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

Shahab Pasha*, 1, Jan Lundgren1, Christian Ritz2, Yuexian Zou3
1STC Research Centre, Mid-Sweden University, Sundsvall 85230, Sweden
2School of Electrical, Computer and Telecommunication Engineering, University of Wollongong, NSW 2500, Australia
3ADSPLAB/Intelligent Lab, School of ECE, Peking University, Shenzhen 518055, China

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 benefiting 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
www.astesj.com
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.](image)

<table>
<thead>
<tr>
<th>Table 1: Ad hoc microphone arrays compared with compact arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distributed</strong></td>
</tr>
<tr>
<td>Unknown structure [9]</td>
</tr>
<tr>
<td>Changeable microphone locations [10]</td>
</tr>
<tr>
<td>Unknown inter-channel time delays [11]</td>
</tr>
<tr>
<td>Inconsistent gain within the array [12]</td>
</tr>
<tr>
<td>Uncertain direction of arrival (DOA) definition [13]</td>
</tr>
<tr>
<td>Large phase differences (i.e., spatial aliasing)</td>
</tr>
<tr>
<td>Inconsistent signal quality [5]</td>
</tr>
</tbody>
</table>

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 $i$th channel forming the $m$th node is modelled as

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

where $s_j(n)$ is the $j$th sound source, $h_{m,i,j}(n)$ represents the room acoustic response between the $i$th channel forming the $m$th 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

$$r_{m,i} = [x_{m,i}, y_{m,i}, z_{m,i}],$$

and the $j$th source location is

$$r_{s,j} = [x_j, y_j, z_j].$$

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

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

For ad hoc scenarios where all the nodes are single channel

$$M_{ch} = M.$$ (5)

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

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

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

$$x_m(n) = s(n) * h_m(n) + v_m(n) + w_m(n).$$ (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
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 $M_n$ channels (4) within an ad hoc array are modeled as

$$x(n) = [x_1(n), ..., x_M(n)]^T. \quad (8)$$

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:

$$[a_{m,0}, ..., a_{m,L}], \quad (9)$$

where $a_0$ is the direct path impulse amplitude, and

$$[\tau_{m,0}, ..., \tau_{m,L}]. \quad (10)$$

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

<table>
<thead>
<tr>
<th>Method/features</th>
<th>Application</th>
<th>Reference(s)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise coherence</td>
<td>Clustering beamforming</td>
<td>Himawan et al. [49]</td>
<td>2011</td>
</tr>
<tr>
<td>Voice activity detection</td>
<td>Speech enhancement</td>
<td>Sakanashi et al. [50]</td>
<td>2013</td>
</tr>
<tr>
<td>Sparsity analysis</td>
<td>Localization</td>
<td>Asaei et al. [21]</td>
<td>2014</td>
</tr>
<tr>
<td>Time delay and sound level</td>
<td>Speech separation</td>
<td>Souden et al. [7]</td>
<td>2014</td>
</tr>
<tr>
<td>Signal power</td>
<td>Traffic monitoring</td>
<td>Toyoda et al. [27]</td>
<td>2014</td>
</tr>
<tr>
<td>Euclidian distance matrix (EDM)</td>
<td>Microphone localization</td>
<td>Dokmanic et al. [51]</td>
<td>2015</td>
</tr>
<tr>
<td>Kurtosis of the linear prediction residuals</td>
<td>Clustering</td>
<td>Pasha et al. [8]</td>
<td>2015</td>
</tr>
<tr>
<td>Generalized cross-correlation</td>
<td>Speaker tracking</td>
<td>Tian et al. [52]</td>
<td>2015</td>
</tr>
<tr>
<td>Non-negative matrix factorization (NMF)</td>
<td>Calibration</td>
<td>Asaei et al. [53]</td>
<td>2015</td>
</tr>
<tr>
<td>Reverberation</td>
<td>DOA estimation</td>
<td>Pasha et al. [54]</td>
<td>2015</td>
</tr>
<tr>
<td>Cepstral features</td>
<td>Interference suppression</td>
<td>Gregen et al. [9]</td>
<td>2016</td>
</tr>
<tr>
<td>$C_{50}$ (short-time reverberation)</td>
<td>Multi-talk detection</td>
<td>Pasha et al. [37]</td>
<td>2016</td>
</tr>
<tr>
<td>Magnitude-squared coherence (MSC)</td>
<td>Crosstalk and multi-talk detection</td>
<td>Pasha et al. [37]</td>
<td>2016</td>
</tr>
<tr>
<td>Echoes</td>
<td>Room geometry reconstruction</td>
<td>Dokmanic et al. [33, 55]</td>
<td>2016</td>
</tr>
<tr>
<td>Coherent-to-diffuse ratio</td>
<td>Multi-talk and crosstalk detection</td>
<td>Pasha et al. [56]</td>
<td>2017</td>
</tr>
<tr>
<td>Power spectral density (PSD)</td>
<td>Spotforming</td>
<td>Habet et al. [57]</td>
<td>2017</td>
</tr>
<tr>
<td>Signal to interference and noise ratio</td>
<td>Noise cancellation</td>
<td>Tavakoli et al. [20]</td>
<td>2017</td>
</tr>
<tr>
<td>Time difference</td>
<td>Microphone localization</td>
<td>Woźniak et al. [58]</td>
<td>2019</td>
</tr>
<tr>
<td>Distributed unscented Kalman particle filter (DUKPF)</td>
<td>Source localization</td>
<td>Zhang et al. [59]</td>
<td>2020</td>
</tr>
</tbody>
</table>
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 \( g = \{g_1, ..., g_M\} \), (1) can be rewritten as

\[
x_m(n) = s(n) * g_m \bar{h}_m(n) + v_m(n),
\]

for one active source, where \( g_m \) and \( \bar{h}_m(n) \) 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 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

\[
\begin{align*}
    TOA_{s,m} &= \frac{|r_x - r_m|}{c} + \delta_m \\
    TDOA_{m,n} &= \frac{|r_x - r_m|}{c} - \frac{|r_x - r_n|}{c} + (\delta_m - \delta_n) + (T_{om} - T_{on}),
\end{align*}
\]

where \( \delta_m \) and \( T_{om} \) represent the internal delay and onset time of microphone \( m \), respectively (Figure 3) [67, 68], and \( r_x = \).
$\{x_m, y_m, z_m\}^T$, $r_m = [x_m, y_m, z_m]^T$ are the source, microphone $m$, and microphone $\bar{m}$ 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

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

(14)

where $\tau_{mm} = 0$ for $m = 1$ to $M$ and $\tau_{mn} = \tau_{\bar{m}n}$ for all $m$ and $\bar{m}$ 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

$$r_m(c) = E(x_m(n)x_m(n + c)), \ c = 0, 1, 2, \ldots$$

where $E$ is the mathematical expectation.

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

(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 $\beta = [\beta_1, \ldots, \beta_M]$, the weighted average autocorrelation function is calculated as

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

(17)

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

$$w_s = \begin{bmatrix} \bar{r}_w(0) & \cdots & \bar{r}_w(P_{\text{short}} - 1) \\ \vdots & \ddots & \vdots \\ \bar{r}_w(P_{\text{short}} - 1) & \cdots & \bar{r}_w(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_w(1) \\ \vdots \\ \bar{r}_w(P_{\text{short}}) \end{bmatrix},$$

(18)

and the pre-whitened signal is

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

(19)

where $w_s = \{w_{s,1}, \ldots, w_{s,P_{\text{short}}}\}$ [13]. Having the source-to-microphone distances $\{q_{1,s}, \ldots, q_{M,s}\}$, the ideal distance weights are $\mathbf{q} = \{1/q_{1,s}, \ldots, 1/q_{M,s}\}$. 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 $y(n)$ (8), the DSB output is calculated as

$$\hat{x}_{\text{DSB}} = \sum_{m=1}^{M} x_m(n - \tau_{m,\text{ref}}),$$

(20)

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

$$\mathbf{W} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{q}_x \mathbf{w},$$

(21)

where $\mathbf{q}_x = E(\mathbf{x} \mathbf{x}^H)$ is the covariance. The solution is

$$\mathbf{W} = \mathbf{q}_x^{-1} \mathbf{h}^H \mathbf{q}_x^{-1} \mathbf{h}$$

(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]:

www.astesj.com
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 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 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].

### 4.5. 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

$$ x(n) = h(n) * s(n) + v(n), $$

where $x(n) = [x_1(n), \ldots, x_M(n)]^T$ contains the multi-channel recording of all $M$ microphones in the array, $h(n)$ is the RIR matrix at each microphone’s location for source $j$, and $v(n)$ is the diffuse noise. The goal is to retrieve $s(n)$ from $x(n)$. 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.

$$ e_{xm}(n) = x_m(n) - \sum_{k=1}^{p} b_k x_m(n - k), $$

where $x_m(n)$ 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

$$ k_{xm}(n) = \frac{E[e_{xm}^4(n)]}{E[e_{xm}^2(n)]^2} - 3 $$

where $E[\cdot]$ denotes the mathematical expectation and $k_{xm}(n)$ 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.
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 Closstalk 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 $\bar{x}_{n,p}(t)$ (29) represents the signals recorded by the two channels, $p \in \{1, 2\}$, at node $n \in \{1, \ldots, N\}$, $\bar{s}(t)$ is the source signal and $\bar{h}_{n,p}(t)$ is the RIR at the nth node as in (6). The CDR features are calculated using the following:

\[
\bar{s}(t) = \bar{s}(t) * \bar{h}_{n,p}(t) + \bar{u}_{n,p}(t),
\]

\[
C_{\nu}(f) = \frac{|\varphi_{\nu_1\nu_2}(f)|^2}{\varphi_{\nu_1}(f) \varphi_{\nu_2}(f)},
\]

\[
C_{x_n,l}(l,f) = \frac{|\varphi_{x_{n,1}x_{n,2}}(l,f)|^2}{\varphi_{x_{n,1}}(l,f) \varphi_{x_{n,2}}(l,f)},
\]

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

\[
CD R_n(l,f) = \frac{C_{\nu}(f) - C_{x_n,l}(l,f)}{C_{x_n,l}(l,f) - C_{\nu}(f)}
\]

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

\[
\overline{CD R_n} = \frac{1}{L(f_B - f_0)} \int_{f_0}^{f_B} \sum_{l=1}^{L} CDR_n(l,f) \, df.
\]

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):

\[
c(l,f) = \frac{|\varphi_{x_{1}x_{2}}(l,f)|^2}{\varphi_{x_{1}x_{1}}(l,f) \varphi_{x_{2}x_{2}}(l,f)},
\]

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

\[
\varphi_{x_{1}x_{2}}(l,f) = \frac{1}{AK} \sum_{a,b} (x_1 * x_2)(Al + a, b) e^{-j2\pi kb},
\]

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

\[
(x_1 * x_2)(\cdot, b) = \sum_{n} x_1(\cdot, n)x_2(\cdot, n + b),
\]

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

\[
y_m(n) = \sum_j s_j(n) * h_{m,j}(n)
\]

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

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 reverbervation 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]:

\[
\hat{X}_c = |F(X_c)|
\]

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

\[
\bar{X}_c = \frac{1}{C_T} \sum_{c=0}^{C_T-1} |X_c|,
\]

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].

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


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


