dataset which facilitates distributed source separation and augmented listening research [102]. The dataset is recorded using 10 speech sources and 160 microphones in a large, reverberant conference room. The applied microphone array includes wearable sensors and microphones connected to tablets.

7. Conclusion

This review paper discussed recent advances in the context of distributed microphone arrays and signal processing. Standard dereverberation, speech separation, and source counting methods have been successfully adapted to the context of distributed signal processing using novel features and machine learning. Most existing ad hoc beamforming methods suffer from limiting assumptions that make them niche applications. Issues such as real-time source localization and DOA estimation are still challenging.

References

[1] S. Pasha, C. Ritz and J. Lundgren, “A Survey on Ad Hoc Signal Processing: Applications, Challenges and State-of-the-Art Techniques,” in *International Symposium on Signal Processing and Information Technology (ISSPIT)*, Ajman, United Arab Emirates, 2019, DOI:10.1109/ISSPIT47144.2019.9001860.

[2] M. J. Taghizadeh, Enabling speech applications using ad hoc microphone arrays: PhD dissertation, Lausanne: EPFL, 2015,https://publications.idiap.ch/downloads/papers/2015/Taghizadeh_THESIS_2015.pdf.

[3] S. Pasha, Analysis and Enhancement of Spatial Sound Scenes Recorded using Ad-Hoc Microphone Arrays, Doctor of Philosophy thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2017, https://ro.uow.edu.au/theses1/450.

[4] J. Benesty, J. Chen and J. Huang, Microphone Array Signal Processing, Verlag Berlin Heidelberg: Springer, 2008, DOI: 10.1007/978-3-540-78612-2.

[5] V. M. Tavakoli, J. R. Jensen, M. Christenseny and J. Benesty, “Pseudo-coherence-based MVDR beamformer for speech enhancement with ad hoc microphone arrays,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015, DOI: 10.1109/ICASSP.2015.7178453.

[6] N. D. Gaubitch, J. Martinez, W. B. Kleijn and R. Heusdens, “On near-field beamforming with smartphone-based ad-hoc microphone arrays,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6953345.

[7] M. Souden, K. Kinoshita, M. Delcroix and T. Nakatani, “Location Feature Integration for Clustering-Based Speech Separation in Distributed Microphone Arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 22(2), 354-367, 2014, DOI: 10.1109/TASLP.2013.2292308.

[8] S. Pasha and C. Ritz, “Clustered multi-channel dereverberation for ad-hoc microphone arrays,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Hong Kong, 2015, DOI: 10.1109/APSIPA.2015.7415519.

[9] S. Gregen and R. Martin, “Estimating Source Dominated Microphone Clusters in Ad-Hoc Microphone Arrays by Fuzzy Clustering in the Feature Space,” in *Speech Communication; 12. ITG Symposium*, Paderborn, Germany, 2016, Print ISBN: 978-3-8007-4275-2.

[10] C. Evers, Y. Dorfan, S. Gannot and P. Naylor, “Source tracking using moving microphone arrays for robot audition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, DOI: 10.1109/ICASSP.2017.7953337.

[11] S. Pasha, Y. Zou and C. Ritz, “Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses,” in *China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Chengdu, 2015, DOI: 10.1109/ChinaSIP.2015.7230367.

[12] Z. Liu, Z. Zhang, L. W. He and P. Chou, “Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays,” in *International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 2007, DOI: 10.1109/ICASSP.2007.366347.

[13] S. Araki, H. Swada, R. Mukai and S. Makino, “DOA Estimation for Multiple Sparse Sources with Arbitrarily Arranged Multiple Sensors,” *Journal of Signal Processing Systems,* 63(3), 265–275, 2011, https://doi.org/10.1007/s11265-009-0413-9.

[14] L. Krishnan, P. Teal and T. Betlehem, “A robust sparse approach to acoustic impulse response shaping,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, DOI: 10.1109/ICASSP.2015.7178067.

[15] T. v. Waterschoot, “Distributed estimation of cross-correlation functions in ad-hoc microphone arrays,” in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362385.

[16] V. M. Tavakoli, J. R. Jensen, M. G. Christensen and J. Benesty, “A Framework for Speech Enhancement With Ad Hoc Microphone Arrays,” *ACM Transactions on Audio, Speech, and Language Processing,* 24(6), 1038-1051, June 2016, DOI: 10.1109/TASLP.2016.2537202.

[17] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: A signal processing perspective,” in *18th Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Ghent, 2011, DOI: 10.1109/SCVT.2011.6101302.

[18] I. Himawan, Speech recognition using ad-hoc microphone arrays, PhD dissertation, Queensland university of technology, 2010, https://pdfs.semanticscholar.org/e95b/ad880c47be6e146dd75b82bc72e51e97db2f.pdf.

[19] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He and A. Colburn, “Distributed meetings: A meeting capture and broadcasting system,” in *ACM Multimedia*, Juan-les-Pins, France, 2002, DOI: 10.1109/ICASSP.2003.1202753.

[20] V. Tavakoli, J. Jensen, R. Heusdens, J. Benesti and M. Christensen, “Distributed max-SINR speech enhancement with ad hoc microphone arrays,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, DOI: 10.1109/ICASSP.2017.7952136.

[21] A. Asaei, H. Bourlard, M. Taghizadeh and V. Cehver, “Model-based sparse component analysis for reverberant speech localization,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, DOI: 10.1109/ICASSP.2014.6853835.

[22] A. Bertrand and M. Moonen, “Robust distributed noise reduction in hearing aids with external acoustic sensor nodes,” *EURASIP Journal on*

*Advances in Signal Processing(*14, 2009, https://doi.org/10.1155/2009/530435.

[23] Y. Jia, L. Yu and I. Kozintsev, "Distributed Microphone Arrays for Digital Home and Office," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006, DOI: 10.1109/ICASSP.2006.1661463.

[24] Diane J. Cooka, Juan C. Augustob and Vikramadit, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive and Mobile Computing,* 5, 277–298, August 2009, https://doi.org/10.1016/j.pmcj.2009.04.001.

[25] A. Brutti, M. Omologo and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *NTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, Corpus ID: 2927411.

[26] J. Lee, C. H. Lee, D. W. Kim and B. Y. Kang, "Smartphone-Assisted Pronunciation Learning Technique for Ambient Intelligence," *IEEE Access,* 2016, DOI: 10.1109/ACCESS.2016.2641474.

[27] T. Toyoda, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Traffic monitoring with ad-hoc microphone array," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6954310.

[28] K. Tsuji, "Migration monitoring of fin whales in the southern Chukchi Sea with acoustic methods during 2012–2015," in *Techno-Ocean (Techno-Ocean)*, Kobe, 2016, DOI: 10.1109/Techno-Ocean.2016.7890746.

[29] D. Stowell, E. Benetos and L. Gill, "On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 25(6), 1193-1206, June 2017, DOI: 10.1109/TASLP.2017.2690565.

[30] A. M. McKee and R. A. Goubran, "Chest sound pick-up using a multisensor array," in *SENSORS*, Irvine, CA, 2005, DOI: 10.1109/ICSENS.2005.1597816.

[31] J. Chien, M. Huang, Y. Lin and F. Chong, "A Study of Heart Sound and Lung Sound Separation by Independent Component Analysis Technique," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY, 2006, DOI: 10.1109/IEMBS.2006.260223.

[32] M. krekovic, I. Dokmanic and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7471627.

[33] I. Dokmanic, L. Daudet and M. Vetterli, "From acoustic room reconstruction to slam," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7472898.

[34] I. Dokmanic, L. Daudet and M. Vetterli, "How to localize ten microphones in one finger snap," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014, Electronic ISBN: 978-0-9928-6261-9.

[35] F. Antonacci, "Inference of Room Geometry From Acoustic Impulse Responses," *IEEE Transactions on Audio, Speech, and Language Processing,* 20(10), 2683-2695, 2012, DOI: 10.1109/TASL.2012.2210877.

[36] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda and P. Naylor, "A Single-Channel Non-Intrusive C50 Estimator Correlated With Speech Recognition Performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 24(4), 719-732, April 2016, DOI: 10.1109/TASLP.2016.2521486.

[37] S. Pasha, C. Ritz and Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, DOI: 10.1109/APSIPA.2016.7820873.

[38] T. K. Hon, L. Wang, J. D. Reiss and A. Cavallaro, "Fine landmark-based synchronization of ad-hoc microphone arrays," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362600.

[39] S. Araki, N. Ono, K. Kinoshita and M. Delcroix, "Estimation of Sampling Frequency Mismatch between Distributed Asynchronous Microphones under Existence of Source Movements with Stationary Time Periods Detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019), 785-789, DOI: 10.1109/ICASSP.2019.8683192.

[40] S. Pazors, M. Hurtado and C. Muravchik, "On Sparse Methods for Array Signal Processing in the Presence of Interference," *Antennas and Wireless Propagation Letters,* 14, 1165-1168, 2015, DOI: 10.1109/LAWP.2015.2394233.

[41] M. Taseska, S. Markovich-golan, E. Habets and S. Gannot, "Near-field source extraction using speech presence probabilities for ad hoc microphone arrays," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014, DOI: 10.1109/IWAENC.2014.6954000.

[42] A. D. Sarwate and K. Chaudhuri, "Signal Processing and Machine Learning with Differential Privacy: Algorithms and Challenges for Continuous Data," *IEEE Signal Processing Magazine,* 30(5), 86-94, Sept. 2013, DOI: 10.1109/MSP.2013.2259911.

[43] R. Wehrmann, "Concepts of improving hands-free speech communication," in *IEEE International Symposium on Circuits and Systems*, San Diego, CA, 1992, DOI: 10.1109/ISCAS.1992.230435.

[44] P. Chevalier and A. Maurice, "Blind and informed cyclic array processing for cyclostationary signals," in *9th European Signal Processing Conference (EUSIPCO 1998)*, Rhodes, 1998, Print ISBN: 978-960-7620-06-4.

[45] E. Vincent, N. Bertin, R. Griboval and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *EEE Signal Processing Magazine,* 31(3), 107-11, May 2014, DOI: 10.1109/MSP.2013.2297440.

[46] C. Vanwynsberhea, P. Challande, J. Marchal, R. marchiano and F. Ollivier, "A robust and passive method for geometric calibration of large arrays," *The Journal of the Acoustical Society of America,* 139, p. 1252, 2016, https://doi.org/10.1121/1.4944566.

[47] I. McCowan, M. Lincoln and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *Transactions on Audio,Speech and Language Processing,* 16(3), 666-670, 2008, DOI: 10.1109/TASL.2007.911428.

[48] D. McCarthy and F. Boland, "A method for source-microphone range estimation, using arrays of unknown geometry, in reverberant room environments," in *15th European Signal Processing Conference*, Poznan, 2007, Print ISBN: 978-839-2134-04-6.

[49] I. Himawan, I. McCowan and S. Sridhan, "Clustered Blind Beamforming From Ad-Hoc Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing,* 19(4), 661-676, May 2011, DOI: 10.1109/TASL.2010.2055560.

[50] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kaohsiung, 2013, DOI: 10.1109/APSIPA.2013.6694323.

[51] I. Dokmanic, J. Raineri and M. Vetterli, "Relax and unfold: Microphone localization with Euclidean distance matrices," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, 2015, DOI: 10.1109/EUSIPCO.2015.7362386.

[52] Y. Tian, Z. Chen and F. Yin, "Distributed Kalman filter-based speaker tracking in microphone array networks," *Applied Acoustics,* 89), 71-77, 2015, https://doi.org/10.1016/j.apacoust.2014.09.004.

[53] A. Asaei, N. Mohamadiha, M. J. Taghizadeh, S. Doclo and H. Bourlard, "On application of non-negative matrix factorization for ad hoc microphone array calibration from incomplete noisy distances," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015, DOI: 10.1109/ICASSP.2015.7178460.

[54] S. Pasha and C. Ritz, "Informed source location and DOA estimation using acoustic room impulse response parameters," in *International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, 2015, DOI: 10.1109/ISSPIT.2015.7394316.

[55] I. Dokmanic, Y. Lu and M. Vetterli, “Can one hear the shape of a room: The 2-D polygonal case,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, DOI: 10.1109/ICASSP.2011.5946405.

[56] S. Pasha, J. Donley, C. Ritz and X. Y. Zou, “Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings,” in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, 2017, DOI: 10.1109/HSCMA.2017.7895582.

[57] M. Taseska and A. P. Habets, “Spotforming using distributed microphone arrays,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2013, DOI: 10.1109/WASPAA.2013.6701876.

[58] S. Woźniak and K. Kowalczyk, “Passive Joint Localization and Synchronization of Distributed Microphone Arrays,” *Signal Processing Letters,* 26(2), 292-296, Feb. 2019, DOI: 10.1109/LSP.2018.2889438.

[59] Q. Zhang, W. Zhang, J. Feng and R. Tang, “Distributed Acoustic Source Tracking in Noisy and Reverberant Environments With Distributed Microphone Networks,” *IEEE Access,* 8), 9913-9927, 2020, DOI: 10.1109/ACCESS.2020.2965210.

[60] S. Vesa, “Binaural Sound Source Distance Learning in Rooms,” *Transactions on Audio, Speech, and Language Processing,* 17(8), 1498-1507, 2009, DOI: 10.1109/TASL.2009.2022001.

[61] I. Himawan, S. Sridharan and I. McCowan, “Dealing with uncertainty in microphone placement in a microphone array speech recognition system,” in *International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, DOI: 10.1109/ICASSP.2008.4517922.

[62] M. Chen, Z. Liu, W. He, P. Chou and Z. Zhang, “Energy-based position estimation of microphones and speakers for ad hoc microphone arrays,” in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 2007, DOI: 10.1109/ASPAA.2007.4393035.

[63] M. J. Taghizadeh, R. Parhizkar, P. N. Garner and H. Bourlard, “Euclidean distance matrix completion for ad-hoc microphone array calibration,” in *18th International Conference on Digital Signal Processing (DSP)*, Fira, 2013, DOI: 10.1109/HSCMA.2014.6843239.

[64] M. J. Taghizadeh, N. Carner, H. Bourlard and A. Asaei, “Ad Hoc Microphone Array Calibration: Euclidean Distance Matrix Completion Algorithm and Theoretical Guarantees,” *Signal Processing,* p. https://doi.org/10.1016/j.sigpro.2014.07.016, 2014.

[65] M. Taghizadeh, A. Asaei, P. Garner and h. Bourlard, “Ad-hoc microphone array calibration from partial distance measurements,” in *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Villers-les-Nancy, 2014, DOI: 10.1109/HSCMA.2014.6843239.

[66] N. Ono, K. Shibata and H. Kameoka, “Self-localization and channel synchronization of smartphone arrays using sound emissions,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, DOI: 10.1109/APSIPA.2016.7820778.

[67] L. Wang, T. K. Hon and A. Cavallaro, “Self-Localization of Ad-Hoc Arrays Using Time Difference of Arrivals,” *IEEE Transactions on Signal Processing,* 64(4), 1018-1033, Feb.15, 2016, DOI: 10.1109/TSP.2015.2498130.

[68] N. D. Gaubitch, W. B. Kleijn and R. Heusdens, “Auto-localization in ad-hoc microphone arrays,” in *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, DOI: 10.1109/ICASSP.2013.6637618.

[69] E. Robledo, T. S. Wada and B. Juang, “On Dealing with Sampling Rate Mismatches in Blind Source Separation and Acoustic Echo Cancellation,” in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2007, DOI: 10.1109/ASPAA.2007.4393044.

[70] H. Chiba, N. Ono, S. Miyabe, Y. Takashaahi, T. Yamada and S. Makino, “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les pins, France, 2014, DOI: 10.1109/IWAENC.2014.6954007.

[71] S. Miyabe, N. Ono and S. makino, “Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation,” *Signal Processing,* p. 185–196, 2015, https://doi.org/10.1016/j.sigpro.2014.09.015.

[72] N. Ono, H. Kohno, N. Ito and S. Sagayama, “Blind alignment of asynchronously recorded signals for distributed microphone array,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2009, DOI: 10.1109/ASPAA.2009.5346505.

[73] P. Pertila, M. S. Hamalainen and M. Mieskolainen, “Passive Temporal Offset Estimation of Multichannel Recordings of an Ad-Hoc Microphone Array,” *IEEE Transactions on Audio, Speech, and Language Processing,* 21(11), 2393-2402, 2013, DOI: 10.1109/TASLP.2013.2286921.

[74] N. J. Bryan, P. Smargadis and G. J. Mysore, “Clustering and synchronizing multi-camera video via landmark cross-correlation,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012, DOI: 10.1109/ICASSP.2012.6288396.

[75] R. Lienhart, I. Kozintsev, S. Wher and M. Yeung, “On the importance of exact synchronization for distributed audio signal processing,” in *International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003, DOI: 10.1109/ASPAA.2003.1285842.

[76] T. Nakatani, T. Yoshioka, K. Kinoshita and M. Miyoshi, “Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation,” in *International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, 2008, DOI: 10.1109/ICASSP.2008.4517552.

[77] S. Pasha, C. Ritz and Y. X. Zou, “Spatial multi-channel linear prediction for dereverberation of ad-hoc microphones,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, DOI: 10.1109/APSIPA.2017.8282306.

[78] S. Pasha and J. Lundgren, “Multi-Channel Compression and Coding of Reverberant Ad-Hoc Recordings Through Spatial Autoregressive Modelling,” in *30th Irish Signals and Systems Conference (ISSC)*, Maynooth, Ireland, 2019, DOI: 10.1109/ISSC.2019.8904918.

[79] M. Delcroix, T. Yosghioka, A. Ogawa and Y. Kubo, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge,” *Proceedings of REVERB Workshop ,* 2014.

[80] V. M. Tavakoli, J. R. Jensen, R. Heusdens, J. Benesty and M. G. Christensen, “Ad hoc microphone array beamforming using the primal-dual method of multipliers,” in *24th European Signal Processing Conference (EUSIPCO)*, Budapest, 2016, DOI: 10.1109/EUSIPCO.2016.7760416.

[81] W. S. Wood, E. Hadad, I. Merks, B. Xu and S. Gannot, “A real-world recording database for ad hoc microphone arrays,” in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2015, DOI: 10.1109/WASPAA.2015.7336915.

[82] I. Himawan, I. McCowan and S. Sridharan, “Clustering of ad-hoc microphone arrays for robust blind beamforming,” in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, DOI: 10.1109/ICASSP.2010.5496201.

[83] T. C. Lawin, S. Stenzel, J. Freudenberger and S. Doclo, “Generalized Multichannel Wiener Filter for Spatially Distributed Microphones,” in *Speech Communication; 11. ITG Symposium*, Erlangen, Germany, 2014, Print ISBN: 978-3-8007-3640-9.

[84] T. Ballal and C. J. Bleakley, “DOA estimation of multiple sparse sources using three widely-spaced sensors,” in *17th European Signal Processing Conference*, Glasgow, 2009, Print ISBN: 978-161-7388-76-7.

[85] M. H. Hennecke and G. A. Fink, “Towards acoustic self-localization of ad hoc smartphone arrays,” in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Edinburgh, 2011, DOI: 10.1109/HSCMA.2011.5942378.

[86] A. Moonen and M. Bertrand, “Energy-based multi-speaker voice activity detection with an ad hoc microphone array,” in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, DOI: 10.1109/ICASSP.2010.5496183.

[87] S. Pasha, J. Donley and C. Ritz, “Blind speaker counting in highly reverberant environments by clustering coherence features,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, DOI: 10.1109/APSIPA.2017.8282303.

[88] Z. Liu, “Sound source separation with distributed microphone phone arrays in the presence of clock synchronization errors,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, 2008.

[89] S. Araki, N. Ono, K. Kinoshita and M. Delcroix, “Projection Back onto Filtered Observations for Speech Separation with Distributed Microphone Array,” in *8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Le gosier, Guadelope, 2019, DOI: 10.1109/CAMSAP45676.2019.9022666.

[90] J. Dmochowsky, Z. Liu and P. Chou, “Blind source separation in a distributed microphone meeting environment for improved teleconferencing,” in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Las Vegas, NV, USA, 2008, DOI: 10.1109/ICASSP.2008.4517553.

[91] V. M. Tavakoli, J. R. Jensen, J. Benesty and M. G. Christensen, “A partitioned approach to signal separation with microphone ad hoc arrays,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, DOI: 10.1109/ICASSP.2016.7472272.

[92] S. Golan, S. Gannot and I. Cohen, “Distributed GSC beamforming using the relative transfer function,” in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, 2012, Print ISBN: 978-1-4673-1068-0.

[93] D. Salvati, C. Drioli and L. Foresti, “On the use of machine learning in microphone array beamforming for far-field sound source localization,” in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, 2016, DOI: 10.1109/MLSP.2016.7738899.

[94] B. Clarke, E. Fokoue and H. Zhang, Principles and Theory for Data Mining and Machine Learning, Springer-Verlag New York Inc, 1st August 2009, https://doi.org/10.1007/978-0-387-98135-2.

[95] S. Gregen, A. Nagathil and R. Martin, “Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays,” in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Vancouver, BC, 2013, DOI: 10.1109/ICASSP.2013.6638347.

[96] “IEEE DCASE challenge,” [Online]. Available: http://dcase.community/. [Accessed 30 3 2020].

[97] S. Gergen and R. Martin, “Linear Combining of Audio Features for Signal Classification in Ad-hoc Microphone Arrays,” in *Speech Communication; 11. ITG Symposium*, Erlangen, Germany, 2014.

[98] “Ad-hoc recording signals,” http://www.eng.biu.ac.il/gannot/speech-enhancement

[99] “Prof. Gannot speech enhancement tutorial,” http://www.eng.biu.ac.il/gannot/tutorials-and-keynote-addresses/introduction-to-distributed-speech-enhancement-algorithms-for-ad-hoc-microphone-arrays-wireless-acoustic-sensor-networks/.

[100] IBM, “Distributed Speech Recognition,” https://www.ibm.com/blogs/research/2019/10/asr-deep-learning/.

[101] J. Donley. https://github.com/JacobD10/CoherenceBasedSourceCounter.

[102] R. M. Corey, M. D. Skarha and A. C. Singer, Composers, *Massive Distributed Microphone Array Dataset.* [Sound Recording]. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-6216881_V1 . 2019.