2017

Analysis and Enhancement of Spatial Sound Scenes Recorded using Ad-Hoc Microphone Arrays

Shahab Pasha
University of Wollongong

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Department of
Engineering and Information Sciences
School of Electrical, Computer and Telecommunication Engineering

Analysis and Enhancement of Spatial Sound Scenes Recorded using Ad-Hoc Microphone Arrays

Shahab Pasha

"This thesis is presented as part of the requirements for the award of the Degree of PhD of the University of Wollongong"

Supervisor:
A/Prof. Christian Ritz

August 2017

Abstract

Ad-hoc microphone arrays formed from the microphones of mobile devices such as smart phones, tablets and notebooks are emerging recording platforms for meetings, press conferences and other sound scenes. As opposed to the Wireless Acoustic Sensor Networks (WASN), ad-hoc microphones do not communicate within the array and location of each microphone is unknown. Analysing speech signals and the acoustic scene in the context of ad-hoc microphones is the goal of this thesis. Despite conventional known geometry microphone arrays (e.g. a Uniform Linear array), ad-hoc arrays do not have fixed geometries and structures and therefore standard speech processing techniques such as beamforming and dereverbearion techniques cannot be directly applied to these. The main reasons for this include unknown distances between microphones an hence unknown relative time delays and the changeable array topology.

This thesis focuses on utilising the side information obtained by the acoustic scene analysis to improve the speech enhancement by ad-hoc microphone arrays randomly distributed within a reverberant environment. New discriminative features are proposed, applied and tested for various signal and audio processing applications such as microphone clustering, source localisation, multi-channel dereverberation, source counting and multi-talk detection. The main contributions of this thesis fall into two categories: 1) Novel spatial features extracted from Room Impulse Responses (RIRs) and speech signals 2) Speech enhancement and acoustic scene analysis methods specifically designed for the ad-hoc arrays.

Microphone clustering, source localisation, speech enhancement, source counting and multi-talk detection in the context of ad-hoc arrays are investigated in this thesis and novel methods are proposed and tested. A clustered speech enhancement and dereverberation method tailored for the ad-hoc microphones is proposed and it is concluded that exclusively using a cluster of microphones located closer to the source, improves the dereverberation performance. Also proposed is a multi-channel speech dereverberation method based on a novel spatial multi-channel linear prediction analysis approach for the ad-hoc microphones. The spatially modified multi-channel linear prediction approach takes into account the estimated relative

distances between the source and the microphones and improves the dereverberation performance. The coherence based features are applied for multi-talk detection and source counting in highly reverberant environments and it is shown that the proposed features are reliable source counting features in the context of ad-hoc microphones. Highly accurate offline source counting and pseudo real-time multi-talk detection results are achieved by the proposed methods.

Acknowledgements

I would like to thank my family: especially my parents for supporting me throughout my studies and all the stages in my life.

I would like to thank my supervisor A/Prof. Christian Ritz for leading me patiently throughout this unpaved way.

I would like to thank all my friends and fellow students especially Gustavo, Jacob, Alanna, Yuxiao and Yi for their feedback, cooperation and of course friendship.

In addition I would like to express my gratitude to all the students and staff at the School of Electrical, Computer and Telecommunication Engineering and the faculty of EIS for making my PhD a great experience.

STATEMENT OF ORIGINALITY

I, Shahab Pasha, declare that this thesis, submitted as part of the requirements for the award of PhD, in the school of electrical, computer and telecommunications engineering, university of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications or assessment at any other academic institution.

Print name: Shahab pasha

STUDENT NUMBER: 4353821

Table of contents

List of figures

List of tables

1 Introduction

New digital devices, such as smart phones and IPads which are increasingly employed as recording tools, are emerging as a convenient alternative to conventional microphone arrays for signal and speech processing applications. Microphone arrays randomly formed by a spontaneous group of recording devices such as sound recorders and smart phones at unknown and changeable locations form a Distributed Microphone Array (DMA) or an ad-hoc array, which is the emerging recording style for applications such as press conferences, lecture halls and meetings (Figure 1-1). The use of microphone arrays in contrast to close talking microphones alleviates the feeling of discomfort and distraction to the user. For this reason, ad-hoc microphone arrays are popular and have been used in a wide range of applications such as teleconferencing, hearing aids, speaker tracking, and as the front-end to speech recognition systems. With advances in sensor and sensor network technology, there is considerable potential for applications that employ ad-hoc networks of microphone-equipped devices collaboratively as a virtual microphone array. By allowing such devices to be distributed throughout the users' environment, the microphone positions are no longer constrained to traditional fixed geometrical arrangements. This flexibility in the means of data acquisition allows different audio scenes to be captured to give a complete picture of the working environment.

Ad-hoc arrays provide wide and flexible spatial coverage for targeting multiple sound sources, however unknown locations, inconsistent sampling frequencies between the microphones, different gains and unsynchronised recordings are the source of uncertainties for joint signal processing methods for applications such as source localisation, speech diarisation, multi-channel noise suppression and dereverberation. Most signal and speech processing applications such as source localisation and separation, speech enhancement and dereverberation are well studied for single channel and conventional microphone arrays of known geometries however there is less literature focusing on the joint analysis of the ad-hoc microphones for these applications.

Although unknown geometry of the ad-hoc arrays causes problems for most of state of the art multi-channel signal processing techniques, it can also be beneficial

for scenarios such as a meeting where participants are spread out in a large area and they might change their positions. The wide and flexible spatial coverage of ad-hoc arrays can be exploited for recording target signals, such as speech, from interfering signals, such as competing speech sources, based on the locations of the sources.

As ad-hoc arrays receive signals at the locations, angles and distances which are not identified and are unique for each microphone therefore the recorded signals cannot be directly applied through standard signal processing tools such as beamformers, Direction of arrival estimators and other acoustic and speech signal application requiring knowledge of the array geometry. For instance, the time differences between the signals received by two adjacent channels in a microphone array of a known geometry (e.g. an Uniform Linear Array) can be easily utilised to calculate the angle of arrival of the source but in the ad-hoc arrays context, even defining adjacent channels and measuring the time differences between the channels can be challenging and sometimes impossible. This example shows that analysis of the signals and the derived information from the signals in the ad-hoc arrays is not straightforward and statistical tools and machine learning techniques are needed to interpret the derived information before any further processing.

Machine learning techniques are believed to be helpful tools for pattern recognition and prediction of unlearned scenarios and they have been successfully applied for binaural source localisation when the inter-channel distance is known or a clean training set is available. These constraints are not easily met in the ad-hoc arrays context where microphones locations and distances are unknown. Despite the fact that machine learning techniques require training data and provide meaningful outputs only under certain circumstances (compliance between the training and the test set), the basic components of machine learning techniques and the artificial neural networks such as feature extraction can be applied in the context of ad-hoc arrays. This thesis investigates the benefits and limitations of different machine learning techniques in order to find suitable techniques and features for speech enhancement and acoustic scene analysis (source localisation, microphone clustering and other similar applications) in the context of ad-hoc arrays.

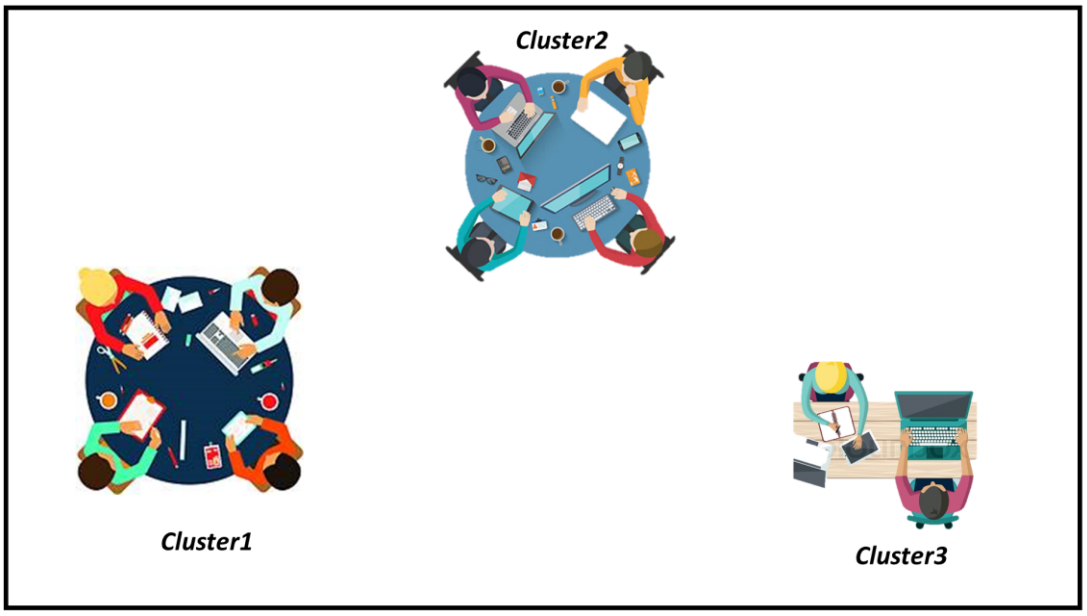


Figure 1-1: Ad-hoc microphone array formed of three clusters

Figure 1-1 illustrates a possible target scenario where a few (usually an unknown number) of meeting participants are spread out at random locations within a reverberant environment (the geometry of the room might or might not be available). Identifying the active source(s) and accordingly choosing the optimised subset of microphones in order to maximise a certain recording quality criteria (Chapter 5). Some side information such as the number of sources, room geometry and relative distances of the microphones and sources can be derived from the raw recorded speech signals and the Room Impulse Responses (RIRs) in order to help the recording process. For instance, in Figure 1-1 the knowledge of having three clusters and the number of participants in each cluster can guide the speech enhancement process by forming clusters of microphones around each source and utilise only one cluster to target each active source. This idea reduces the level of interference in the recorded signal. The knowledge of the number of sources might be available or might be derived from the recorded signals.

Ad-hoc arrays advantages and disadvantages in different applications can be categorised as follows:

Ad-hoc recording advantages:

- Flexible and wide spatial coverage
- Acoustic scene analysis for changeable setups

Ad-hoc recording disadvantages:

- Unknown relative distances and time delays
- Unsynchronised channels
- Unequal microphone gains, internal delays and qualities

In this thesis the following applications of the ad-hoc arrays are investigated and suitable methods for the joint analysis of the ad-hoc recording are proposed:

- Microphone clustering
- Source localisation
- Speech dereverberation
- Multi-talk detection and source counting

1.1 Scope of the research

This thesis focuses on signal processing and acoustic scene analysis techniques for ad-hoc microphone arrays spontaneously formed by digital recording devices at unknown locations. It is assumed that the microphones and other recording devices are not partially or fully connected and therefore they cannot transmit synchronisation timestamps or location cues. In other words the ad-hoc microphones do not form a Wireless Acoustic Sensor Network (WASN) however the joint analysis of the independently recorded signals is discussed.

1.2 Aim of the research

Array signal and speech processing is a well-studied topic however the existing methods are not applicable where the microphone array structure is unknown and the microphones cannot communicate within the array.

The aim of this research is to establish a framework for multi-channel signal processing and acoustic scene analysis for the ad-hoc arrays where the microphones and the source locations are not available. Proposing and extracting novel features from the speech signals and room acoustic responses for each specific task (e.g. Microphone clustering) is the objective of this research.

1.3 Outline of the thesis

This thesis aims to establish a framework for multichannel informed speech enhancement and acoustic scene analysis in the context of ad-hoc arrays. One requirement for this is proposing signal processing methods to obtain side information and cues tailored for the ad-hoc arrays. The proposed clustered dereverberation method for the ad-hoc arrays makes use of derived information such as the source to microphone relative distances and microphone clusters. Although in this thesis this side information is utilised to improve the dereverberation performance but they can be applied separately for other applications in the context of ad-hoc arrays.

Chapter 2 of the thesis reviews the literature published on ad-hoc arrays signal processing, beamforming, microphone clustering, speech enhancement and other applications of ad-hoc arrays such as traffic control. These applications might not be directly related to the speech enhancement application but side information and the applied techniques can help the target application of this thesis. Machine learning techniques previously applied to these applications and also discriminative features derived from speech signals and RIRs for various applications are briefly explained as well. The limitations of the state of the art speech enhancement and source localisation techniques are also briefly explained.

Chapter 3 focuses on microphone clustering, discriminative features and the advantages of clustered signal processing approaches. The novel code-book based clustering and the proposed discriminative features derived from acoustic impulse responses are introduced and compared with the baseline methods and features. This chapter provides the underlying method for clustered dereverberation and also proposes a systematic approach to the microphone clustering evaluation.

Chapter 4 is dedicated to source localisation. The novel surface fitting method for multiple sources is explained. Different features extracted from Room impulse responses and speech signal for source localisation are also investigated and compared. The derived source location information can lead to a more successful microphone clustering and speech enhancement. This chapter introduces a novel source localisation method by the ad-hoc microphones which was missing from the literature.

Chapter 5 of this thesis proposes a novel dereverberation method based on spatial multi-channel linear prediction analysis. The proposed method is compared with the baseline dereveberation methods and recent top performance methods. The clustered dereverberation is also introduced as an informed speech enhancement method. Spatial modification of the Linear Prediction (LP) for the dereverberation task is the main contribution of this chapter.

Chapter 6 uses the estimated coherence features derived from dual ad-hoc nodes for overlap detection and source counting in the context of ad-hoc arrays. Accurate overlap detection and offline source counting results are obtained in the context of ad-hoc arrays where the microphone locations, microphone array geometry and the room geometry are all unknown.

1.4 Contributions of the thesis

- Code-book based microphone clustering algorithm by utilising discriminative features derived from the Room Impulse Responses (RIRs). The proposed clustering method flexibly chooses the number of clusters to form, based on the microphones spatial distribution.
- Surface fitting method for source localisation. The derived features from the RIRs are exploited to localise a source within a room of known geometry. It is shown that the derived features can pinpoint the source location and estimate the Direction of arrival at each microphone location. The accuracy of this method depends on the number of ad-hoc microphones and their distribution pattern within the room.
- Speech enhancement framework based on the multi-channel linear prediction for ad-hoc arrays. A two-phase speech dereverberation scheme is proposed for ad-hoc arrays where the array geometry, source location and the room dimensions are unknown. The proposed method targets the short term and the long term reverberation.
- Clustered multi-channel dereveberation for ad-hoc arrays. The derived side information such as relative microphone to source distances is applied to increase the dereverberation performance by excluding the microphones located far from the source.

- The spatial multi-channel linear prediction as the optimised multi-channel linear prediction for ad-hoc microphones is proposed and applied for short-term dereverberation of speech.
- Multi-talk detection and source counting by utilising cues derived from ad-hoc microphones at unknown positions. Coherence to Diffuse Ratio (CDR) is applied for multi-talk detection and source counting and the results suggest that CDR can effectively discriminate the single talk frames from multi-talk frames and can also be applied for estimating the number of sources.
- Offline source counting in the context of ad-hoc microphones for counting the number of speakers in a meeting based on the coherence features.

1.5 Publications arising from the research

- S. Pasha, J. Donley, C. Ritz and Y. X. Zou, "Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings," *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017, pp. 161-165.
- S. Pasha, C. Ritz and Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, pp. 1-6.
- S. Pasha and C. Ritz, "Informed source location and DOA estimation using acoustic room impulse response parameters," *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, 2015, pp. 139-144.
- S. Pasha and C. Ritz, "Clustered multi-channel dereverberation for ad-hoc microphone arrays," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Hong Kong, 2015, pp. 274-278.
- S. Pasha, Y. X. Zou and C. Ritz, "Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses," *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Chengdu, 2015, pp. 84-88.

1.6 Papers in Preparation

- S. Pasha, J. Donley, C. Ritz, "*Recent advances on ad-hoc signal processing: Applications, challenges and techniques"*, APSIPA transaction, 2017 [Under revision]
- S. Pasha, C. Ritz, Y.X. Zou "Spatial multi-channel linear prediction analysis for dereverberation of ad-hoc microphone arrays" *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* [Under revision]
- S. Pasha, Jacob Donley and C. Ritz, "Speaker counting and diarisation through analysis of the magnitude squared coherence frequency response for highly reverberant signals" *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* [Under revision]

2 Ad-hoc arrays for recording and analysing sound scenes

2.1 Overview

This chapter defines the fundamentals of ad-hoc microphone arrays and reviews their advantages, limitations and applications according to the existing literature. A comparison between blind and informed signal processing is made. The machine learning and data mining techniques applied for signal classification, microphone clustering and other informed approaches to speech enhancement are also mentioned and compared in this chapter. It is also justified why certain machine learning techniques are more suitable for specific signal processing applications and why it is preferred to avoid supervised methods in the context of ad-hoc arrays.

2.2 Ad-hoc arrays and room acoustics

In this section recording by ad-hoc arrays in a general scenario is explained and the main issues and challenges are reviewed.

2.2.1 What is an ad-hoc microphone array?

Let's consider the context of a microphone array in which a set of $m \in \{1,2,\dots,M\}$ randomly distributed microphones (which can be a compact array or a single microphone) is recording an active source. In this thesis each element of the array is referred to as a node, a node can contain a single channel microphone or a multi-channel compact microphone array [1], [2]. At each time index n the m^{th} microphone in the array records it's unique version of the reverberated source signal distorted by the noise and interference which can simplistically be modelled as

$$x_m(n) = s(n) * h_m(t) + w_m(n) + v_m(t) \quad 2\text{-}1$$

where $s(n)$ is the target source signal and $h_m(t)$ is the room impulse response at the m^{th} microphone's location which is the function of room RT_{60}, microphone and source location, room geometry and the walls reflection factor [3] . $w_m(n)$ and $v_m(t)$ represent interference and the noise respectively. $w_m(n)$ is not coherent with the target speech and represents the sum of multiple interfering sources arriving from different locations to the target source.

In this thesis truncated RIRs of length L are mathematically modelled as a train of impulse responses with different time delays, t_k, and amplitudes, a_k:

$$h_m(t) = \sum_{k=1}^{L} a_k \delta(t - t_k) \quad 2\text{-}2$$

Unlike conventional microphone arrays, ad-hoc arrays do not have standard structures and sizes (in terms of the number of the channels and the geometry) and one or more nodes might move during the recording and basically the structure of the array might change. For instance, for a 4-channel ULA with d=2cm inter-channel spacing, the Time Difference of Arrival (TDOA) information are easily obtainable and utilised for applications such as source Direction of Arrival (DOA) estimation and beamforming using the following equations

$$TDOA_{1,2} = \frac{d}{c}, \quad TDOA_{1,3} = \frac{2 \times d}{c}, \quad TDOA_{1,4} = \frac{3 \times d}{c} \qquad 2\text{-}3$$

However in the ad-hoc arrays retrieving this information is computationally expensive and sometimes impossible as d is unknown. The main issues with recording with nodes of microphones that are not connected are synchronisation, sampling frequency mismatch, gain and quality differences. However, recording with a few widely distributed microphones [4] enables the recording of more information about the room geometry and characteristics, source locations and the acoustic setup.

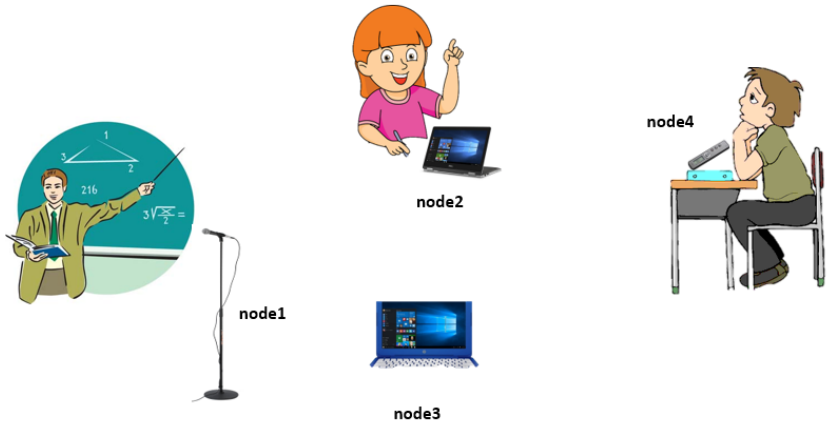


Figure 2-1: An ad-hoc microphone array with four nodes

An example of recording with ad-hoc arrays is discussed in [5] where advantages of applying ad-hoc arrays to record simultaneously active sources are investigated. It is shown that ad-hoc arrays facilitate recording of two competing sources and classifying the recorded signals. It is also concluded that the formation of microphone clusters around each source and assigning one cluster to each source improves the recording quality.

2.2.2 Recording with ad-hoc arrays

In a general meeting scenario where an unknown number of sources (N) or participants are being recorded by a distributed microphone array of M nodes (nodes can contain one or more microphones) at unknown locations, the m^{th} node recording can be represented mathematically as

$$y(n) = \sum_{k=1}^{N} \sum_{m=1}^{M} s_k(n) * h_{mk}(n) + v(n) + w_m(n) \qquad 2\text{-}4$$

where $y(n) = [x_1(n), \dots, x_M(n)]^T$ (from 2-1), contains the multi-channel recording of all M microphones in array and $h_{mk}(n)$ is the Room Impulse Response (RIR) at microphone m location when source k is active. $v(n)$ and $w_m(n)$ are the diffuse noise and the interfering source(s) at the m^{th} microphone location, respectively. It is assumed that the room acoustic impulse response is time invariant and room characteristics do not change during the meeting (closing the blind or curtain change the reverberation time significantly). It is also assumed that $s_k(n) * h_{mk}(n)$ and $w_m(n)$ are not mutually coherent as they are speech signals from different sources with different pitch frequencies.

The objective of recording with ad-hoc arrays is to retrieve the best estimate of $s_k(n)$ from $y_m(n)$. This can be done blindly through utilising all the microphones regardless of their relative distance to the source or by taking into account the spatial information and cues derived from $h_{mk}(n)$ and $y_m(n)$.

The matrix of the sources signals in the discrete time domain can be represented as:

$$S(n) = \begin{bmatrix} s_1(1) & \cdots & s_1(L) \\ \vdots & \ddots & \vdots \\ s_N(1) & \cdots & s_N(L) \end{bmatrix} \qquad 2\text{-}5$$

where L is the frame length which can be very small (e.g. 320 samples at 16kHz sampling rate, 20ms) for real time applications or large for full utterance recordings (e.g. 80000 samples at 16kHz sampling rate, 5s). The matrix S is of size $N \times L$. The recorded signals matrix $\mathbf{X}$ by the microphones can be of a different size as the number of microphones and sources are not always equal.

$$\mathbf{X}(n) = \begin{bmatrix} x_1(1) & \cdots & x_1(L) \\ \vdots & \ddots & \vdots \\ x_M(1) & \cdots & x_M(L) \end{bmatrix} \qquad 2\text{-}6$$

The microphone recording matrix X is of size $M \times L$.

2.2.3 Ad-hoc arrays and the synchronisation problem

The first issue with the recording matrix (2-6) is the problem of unsynchronised signals. Finding the delays between the channels and time-alignment of the signals are essentials to the tasks such as beamforming, Dereverberation and Direction of Arrival (DOA) estimation. If the microphone array geometry and the source-to-microphone distances are available the Time of Arrival between the source and microphone *i* (TOA_{si}) and the Time Difference of Arrival (TDOA) between each two microphones can be calculated by

$$TOA_{si} = \frac{|r_s - r_i|}{c} + \delta_i + T_{oi} \qquad 2\text{-}7$$

$$TDOA_{ij} = \frac{|r_s - r_i|}{c} - \frac{|r_s - r_j|}{c} + (\delta_i - \delta_j) + (T_{oi} - T_{oj}), \qquad 2\text{-}8$$

where δ_i and T_{oi} represent the microphone i internal delay and the onset time respectively [6], [7] and $r_s = [x_s, y_s\ z_s]^T, r_i = [x_i, y_i\ z_i]^T$ and $r_j = [x_j, y_j\ z_j]^T$ are the source, microphone *i* and microphone *j* Cartesian locations in the space, respectively. However in the context of ad-hoc arrays due to the unconventional, unknown and sometimes variable geometry of the array, calculation of the delays is not easily possible. In this thesis it is assumed that all the internal delays and onset times are negligible or exactly the same for all the microphones which leads to a simpler equation

$$TDOA_{ij} = \frac{|r_s - r_i|}{c} - \frac{|r_s - r_j|}{c}. \qquad 2\text{-}9$$

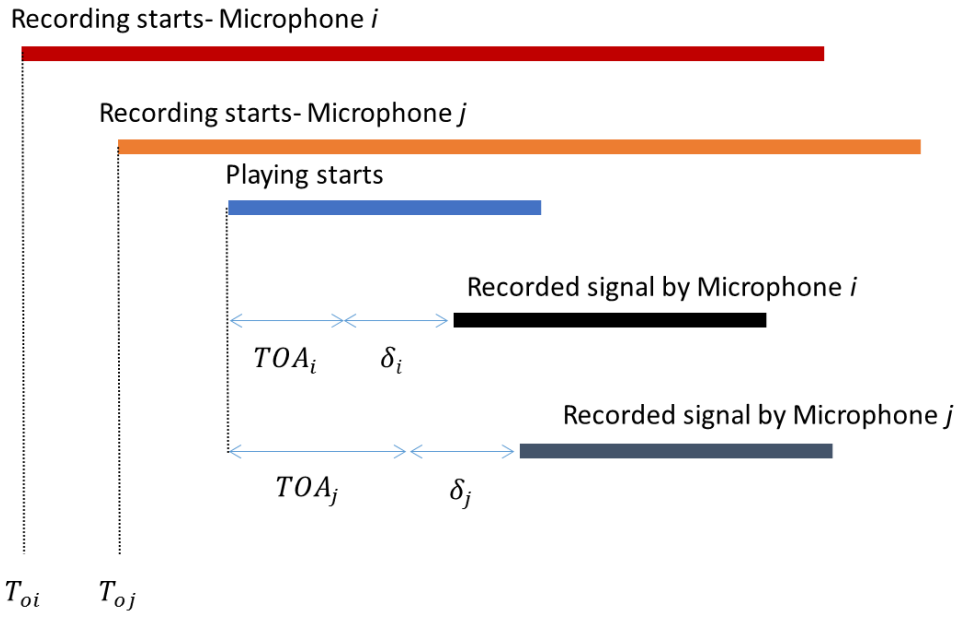


Figure 2-2: Time of Arrival and internal delays

In order to overcome the issues caused by these unsynchronised recordings signal processing methods have been proposed to time-align the signals by iteratively shifting one relative to the other until the highest similarity between the two is achieved. These methods obviously suffer from reverberation and noise and are not computationally feasible for real-time applications.

The goal of synchronisation is to calculate the delay between each two microphones where the acoustic scene is unknown.

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{11} & \cdots & \tau_{1M} \\ \vdots & \ddots & \vdots \\ \tau_{M1} & \cdots & \tau_{MM} \end{bmatrix} \qquad 2\text{-}10$$

where $\tau_{ii} = 0$, for i=1 to M and $\tau_{ij} = -\tau_{ji}$ for all i and j values.

Researchers have used time-alignment of ad-hoc channels for source localisation through Generalised Cross Correlation (GCC) [8] [9] and defining the square errors of time differences based on some parameters [10]. It is concluded that GCC is the computational cost and that it is more suitable for microphones that are already coarsely synchronised so that a full search of all possible correlation lags does not need to be searched.

Even if the ad-hoc recordings are time-aligned, as they each device might have a different sampling rate and they might start the sampling at different times the obtained samples might not align properly [11]. This issue can be critical for

dereverberation and beamforming applications. In this thesis the problem of the signals time alignment is addressed when necessary by state of the art methods and the sampling frequency mismatch is not investigated.

Some more advanced methods use least squares method for temporal offset estimation of static ad-hoc microphone arrays [12] and audio fingerprinting [13]. The proposed fingerprinting methods are inspired by methods that were previously applied to clustering and synchronising unsynchronised multi-camera videos [14] and are based on matching the time-frequency landmarks between two channels. The TDOA then is detected as the peak of the correlation function calculated for audio landmarks. The synchronisation accuracy achieved by conventional audio fingerprinting methods is limited by the time-frequency analysis hop size, with typical values between a few and tens of milliseconds.

Although the focus of this thesis is on the ad-hoc microphone arrays and not ad-hoc wireless acoustic sensor networks with inter-device transmission and synchronisation it is noteworthy that the effect of synchronisaton on Blind Source Separation (BSS) is investigated in [15] and it is concluded that full synchronisation increases the separated source signals quality by an average of 4dB (Signal-to-Interference (SIR)).

As most of the proposed synchronisation methods are able to time-align the signal and calculate the TDOA with an error between 1 to 10 milliseconds, the important factor is the computational cost. The watermark based algorithms are typically more efficient and faster compared to GCC methods [14] as they try to maximise the correlation between the landmarks and not the whole frames [13]. This thesis does not focus on time-alignment and synchronisation and instead applies the state of the art methods.

2.3 Speech enhancement and dereverberation

Speech enhancement [16] covers variety of applications such as noise compensation [16] and dereverberation [17]. Single channel speech enhancement methods [18] [19], [20] do not benefit from the multiple spatial recordings and are based on the prediction and removal of the noise and reverberation in time or frequency domain whereas multichannel speech enhancement methods can discriminate the target signal based on the DOA by the joint analysis and spatial selectivity [21] of the channels.

Speech enhancement methods proposed for the ad-hoc arrays are limited to certain scenarios such as scenarios with nodes of the same structure [1] and are based on basic beamforming techniques [22]. Some noise cancellation methods aim to form clusters around the target speech source and discriminate the speech and the noise by clustering [23].

This thesis proposes a novel speech dereverberation method tailored for the ad-hoc arrays by removing the reverberation in the LP residual signals prior to the beamforming stage (Chapter 5). The proposed method targets the long term reverberation and the short term reverberation [24], [18] separately in order to maximise the dereverberation performance. The clustered dereverberation is also applied in order to increase the dereverberation performance by excluding the highly reverberant signal form the array estimated by the kurtosis of the LP residual signals [25].

2.4 Speech source counting and localisation

Speech processing algorithms need a voice activity detector (VAD), to distinguish the time frames with an active speech source [26] for applications such as speech diarisation and source separation. However, most state of the art VAD methods assume that there is only one speech source and the output of the VAD is a binary value evaluated by precision and recall measurements [27]. In applications such as speech diarisation for meetings and press conferences, it is important to localise the active speaker and distinguish between different speakers. In some speech enhancement methods also distinguishing between the active speech source and interfering sources or the background noise is an essential to applications such as microphone clustering and distributed recording [23].

Inspired by the VAD algorithms, researchers have proposed multi-talk detectors based on some extracted features from ad-hoc recordings where the source and the microphone locations are not known. In [28] a multi-speaker voice activity detection technique, which tracks the power of multiple simultaneous speakers using an ad-hoc microphone array with unknown microphone positions, is proposed and tested. It is concluded that by using short-term power measurements at the different microphones, the multi-speaker VAD problem can be converted into a non-negative blind source separation (NBSS) problem. Other than power, Coherent to Diffuse Ratio (CDR) [29] values calculated or estimated at dual microphone node locations

are also applied for source counting and multi-talk detection when the microphone arrays geometry, source location, and the room dimensions are unknown.

Source localisation with multichannel microphones [30] is a well-studied topic based on binaural analysis and the joint analysis of the channels which is possible if the microphone array geometry is known. The proposed source localisation methods for ad-hoc arrays [31] are applicable to limited scenarios where microphones and sources are collocated.

This thesis overcomes the limitations of the state of the art methods and proposes a surface fitting source localisation method (Chapter 4) that pinpoints the source location within the room.

In Chapter 6 a novel multi-talk detection and source counting method specifically tailored for ad-hoc nodes is proposed and tested.

2.5 Blind and informed acoustic scene analysis approaches

Over the past years, researchers have been trying to exploit the properties of audio sources and signals in order to propose more sophisticated models and algorithms that consume side information (or the estimates of the side information) to guide the scene analysis process. Recently some of the most advanced source separation systems, integrate the feature extraction and the source separation blocks together to achieve an informed process [32]. In Figure 2-3 the process of moving from a blind approach to an informed process is depicted.

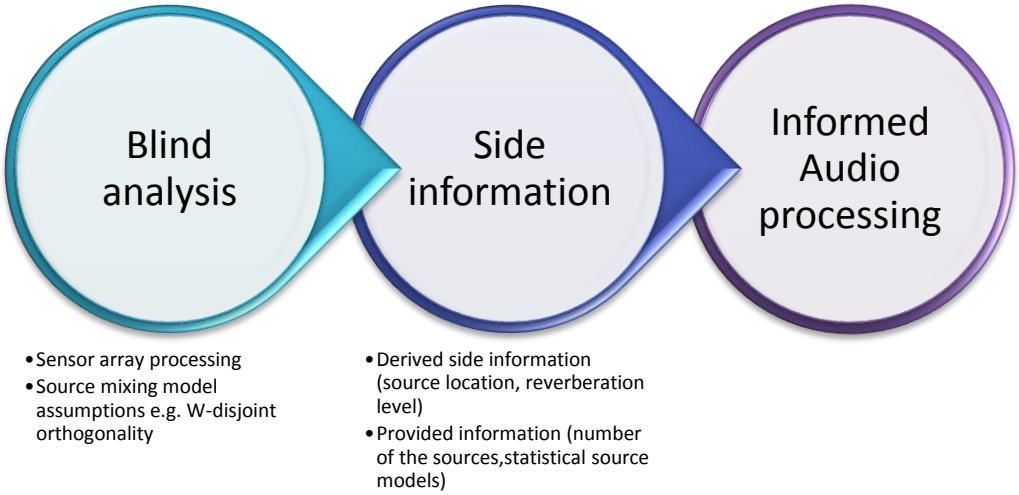


Figure 2-3: from blind to informed speech processing approach

According to the literature, blind approaches such as blind source separation do not exploit any information about the sources nor about the mixing process and analyse the signals without any prior or derived knowledge of the recording scene. Terms such as semi-informed have been previously used for separation techniques relying on highly precise side information, coded and transmitted along with the audio, e.g., the mixing filters and the short-term power spectra of the sources, which can be seen as a form of audio coding. The term guided is used specifically in [32] for source separation approaches which benefit from side-information such as room acoustic. Modelling and exploiting the spatial side-information for signal processing applications is one of the objectives of this thesis. The derived types of side information beneficial for speech enhancement applications are source location, source-to-microphone relative distances, Room acoustics (e.g. reverberation time) and estimation of cross-talk segments. In this thesis the above side information is derived from ad-hoc recordings and is exploited for the informed speech enhancement process.

2.6 Machine learning techniques for informed signal processing

Generally speaking, machine learning techniques are categorised as: 1) Supervised techniques; and 2) Unsupervised techniques. Supervised techniques require a training set which in speech and signal processing applications, is a set of clean utterances spoken by male and female speakers at different locations and setups. It is shown that utilising raw speech signals and utterances does not lead to an optimised training and testing procedure and it is required to extract some discriminative features from this raw data suitable for each application. The discriminative features are highly dependent on the application and the scenario and it can target different aspects of the signal (e.g. cepstral features, relative time delays). On the other hand, unsupervised techniques do not require training and they usually try to use the similarities and dissimilarities between the data points (speech utterances or any other types of acoustic signals such as RIRs). The extracted discriminative features are analysed by the unsupervised methods and based on the mixture and their proximities the categorised output is formed. The main differences between the supervised and unsupervised techniques are: 1) Training requirements; and 2) predefined categories.

2.6.1 Supervised and unsupervised machine learning techniques

Supervised techniques learn the pattern and classify an unseen data point based on the predefined classes. An example of this category can be a classifier (K Nearest Neighbour) or a decision tree [33] that use training sets to learn about the data and then they can categorise an unseen sample based in the training set.

The following figures illustrate the difference between a supervised approach and an unsupervised approach. It is shown that supervised techniques require training (Figure 2-4) and they classify unseen samples based on the predefined classes (Figure 2-5) whereas unsupervised techniques (Figure 2-6) do not require training and cluster the similar samples based on some similarity function (Figure 2-7).

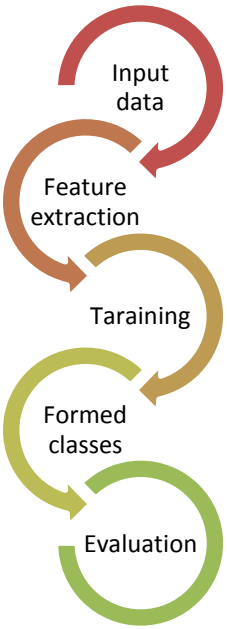


Figure 2-4: Supervised methods based on training

The following examples are supervised machine learning techniques applied for speech enhancement applications:

- Deep learning for binaural speech enhancement [34]
- Speech enhancement based on speaker gender, noise type and the SNR. [35]
- Non-negative matrix factorisation and deep neural networks combined for speech enhancement applications. [36]

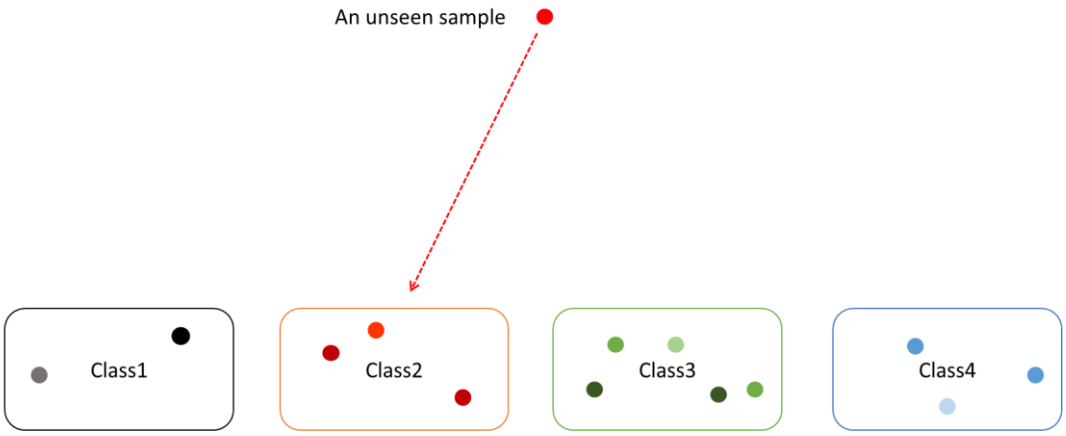


Figure 2-5: Supervised classification

The second type of machine learning and data mining techniques, are the unsupervised methods which do not utilise training and group/cluster similar data-point based on a similarity (dissimilarity) function (e.g. Euclidian distance). Clustering methods such as K-means [33] is an example of these techniques.

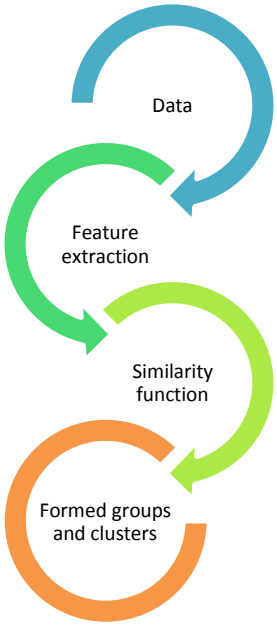


Figure 2-6: Unsupervised methods

Examples of unsupervised machine learning techniques for speech enhancement are:

- Clustering for noise cancellation [37]
- Speaker discrimination by Support Vector machine (SVM) [38]
- Source separation by clustering [39]

The main difference between the supervised and unsupervised techniques is that supervised techniques compare data-points against predefined classes and choose the most suitable class for the unseen sample whereas unsupervised methods analyse the whole data set and form meaningful clusters based on the data distribution.

Figure 2-7: Unsupervised clustering

2.6.2 Extracting discriminative features

Almost all machine learning techniques do not analyse the raw signals and instead extract discriminative features from the data points that 1) have smaller sizes than the data points and 2) Discriminate data points based on the target application. It is also important that the extracted features are easy to calculate especially for real-time applications. Mathematically intensive features might be effective in terms of discriminating the data points but are not suitable for real-time applications.

2.6.3 Performance measures

The formed classes or clusters can be meaningful or just based on the poor selection of the similarity function or the extracted features. The test set and the ground truth are required for the evaluation. For supervised classifier accuracy, confusion matrix, True Positive Ratio (TPR), Receiver Operating Characteristics (ROC) graphs are all used [40]. For unsupervised clustering, cluster purity is the main evaluation measurement [41].

2.7 Ad-hoc arrays applications

The following applications are investigated in the context of ad-hoc arrays. It is briefly mentioned how the outcomes of this research are helpful for speech enhancement and acoustic scene analysis applications. A general survey of ad-hoc arrays applications and challenges with focus on synchronisation and localisation is provided in [42].

2.7.1 Source localisation

Compared with a compact array located at a fixed location ad-hoc arrays can collect more distance cues from the source. These distance cues can be utilised for source localisation applications [31]. Having the knowledge of the source location guides the process of beamforming and microphones clustering for speech enhancement. This problem is investigated in Chapter 4 where a novel surface fitting method for pinpointing the source in a room is proposed and successfully tested.

2.7.2 Microphone localisation

In order to beamform the microphones' signal it is critical to localise the microphones or estimate their distances. Having the microphones' distances, it is possible to calculate the time delays and beamform the signals. [41] [43]. In other words, microphone localisation leads to an informed beamforming and speech enhancement process. Similar to this application in Chapter 3 of this thesis a novel code-book based microphone clustering and segmentation method is proposed.

2.7.3 Noise cancellation and speech enhancement

In an ad-hoc arrays as the channels are not collocated, each microphone receives a different level of noise and one of them is the closest microphone to the noise source. Signals obtained by this microphone can be applied within adaptive methods to estimate and suppress the noise more effectively [1] [44]. The input SNR at each node location is considered as a discriminative feature in order to pick the closest node to the source and achieve a higher noise cancellation outcome. A two stage dereverberation method that targets short-term

reverberation and long term reverberation separately is proposed in Chapter 5 and it is shown that the proposed method outperforms the state of the art dereverberation methods when applied to ad-hoc microphones.

2.7.4 Multi-talk detection

A distributed array of microphone nodes which might be located close to the sources can track the activity of the corresponding sources more accurate than a single compact array which might not be close to any source [28] . Having the knowledge of double-talk and multi-talk frames can help the speech diaraisation and source separation process. A coherence based feature is applied in this thesis (Chapter 6) as a new feature for multi-talk detection and source counting.

2.7.5 Blind source separation

The problem of blind source separation of acoustic mixtures is often addressed using independent component analysis in the frequency domain. Solutions to this problem have been proposed that exploit known properties of both the source signals and the mixing system, but require the microphones to be in a constrained geometry. Methods proposed for this problem in the context of ad-hoc arrays utilises the source estimates to provide a reliable permutation alignment [23] [45].

2.7.6 Speech recognition and acoustic scene analysis

While close talking microphones give the best signal quality and produce the highest accuracy from current Automatic Speech Recognition (ASR) systems, the speech signal enhanced by microphone array has been shown to be an effective alternative in a noisy environment [46]. The process of feature extraction and utilising the Hidden Markov Model (HMM) for this particular pattern recognition problem by analysing the speech model parameters is proposed in [46] for the ad-hoc arrays.

2.7.7 Other applications

In a novel application for ad-hoc arrays, vehicle sounds are recorded by ad-hoc microphone arrays and through peak detection of the power envelope, the

number of vehicles is counted [47]. The issues of different sampling frequencies and asynchronous recordings are also discussed. The focus of that research is on counting the number moving vehicles but as the only applied feature is the signals power, the proposed method might be applicable to source counting application as well.

Video and audio recording with more than one microphone and camera is another application of the ad-hoc microphone arrays and it is reviewed in [14]. The issue of synchronisation is also investigated in that research.

2.8 The applied discriminative features and their applications

Rather than applying the machine learning and data mining techniques on the raw audio or speech signals directly, the data is typically transformed to a reduced parametric representation [48]. As the feature extraction is an inevitable part of any machine learning process, here a brief review of the applied features and their applications in the speech processing literature is presented. Some features such as phase information have been shown to be unreliable for microphone discrimination applications in the context of the ad-hoc arrays [49].

2.8.1 Norm of the pseudo-coherence-vector

The pseudo-coherence-vector is applied in [22] to choose the node that yields the highest output quality after beamforming. This feature is defined as

$$\rho_{x_{n,1},X_{n,2}}(k,t) = \frac{E[x_{n,1}(k,t)X_{n,2}^{*}(k,t)]}{E\left[\left|X_{n,1}(k,t)\right|^{2}\right]} \qquad 2\text{-}11$$

where E[.] and * denote mathematical expectation and complex conjugate respectively and $\rho_{x_{n1},X_{n,2}}(k,t)$ is the pseudo coherence vector of length M between $x_{n,1}(k,t)$ and $X_{n,2}(k,t)$.

The norm of the pseudo-coherence-vector reflects the input signal quality at each compact array location. In the literature, this feature is only calculated for the dual compact arrays and not single microphones. This feature has been applied for distinguishing between high quality input nodes and highly distorted nodes where all the nodes have the same structure. Assuming that all th endes are of the same structure is the limitation of [22].

2.8.2 MFCC

The Mel Frequency Cepstral Coefficients (MFCCs) feature is a cepstral feature which has been successfully applied for speaker profiling [50] and emotion detection. This feature has proven to give very good results in the context of (anechoic) speech/music/noise classification tasks and constitute a very compact representation of the signals. It is also applied for microphone clustering [5]. It is important to note that MFCC has been applied within supervised and unsupervised machine learning techniques.

In order to calculate the MFCC coefficients the speech sample is broken down into frames of length such that the information in a frame does not vary statistically (e.g. 20ms). For each short time frame,a periodogram estimate of the power spectrum is calculated as :

$$P_i(k) = \frac{1}{N}|X_i(k)|^2 \qquad 2\text{-}12$$

$$X_i(k) = \sum_{n=1}^{N} x_i(n)h(n)e^{-i2\pi kn/N}, \quad 1 \leq k < K \qquad 2\text{-}13$$

where P_i is the power spectrum X_i is the length K discrete Fourier transform of $x_i(n)$ and *i* is the frame index. The Mel filter bank is applied to the power spectra and the energy in each filter is added. $h(n)$ is an *N* sample long analysis window.

MFCC as a cepstral feature has been applied to speech signals, noise, music and RIRs [51]. Although it has been applied to microphone clustering but it does not contain any information about source to microphone distance [52] [53].

2.8.3 LP CMRARE

The Legendre Polynomial-based Cepstral Modulation RAtio REgression (LP-CMRARE) is a cepstral feature for compact representation of the (anechoic) speech , noise and music signals for signal classification and microphone clustering. It is important to note that LP CMRARE has been applied within supervised and unsupervised machine learning techniques. [5]

To obtain the LP-CMRARE features, the spectrum is transformed into the cepstral domain. In order to analyse the spectro-temporal changes of the cepstrum a sliding window Discrete Fourier Transform (DFT) is applied as :

$$\hat{X}_c = \sum_{m=0}^{M-1} X_c e^{-i2\pi vm/M} \quad 2\text{-}14$$

where X_c is the cepstral domain signal and *v* represents the modulation frequency bin index. The magnitude of the modulation spectrum is averaged over all windows as:

$$\bar{X}_c = \frac{1}{C_T}\sum_{c=0}^{C_T-1}\left|\hat{X}_c\right| \quad 2\text{-}15$$

LP CMRARE has been used for speech, noise and music signals and it has been successful for speaker recognition and discrimination but it does not contain any information about the signal quality, reverberation level and source to microphone distance.

2.8.4 Time of Arrival

Time of Arrival (TOA) or Time of Flight (TOF) information if available or retrievable can accurately calibrate microphone arrays [54] which can be useful for microphone clustering, clustered dereverberation and source targeting applications however in the target scenarios of this research the nodes are independent and do not communicate and the source's start and stop times are assumed unknown. TOA can be calculated if the microphones are synchronised and the source start time is known which are not practical assumptions for ad-hoc arrays and spontaneous meetings.

In the context of ad-hoc arrays TOA information derived from RIRS can be applied for microphone clustering however this method requires full knowledge of RIRs which might not be available for all scenarios. TOA at microphone *m* location (r_m) from source location (r_s) is mathematically defined as:

$$TOA_m = \frac{|r_s - r_m|}{c} \quad 2\text{-}16$$

2.8.5 Time Difference of Arrival

The Time Difference of Arrival (TDOA) is applied in the literature for source localisation [55], microphone localisation [56] and joint localisation of the source and the microphones. Although TDOA overcomes the limitation of unknown start time (t=0 timestamp) the main issue with TDOA feature for such applications is that it requires communication among the nodes, which is not available in many recording devices and scenarios. Another challenge that arises with ad-hoc arrays

due to their unknown geometrical configuration and inconsistency of the devices is that the nodes are usually not synchronised and might use different frequency rates. Under certain circumstances the calculation of TDOA is straightforward but for unsynchronised devices without inter-node communication mathematically intensive solutions are suggested [6], which are not recommended for real time applications.

The problem of sensor and source joint localisation using time-difference of arrivals (TDOAs) of an ad-hoc array is investigated in the literature. The major challenge is that the TDOAs contain unknown time offsets between asynchronous sensors but it is shown that this issue can be addresses by further mathematical processing [6], [57] , [58].

TDOA information is successfully used for localisation applications but in terms of signal quality and dereverberation TDOA information are not helpful.

2.8.6 Speech Energy

Energy is the simplest feature to calculate/estimate for both full utterance and frame based analysis however a few critical issues confine it's applications as discussed in the literature [31], where an energy-based method for source and microphone localisation is proposed for an ad hoc network of microphones. The target scenario is a meeting that sources (participants) and the microphones (laptops) are collocated. Compared with traditional sound source localisation approaches based on time of flight, this technique does not require accurate synchronisation, and it does not require each laptop to emit special signals.

In a multi-channel recording scenario, the energy of a signal can be calculated independently of other channels, signal synchronisation and time alignment are not required. Energy levels can be compared and if the microphones have the same gain (which is not always verifiable), the node with the highest energy level is the closest node to the active source during that time frame or utterance.

$$E\big(x(n)\big) = \langle x(n), x(n)\rangle = \sum_{i=1}^{\infty} x^2(i) \qquad \text{2-17}$$

For the full utterance analysis and calculated over a short time frame of length (L) in time domain as:

$$E_L = \sum_{i=1}^{L} x^2(i) \qquad \text{2-18}$$

Energy can also be calculated in the time-frequency domain. The main limitation of the energy feature is that it is not possible to control the microphones gains or verify if they all have the same gain. Under special circumstances (i.e. microphone and source being collocated) it is possible to overcome this limitation and use the energy level for microphone localisation and clustering.

2.8.7 Kurtosis of linear prediction residual signal

The kurtosis of the Linear Prediction (LP) residual signal was proposed as a discriminative feature for target speech discrimination in teleconferencing systems where interference is a common issue that decreases the teleconferencing experience significantly. [25] Conventional methods of voice activity detection (VAD) utilise the location cues of sound sources to distinguish desired from undesired speech and utilise multiple microphones to estimate the directions of sound sources. Research in [25] has proposed a novel source discrimination method that exploits only one microphone to discriminate desired from undesired speech assuming that the desired source is located closer to the microphone than the interfering source. Kurtosis of the linear prediction residual signals is applied as the discriminative feature in the research by [25] as their observations show that this feature has an inverse relationship with source to microphone distance in a variety of room types in terms of sizes and the reverberation times including conference rooms, sound proof room, elevator hall and laboratory. The experimental results revealed that the proposed method could distinguish close-talking speech from distant-talking speech within a 10% equal error rate (EER) in ordinary reverberant environments. The main drawback of this feature and the proposed method is the dependency on a predefined threshold. As kurtosis values are calculated based on the residual signals obtaining the prediction coefficients is the first step. For the recorded signal $x_m(n)$ from (1), the predicted signal $\hat{x}_m(t)$ obtained by the LPC method is:

$$\hat{x}_m(t) = \sum_{j=1}^{J} a_j \, x_m(t-j) \qquad 2\text{-}19$$

where *J* is the LPC prediction order and the LPC prediction coefficients (a_j) can be calculated by any conventional method for each channel. The resulting LPC residual (error) signal is

$$e_m(t) = x_m(t) - \hat{x}_m(t) \quad \text{2-20}$$

The kurtosis values for each frame or utterance can be obtained by:

$$k_m(t) = \frac{E\{e_m^4(t)\}}{E^2\{e_m^2(t)\}} - 3 \quad \text{2-21}$$

The kurtosis value can be calculated in both utterance mode and frame based mode and the discriminative feature for each node with more than one channel is calculated by averaging the kurtosis values within each node. Kurtosis can be calculated and applied as a discriminative feature when the source is a speech signal and the nodes located closer to the source have higher kurtosis values [25]. The disadvantage of this feature is that it can only be applied to speech signal as it is based on LP coding and cannot be applied to noise, RIRs or other signal types. Another limitation of the proposed method by [25] is the dependency on the predefined threshold which requires training for each recording setup and room.

2.8.8 The clarity feature (C_{50})

The C_{50} or Clarity measurement is the ratio of early to late reverberation expressed in dB. This measure is higher when the microphone to sources distance is relatively small and the recorded signal by the microphone is dominated by the direct path signal [59] [60]. In contrast it is lower when microphone to source distance is relatively large and the second and third order reverberations are no longer negligible. It is shown that the C_{50} has an inverse relationship to the microphone to source distances and for calculating C_{50} the clean signal is not required (in contrast to the Direct to Reverberation ratio (DRR)). The C_{50} is defined in as:

$$h(t) = h_{direct}(t) + h_{early}(t) + h_{late}(t) \quad \text{2-22}$$

$$C_{50} = 10 \times \log\left(\frac{E_{direct}+E_{early}}{E_{late}}\right) \quad \text{2-23}$$

with $E_{Direct} = a_1\delta(n)$, $E_{early} = \sum_0^{t=50ms} h(n)$, and $E_{late} = \sum_{50ms}^{\infty} h(n)$ and n is the frame index. Using (2), C_{50} can be calculated for each RIR without synchronisation by:

$$C_{50} = 10 \times \log\left(\frac{\sum_0^{t=50ms} h(t)}{\sum_{50ms}^{\infty} h(t)}\right) \quad \text{2-24}$$

The clarity feature is robust against fluctuations of the source energy level and can reliably be used when there are sources with different levels of energy.

$$C_{50} = 10 \times \log\left(\frac{\propto.E_{direct}+\propto.E_{early}}{\propto.E_{late}}\right) \quad 2\text{-}25$$

$$= 10 \times \log\left(\frac{\propto\left(E_{direct}+E_{early}\right)}{\propto E_{late}}\right) \quad 2\text{-}26$$

$$= 10 \times \log\left(\frac{E_{direct}+E_{early}}{E_{late}}\right) \quad 2\text{-}27$$

The limitation of C_{50} is that it requires the full length RIRs and hence cannot be applied to real time applications.

2.8.9 Magnitude square Coherence (MSC)

Reverberation and interference recorded by each microphone are functions of its location in the room and as the microphones of each node are not exactly collocated they record slightly different echoes and interferences [61], [62], [63]. When microphone's signals are distorted by reverberation and interference they become statistically more independent and they will have lower intra MSC values calculated by:

$$C_{ij}(f) = \frac{|\varphi_{m1m2}(f)|^2}{\varphi_{m1m1}(f)\,\varphi_{m2m2}(f)} \quad 2\text{-}28$$

where $\varphi_{m1m1}(f)$ and $\varphi_{m1m2}(f)$ are auto and cross power spectral densities between microphone m_1 and m_2 respectively from (1). If nodes in the ad-hoc array contain dual-channel microphone systems, it is possible to discriminate highly distorted nodes (located far from the active sources) and the node's signals predominated by the speech signals (located closer to one of the sources). This fact about MSC is utilised here as a distance cue to estimate the distances between the active sources and the nodes. "*The idea is that when the magnitude [square coherence] is close to one, the speech signal is present and dominant and when it is close to zero, the interfering signal is dominant.*" [61].

2.8.10 Room impulse responses

RIRs as they contain echo time delays and attenuation information, can be considered for feature extraction [63], [64], [65].

In the general form of the problem let *M* microphones be distributed in a room of unknown geometry and labelled $m_1, m_2, \dots m_j \dots, m_M$, which record *N* sources $s_1, s_2, \dots s_k \dots, s_N$. The sound recorded by each of these microphones is the convolution of the acoustic RIR corresponding to its location in the room and the source signal. It is assumed that all microphones are synchronized and the lengths of the RIRs are equal. These RIR sequences contain impulses received from direct paths between sources and microphones and reflections from the walls, ceiling and floor and can be modelled mathematically as a train of impulses as:

$$\hat{h}_{sk,mj}(n) = \sum_l a_{sk,mj}(l)\delta\left(n - d_{sk,mj}(l)\right) + N(n) \qquad 2\text{-}29$$

where $d_{sk,mj}(l)$ represents the propagation delay from source and reflectors to the microphone m_j when source *k* is active, $a_{sk,mj}(l)$ represents the amplitudes of each impulse corresponding to an echo and *l*=0 to *L* represents the number of impulses. $N(n)$ represents the noise in the general form. In practice, RIRs can be estimated by techniques such as recording a sine-sweep covering a range of frequencies (e.g. 20Hz to 20 kHz) and digitally sampling this signal as a pre-recording phase or they can be extracted from speech signals by the proposed method in [66].

2.8.11 Direct to reverberant ratio

Reverberation affects the speech signal quality and intelligibility in the reverberant environment. Direct to Reverberant Ratio (DRR) is a function of reverberation and the distance from the source [60]. The microphones located close to the source have higher signal quality and The DRR. It is also shown that DRR can be estimated accurately [67]. DRR for microphone *m* in the array is defined as

$$DRR = \frac{\sum h_{d,m}(n) * s(n)}{\sum h_{r,m}(n) * s(n)} = \frac{\sum h_{d,m}(n)}{\sum h_{r,m}(n)} \qquad 2\text{-}30$$

where $h_{d,m}(n)$ and $h_{r,m}(n)$ are the direct and the reverberant components of the RIR. As it is observed from the equation the DRR is independent of the source signal and the energy level.

2.9 Chapter summary and conclusion

In this chapter the ad-hoc arrays and their applications have been reviewed based on the most recent literature. As was mentioned, the final objective is to develop a multichannel dereverberation method for ad-hoc arrays however microphone clustering, source counting, targeting and localisation can help the analysis of the acoustic scene as a prior stage to the dereverberation task. As the machine learning techniques and the extracted features from the multi-channel and multi-node recordings are important parts of microphone clustering they have been separately reviewed in the literature as well. It is important to conclude that each feature is suitable for certain applications and each machine learning technique can be helpful

in a specific task and hence it is not possible to come up with one feature and one technique which can be applied in general to applications. In the next chapters, the state-of-the-art and the proposed features extracted from the ad-hoc array recordings will be applied to microphone clustering prior to applying the proposed multi-channel dereverberation method.

3 Microphone clustering

3.1 Introduction

This chapter investigates the formation of ad-hoc microphone arrays for the purpose of recording and processing multiple sound sources by clustering microphones spatially distributed within a room. In the context of ad-hoc microphones, clustering is important as microphones located close to a source record the signal with higher quality [5]. On the other hand the microphones located far from the source are usually highly distorted by noise, reverberation and interfering sources and it is suggested to exclude them from the recording and post-recording process [22]. In other words, utilising all the available nodes and microphones for applications such as dereverberation and source localisation is not the optimal approach as the higher number of channels usually means more processing load and also including highly distorted microphones only decreases the overall system performance [5], [68].

This hypothesis is investigated in this thesis for the specific task of dereverberation and analysis of acoustic scenes by investigating several ad-hoc scenarios and evaluating the recording quality and speech enhancement performance by the conventional measurements. A novel codebook-based unsupervised method for cluster formation using features derived from the Room Impulse Responses (RIRs) corresponding to each microphone is proposed and compared with baseline clustering and classification methods.

The estimated coherence feature [69] is also proposed in this chapter as a novel feature for microphone clustering where all nodes have the same structure, which is an acceptable assumption for most conference tables with a built-in microphone at each seat location. Based on this feature a novel clustering method is proposed which overcomes the limitations of the state of the art clustering methods such as prior knowledge of the number of clusters to form [5] and the training phase. The proposed clustering methods in this chapter obtain high clustering accuracy with less limiting constraints and required prior information.

The objectives of this chapter are:

- Extracting microphone clustering discriminative features from RIRs where the RIR recordings are available or retrievable.
- Extracting microphone clustering discriminative features from speech signals or situations where RIRs are unknown or cannot be reliably estimated.
- Proposing a microphone clustering method that does not require the prior knowledge of the exact number of clusters to form (limitation of [5]).

The contributions of this chapter primarily overcome the limitations of the previous research

- Clustering the ad-hoc microphones without requiring the prior knowledge of the number of sources or pre-assigning the number of clusters.
- Proposing the inter-microphone coherence based clustering method for speech signals without using standard clustering techniques.
- Proposing new clustering evaluation measurements. (average intra cluster distance and Magnitude square coherence for more than two signals)
- Proposing a systematic microphone clustering evaluation scheme for ad-hoc scenarios.

Publications arising from this chapter include

- S. Pasha, Y. X. Zou & C. Ritz, "Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses," in Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on, 2015, pp. 84-88.
- S. Pasha & C. H. Ritz, "Clustered multi-channel dereverberation for ad hoc microphone arrays," in Proceedings of APSIPA Annual Summit and Conference 2015, 2015, pp. 274-278.

3.2 Motivation and Problem formulation

Some recent recording methods [70], [71] using ad-hoc microphone arrays utilise partial information to help guide applications such as sound source separation and classification. These informed signal processing approaches are more effective compared to blind approaches for the analysis of complex acoustic scenes and sound source separation. As an example, in [5] a novel method for exploiting relative microphone and source spatial locations was introduced and evaluated for microphone clustering and signal classification. This method relies on accurate knowledge of the total number of sources as well as the total number of clusters to form. In [72] A maximum likelihood approach using time of arrival measurements of short calibration pulses is proposed to solve this self-localisation problem.

In [41] the authors showed that rather than using all microphones in a room, forming ad-hoc microphone arrays using small clusters of microphones each located close to one source can yield better separation quality. The approach removes microphones from the ad-hoc array that are located far from target sources, which may be corrupted by other sources and hence have a low target-to-interference signal ratio. Such an approach also reduces the beamforming steering error and is based on measuring the coherence between microphones in noise-only periods as well as the relative Time Difference of Arrival (TDOA) between neighbouring microphones during speech periods. Their approach assumed small subsets of microphones were located close to desired speakers. Herein in a general scenario of ad-hoc arrays the main goal is to propose a novel codebook microphone clustering method based on time delay and gain information derived from microphones at unknown locations

Microphone clustering is a way to group microphones spatially close to each other for applications such as beamforming and source separation. Microphone clustering does not need the exact localisation of all nodes (different to [73], [74]) and it is only based on the similarities of the features derived from the recorded signals or RIRs [68], [75].

As explained in Chapter 1, an ad-hoc microphone array is formed from sets of microphones randomly positioned in a room and can be used to record multiple spatially distributed sound sources with a better and more flexible spatial coverage compared with a single microphone array located at one position.

Assuming that there are *M* microphones (or nodes) in an ad-hoc array and based on their relative distances to the source they receive a unique version of the source clean signal, the objective is to choose a subset of nodes such that applying them exclusively for a task such as dereverberation or recording yields the highest output quality in terms of the conventional measurements of each task.

Figure 3-1: Examples of microphone clusters

In Figure 3-1, *X* is the matrix of the ad-hoc channel recordings. It is shown that each source can be recorded with a higher intelligibility if only microphones close to the target source are utilised and the other channels which are highly distorted by interference from other speakers are removed from the array [5].

The goal of signal clustering is to assign objects to groups with small intra-group differences and large inter-group differences (3-2). Assuming that $X = \{x_1, x_2, \dots x_M\}$ is the set of recorded signals by all the *M* channels in the array, the clustering objective is to form the subsets $X_c \subset X$ that minimises the following cost function *J*.

$$X = \{x_1, x_2, \dots x_M\} \quad 3\text{-}1$$

$$J = \sum_{j=1}^{N} \sum_{m \in X_c} |x_m - \mu_j|^2 \quad 3\text{-}2$$

where *N* is the number of clusters to form and μ_j is each cluster centroid. [33] and improves a certain criterion for each application. (3-3)

Assuming that function F is a performance measurement specifical to a particular application such as SNR for noise cancellation, the clustering criterion can be modelled mathematically as:

$$F(X_c) > F(X) \qquad 3\text{-}3$$

Which means utilising the microphones within the chosen cluster (i.e. X_c) yields better results compared to the blind use of all M nodes in X.

As the aim is to cluster microphones, raw signals cannot be exploited as the process will be inefficient and time consuming [5]. As an alternative, discriminative features should be chosen and derived from the raw signals that discriminate microphones according to their spatial location (Figure 3-2).

Figure 3-2: Unsupervised microphone clustering process

In this research the required signals and RIRs are simulated under different circumstances in terms of RT_{60}, SNR, the number of simultaneously active speakers and the source to microphone distances. Each recording simulation is labelled based on the recording attributes.

3.3 Discriminative features

Rather than performing clustering on the recorded audio signals or RIRs directly the data is typically transformed to a reduced parametric representation which is referred to as feature extraction. In case of clustering into groups without training data, unsupervised methods can be used to generate unlabelled clusters of objects [5].

In this section novel discriminative features for microphone clustering derived from RIR recordings and speech signals are described and the process and requirements of their extraction are discussed.

3.3.1 Discriminative features derived from RIR recordings

The base-line feature for microphone clustering is the Time Difference of Arrival (TDOA) which is based on the difference in the arrival time of the direct path signal at two microphones and is generally calculated using cross correlation-based methods [41], [76], [77]. These methods suffer from room reverberation and hence

techniques to suppress the effects of reverberation on TDOA estimation accuracy are often required [5], however researchers have recently shown that it is possible to make use of reverberation for extracting distance cues [25]. TDOA is also not reliable when the room reverberation time is relatively large which causes TDOA outliers [78]. It is noteworthy that the main constraint for defining a discriminative feature is the feasibility of the feature extraction in the ad-hoc scenarios where microphones might freely move at any time and information such as source location and the start time of the speech signal are unknown. Features such as TOA suffer from dependency on source start time and time alignment of the microphones which make it less practical in ad-hoc scenarios.

Herein a novel feature is derived from RIR recordings rather than recorded speech signals that does not require complex calculations of noise coherence and inter-microphone cross correlations. This method does not require the information about the sources, room and microphone array and is solely based on similarity and dissimilarity of the extracted features from RIRs. In contrast to [31], where reverberation was causing error and was needed to be supressed, the proposed RIR clustering method exploits reverberation to cluster the microphones [63]. This is motivated by the approaches in [64], where similarly they estimate the echoes as the peaks in the RIR recordings. These are then used within alternative clustering algorithms for forming the ad-hoc microphone arrays. Discrimination of symmetric clusters by using two asynchronous sources located at two different locations is the novelty of this proposed method.

In the general form of the problem let *M* microphones be distributed in a room of unknown geometry and labelled $m_1, m_2, \dots m_j \dots, m_M$, which record *N* sources $s_1, s_2, \dots s_k \dots, s_N$. The sound recorded by each of these microphones is the convolution of the acoustic RIR corresponding to it's location in the room and the source signal. It is assumed that all microphones are synchronised and the lengths of the RIRs are equal. These impulse sequences contain impulses received from direct paths between sources and microphones and reflections from the walls, ceiling and floor and can be modelled mathematically as a train of impulses as :

$$h_{sk,mj}(n) = \sum_{l=0}^{L} a_{sk,mj}(l)\delta\left(n - d_{sk,mj}(l)\right) + N(n) \qquad 3\text{-}4$$

where $d_{sk,mj}(l)$ represents the propagation delay from source *k* and reflectors to the microphone *j* when only source *k* is active, $a_{sk,mj}(l)$ represents the amplitudes of each impulse corresponding to an echo and *l*=0 to *L* represents the number of impulses. The number of counted impulses depends on the room RT_{60} and the room dimensions. $N(n)$ represents the noise in the general form. In practice, RIRs can be estimated by techniques such as recording a sine-sweep covering a range of frequencies (i.e. 20Hz to 20 kHz) and digitally sampling this signal as a pre-recording phase [79] or they can be extracted from speech signals by the proposed method in [51]. Assuming that the RIR recordings are available or estimated, the RIR of length *L+1* at microphone *j* when source *k* is active can be represented as:

$$\hat{h}_{sk,mj} = [h_{sk,mj}(0), \dots, h_{sk,mj}(L)] \quad 3\text{-}5$$

In a general scenario of *M* microphones and *N* sources, a matrix of $\hat{h}_{sk,mj}$'s can be constructed as :

$$H = \begin{bmatrix} \hat{h}_{s1,m1} & \dots & \hat{h}_{sN,m1} \\ \vdots & \ddots & \vdots \\ \hat{h}_{s1,mM} & \dots & \hat{h}_{sN,mM} \end{bmatrix} \quad 3\text{-}6$$

The peak sample numbers representing the propagation delays [80], $d_{sk,mj}(l)$, corresponding to the peaks of $\hat{h}_{sk,mj}$ of (2) are represented here by the vector of delays, $\hat{d}_{(sk,mj)}(l) = \left[d_{sk,mj}(0), d_{sk,mj}(1), \dots, d_{sk,mj}(L)\right]$, where $d_{sk,mj}(0)$ is the arrival time from the source *k* to the microphone m_j for the direct path signal and $d_{sk,mj}(1), \dots, d_{sk,mj}(L)$ represent the delays for the first *L* echoes. The delay matrix for microphone m_j can be constructed as D_j, where *j*=1 to *M*:

$$D_j = \begin{bmatrix} d_{s1,mj}(0) & \dots & d_{sN,mj}(0) \\ \vdots & \ddots & \vdots \\ d_{s1,mj}(L) & \dots & d_{sN,mj}(L) \end{bmatrix} \quad 3\text{-}7$$

The magnitudes of the direct path impulses and *L* echoes received from *N* sources to microphone m_j from the array can be represented as A_j:

$$A_j = \begin{bmatrix} \left|\hat{h}_{s1,mj}(0)\right| & \dots & \left|\hat{h}_{sN,mj}(0)\right| \\ \vdots & \ddots & \vdots \\ \left|h_{s1,mj}(L)\right| & \dots & \left|\hat{h}_{sN,mj}(L)\right| \end{bmatrix} \quad 3\text{-}8$$

$\left|\hat{h}_{sk,mj}(0)\right| = a_{sk,mj}(0)$ and the extension to more sources and microphones is straightforward.

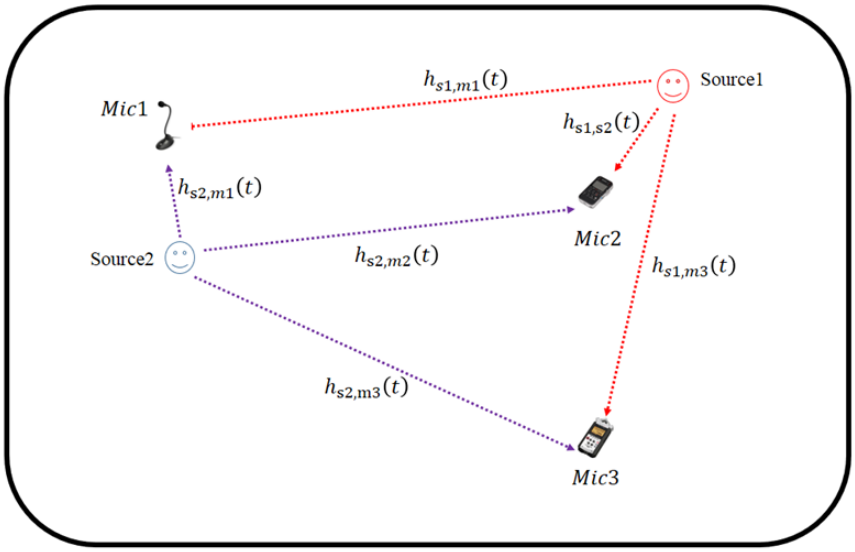


Figure 3-3: Two speech sources being recorded by three ad-hoc microphones

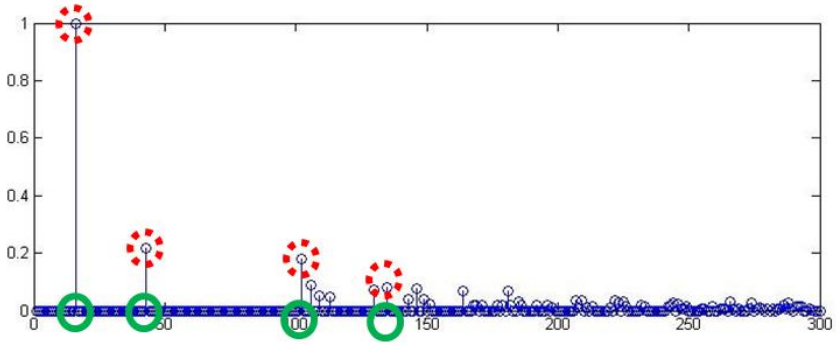


Figure 3-4: RIR time delays and peaks

A set of extracted features from one microphone RIR for two echoes can be represented as

$$K_j = \begin{bmatrix} d_{s,mj}(0) \\ d_{s,mj}(1) \\ d_{s,mj}(2) \\ \hat{h}_{s,mj}(0) \\ \hat{h}_{s,mj}(1) \\ \hat{h}_{s,mj}(2) \end{bmatrix} \quad 3\text{-}9$$

This vector is calculated for all the j values $1 \leq j \leq M$, and the obtained feature vectors are clustered by the clustering methods. The distance function (e.g. Euclidian distance) is applied to these vectors in order to measure their similarities:

$$\|K_1 - K_2\| = \sqrt[2]{(K_1 - K_{2)}{}^2} \quad 3\text{-}10$$

The feature matrix for the array is represented by

$$\mathbf{K} = \begin{bmatrix} d_{s,m1}(0) & \cdots & d_{s,mM}(0) \\ d_{s,m1}(1) & \cdots & d_{s,mM}(1) \\ d_{s,m1}(2) & \cdots & d_{s,mM}(2) \\ \hat{h}_{s,m1}(0) & \cdots & \hat{h}_{s,mM}(0) \\ \hat{h}_{s,m1}(1) & \cdots & \hat{h}_{s,mM}(1) \\ \hat{h}_{s,m1}(2) & \cdots & \hat{h}_{s,mM}(2) \end{bmatrix}. \quad 3\text{-}11$$

3.3.2 Discriminative features derived from speech signals

The following proposed features are extracted by utilising the speech signals for a microphone clustering application using the baseline and the proposed code-book methods.

3.3.2.1 The kurtosis of LP residual signal

Microphones located close to each other receive similar levels of reverberation and microphones far from each other (e.g. one microphone close to the source and the other close to a wall) have different levels of reverberation in their recorded signals. As the kurtosis of the LP residual signal is a function of reverberation level [25] this feature is applied to cluster microphones. In a sample echoic recording room a source is recorded by a grid of microphones across the *x* and *y* axis. The grid step size is 0.5m and all the microphones and the source are at the same height (2m). It is observed that the Kurtosis of the LP residual signal drops as the source to microphone distance increases (Figure 3-5) and hence can be applied as a microphone clustering feature to discriminate microphones located close to the source and the microphone far from the source [68].

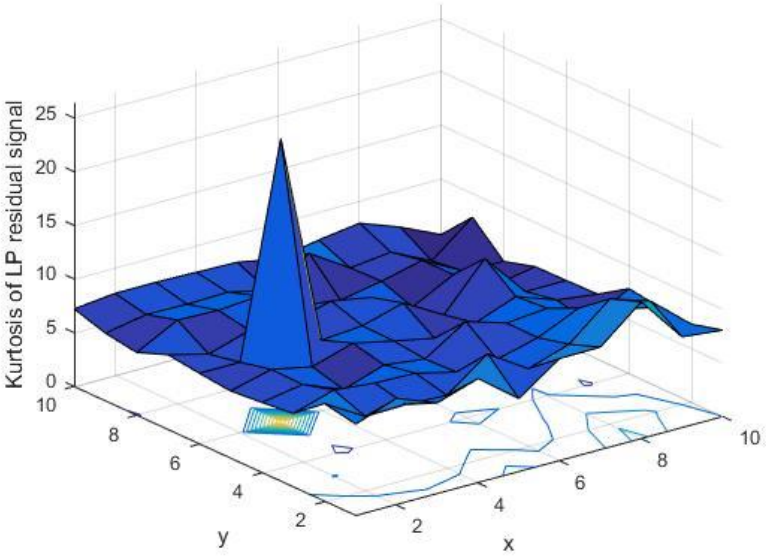


Figure 3-5: Kurtosis values for a source located at (3,6,2) in a 10m by 10m by 3m room. fs=16k, $RT_{60} = 600ms$, calculated for 32ms frames and averaged across one second of speech signal.

According to the results, it is possible to cluster the microphones into two categories based on their locations in the room. 1) Anechoic clean signal (peak area) 3) highly reverberated area (flat area)

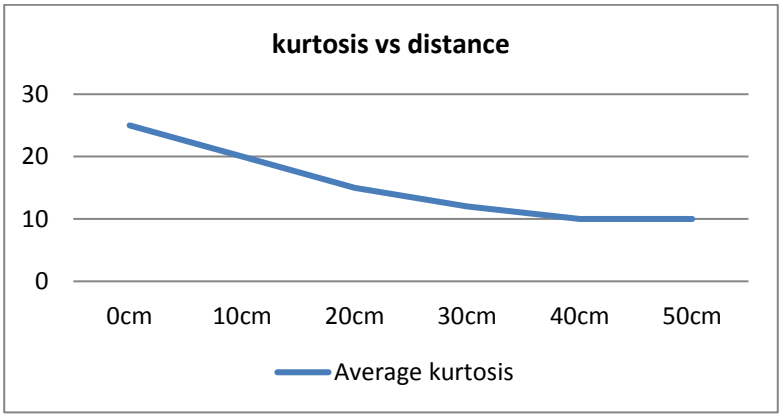


Figure 3-6: Kurtosis vs distance

$RT_{60} = 600ms$, full utterance (3s)

The first data-point (distance =0cm) represents the source clean signal kurtosis value.

This graph and similar results from [25] show that the kurtosis of the LP residual signal has an inverse relationship to the microphone to source distance.

3.3.2.2 Magnitude Coherence Square (MSC)

The relationship between the MSC and the source to microphone distance is investigated in [81] and it is concluded that the MSC can localise sources. In this thesis this feature is applied as a clustering feature to cluster microphones based on their distances to the source (Figure 3-7). The limitation of this feature is that it requires dual channel nodes and all the nodes should be of the same structure and inter channel distance. The advantage of the coherence feature is that it can be estimated based on short frames of the speech signals [29].

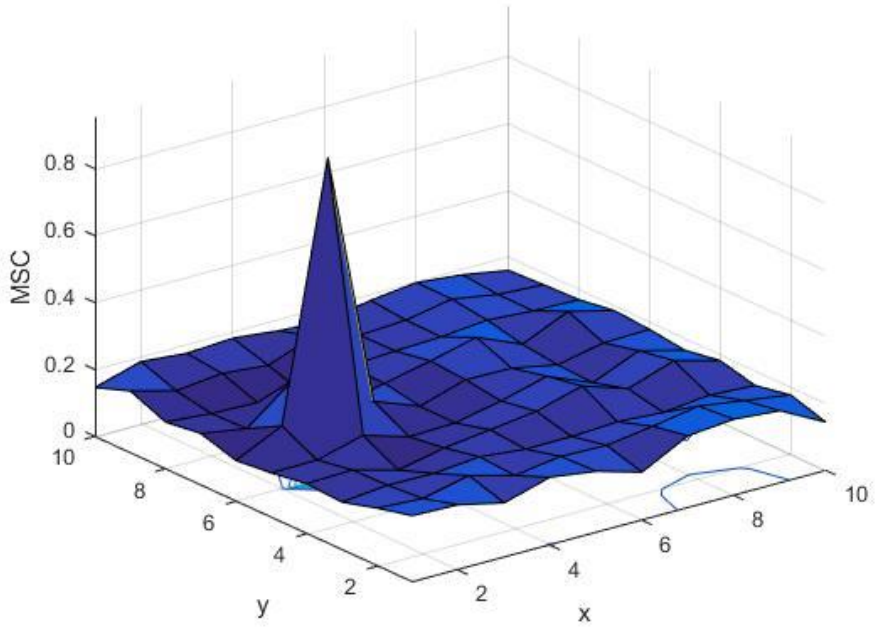


Figure 3-7: MSC values calculated across the room (source at 3m,6m,2m)

3.4 Proposed clustering methods

3.4.1 Code-book based methods

In a randomly distributed microphone array, the objective is to extract and compare microphones features that are used to cluster microphones. All the proposed features in the previous section can be applied as discriminative features within the proposed code-book based clustering algorithm. The process starts with generating a code-book of 5 centre points features across the room. Unseen microphones signals are then processed and the discriminative features are extracted. The extracted set of

features from each unseen microphone is then compared with the code-book centre points to find the best cluster for this microphone.

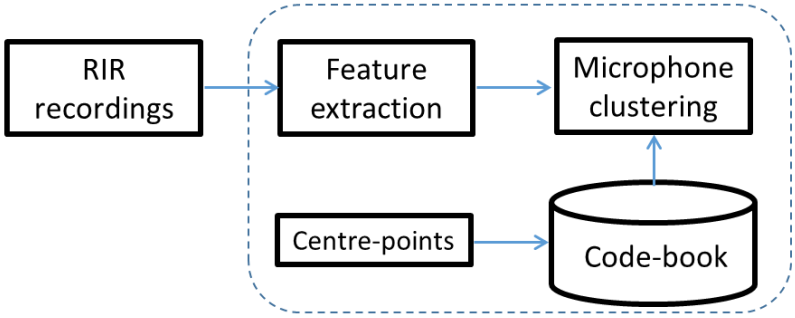


Figure 3-8: Code-book based clustering algorithm

The assumption is that while recording RIRs or extracting other speech features, only one source is active (for both code-book generation and microphone clustering phase). The proposed codebook based clustering method [75] is summarised in Table 3-1.

Table 3.1: Code-book based clustering method
Input: RIR of each microphone, Codebook ***Output: Clustered microphones based on spatial locations***
1. Choose P centre points in the room, obtain arrival time and echo delays and assign a zone label to each centre point (Codebook generation)
2. For each randomly distributed microphones in the microphone array:
A. Obtain the recorded RIR ***B. Derive discriminative features***
C. Compare each microphone's feature vector with the generated codebook
D. Assign the closest centre point's zone to the microphone ***3. The number of assigned zones labels show the number of clusters***

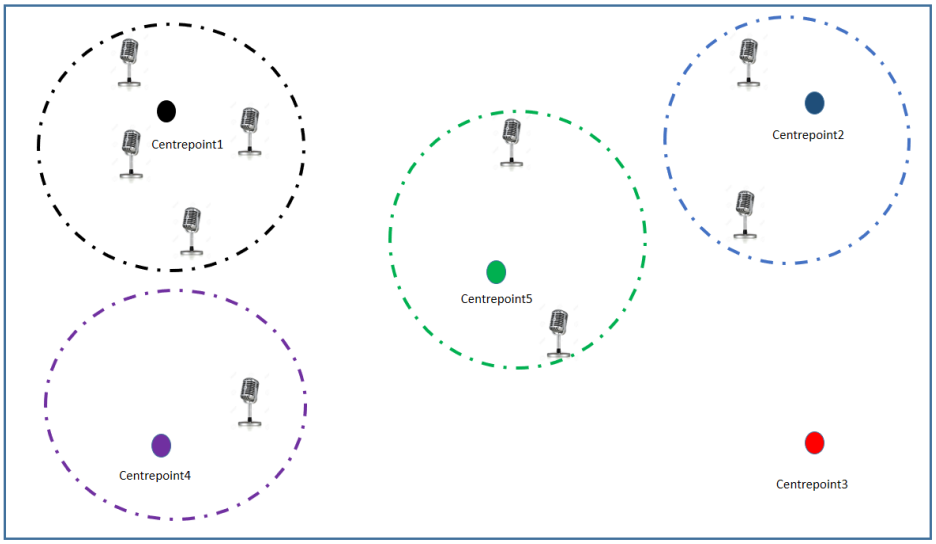


Figure 3-9: Centre points and formed clusters

In this approach it is assumed that *P* reference RIRs or speech signal centre points are known (or have been previously recorded) within the room (referred here also as centre point of a cluster). These centre points can be chosen blindly with a uniform distribution within the room however if there is prior information about possible locations of sources and microphones they can be chosen in an informed manner. For *M* microphones the goal is to assign each data point for each microphone at an unknown position to the closest centre point based on similarities between features. Similar to Vector Quantization (VQ), microphones are clustered based on the closest matching centre points estimated by the Euclidian distance measure:

$$d_{i,p} = \sqrt[2]{(f_i - f_p)^2}$$

where $d_{i,p}$ is the Euclidian distance between the microphone i and centre point p ($1<p<P$) and $\mathbf{f}_i$ and $\mathbf{f}_p$ represent feature vectors from microphone i and centrepoint p respectively.

The main issue with clustering microphones in symmetrical rooms and setups is that microphones located far from each other might get clustered together due to the symmetry. Clustering symmetrically positioned microphone, clusters together is also addressed by using two asynchronous sources at different positions and concatenating the feature vectors. The symmetry issue is depicted in Figure 3-10 where two microphones at two facing corners of the room have similar RIRs and discriminative features.

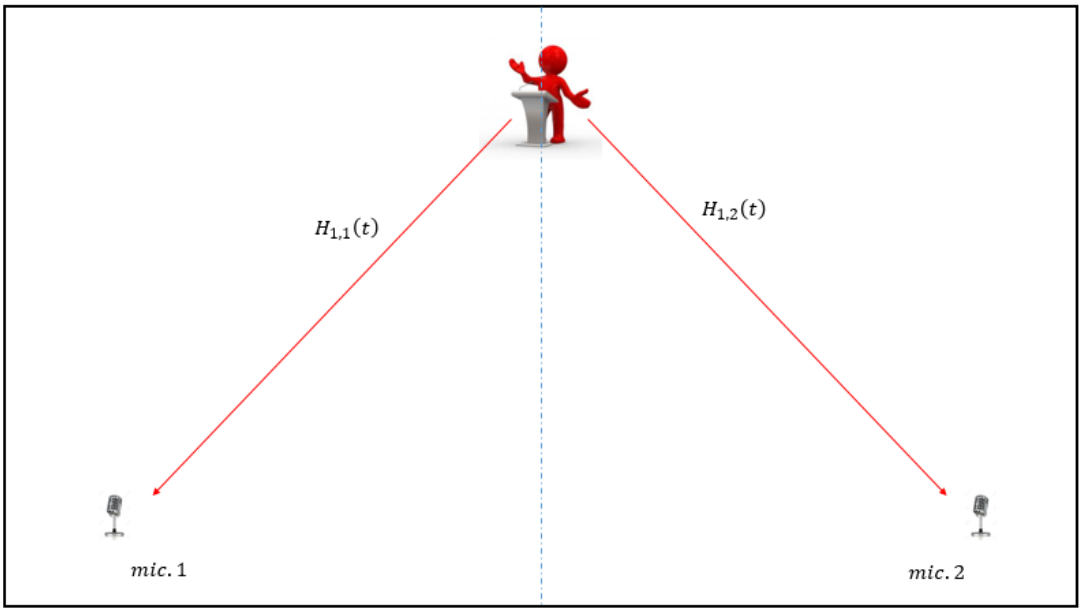


Figure 3-10: Symmetry issue for clustering microphones

The advantages of the code-book based clustering method are:

- Forming a flexible number of clusters (between 1 and the number of centre points) whereas baseline clustering methods (e.g. Kmeans) require predefined number of the clusters
- Clustering microphones based on their features similarities to the center points without training

Limitations of the code-book based clustering method are:

- Requiring features derived from the RIRs or the speech signals at certain points of the room (Centre points) which might not be practical for all setups and scenarios.

3.4.2 Coherence based clustering method

Assuming there are *M* microphones (nodes) randomly distributed in a room, the objective is to cluster them into a flexible number of clusters based on the coherence of their signals (estimated/calculated over short time frames). This proposed clustering method is based on this observation that microphones that record similar

signals have higher coherence compared to microphones located far from each other. In other words, signal coherence is a function of microphone separation distances.

The coherence between two microphones' signals ($m1$ and $m2$) is mathematically defined as:

$$C_{m1m2}(f) = \frac{|\varphi_{m1m2}(f)|^2}{\varphi_{m1m1}(f)\,\varphi_{m2m2}(f)} \quad 3\text{-}12$$

$$\varphi_{m1m2}(f) = F\left(\sum_{\tau=-\infty}^{\infty} R_{m1m2}(t)\right) \quad 3\text{-}13$$

where $\varphi_{m1m2}(f)$ and $R_{m1m2}(t)$ represent the cross power spectral density and the cross correlation functions respectively. F indicates the Fourier transform.

Coherence function obtains its maximum value (the maximum value of the coherence function is one) when two signals are identical and therefore:

$$\varphi_{m1m1}(f) = \varphi_{m1m2}(f) \quad 3\text{-}14$$

$$\varphi_{m2m2}(f) = \varphi_{m1m2}(f) \quad 3\text{-}15$$

and:

$$C_{m1m1}(f) = \frac{|\varphi_{m1m1}(f)|^2}{\varphi_{m1m1}(f)\,\varphi_{m1m1}(f)} = 1 \quad 3\text{-}16$$

The two microphones can only have identical signals if they are collocated and if they are located far from each other the value of the coherence will decrease as a function of the distance between them, interfering sources, noise and reverberation time.

For a scenario that microphones are located at different distances and the source is at the center of the room, the relationship between the coherence and microphones distances are depicted in Figure 3-11.

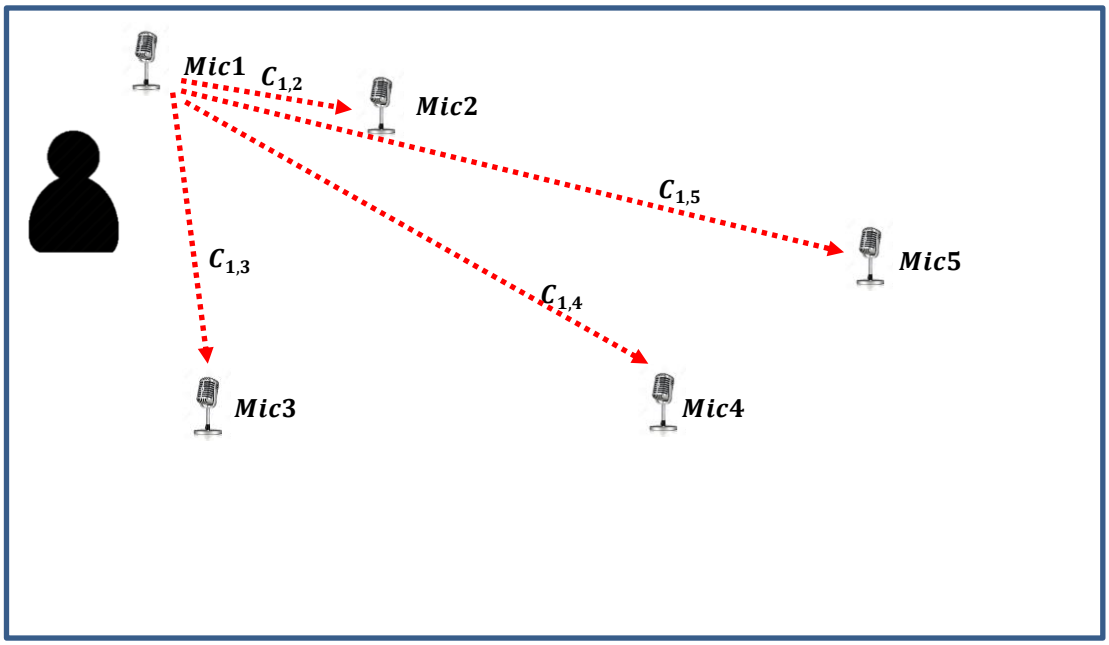


Figure 3-11: Coherence for ad-hoc arrays

Table 3-1: Coherence for the ad-hoc microphones

	Mic.1	Mic. 2	Mic.3	Mic.4	Mic.5
Mic.1	1	0.87	0.73	0.68	0.63
Mic.2	0.87	1	0.71	0.69	0.7
Mic.3	0.73	0.71	1	0.66	0.62
Mic.4	0.68	0.69	0.66	1	0.78
Mic.5	0.63	0.7	0.62	0.78	1

Based on the coherence valued form the table it is concluded that microphone 1 and microphone2 are clustered together and microphone 4 and microphone 5 form a cluster as well. Microphone 3 does not cluster with any microphone as the recorded signal by microphone3 is not similar to any other microphone.

This observation shows that the calculated coherence (or estimated) coherence values obtained for all the microphone pairs can be applied as an indicator for microphones relative distances and their signal similarities however it is noteworthy

that the symmetry issue can still decrease the clustering success rate of this proposed method.

Table 3.2: Coherence based clustering algorithm for source targeting

1. Start with a random microphone as the reference microphone (m_{ref})
2. Estimate the coherence between the reference microphone and all the other microphones, $C_{m,m_{ref}}$, $m = 1, \dots, M$ and $m \neq m_{ref}$.
3. Obtain $C_{m,m_{ref}}(min), C_{m,m_{ref}}(max)$
4. For $m = 1, \dots, M$ and $m \neq m_{ref}$. Cluster the m[th] microphone with the reference microphone, m_{ref}, if $C_{m,m_{ref}} \geq C_{m,m_{ref}}(min) + \frac{C_{m,m_{ref}}(max) - C_{m,m_{ref}}(min)}{2}$
5. Exclude microphone m_{ref} (the reference microphone) and all the microphones clustered with it and return to 1, M times
6. Microphones that are not clustered with any other nodes form a single node cluster

Inspired by [82] the concept of coherence can be expanded to more than two signals by defining the Cross Spectral Density (CSD) for three signals by:

$$C_{m1m2m3}(f) = \frac{|\varphi_{m1m2m3}(f)|^3}{\varphi_{m1m1}(f)\,\varphi_{m2m2}(f)\varphi_{m3m3}(f)} \quad 3\text{-}17$$

where:

$$\varphi_{m1m2m3}(t,k) = F\left(\sum_{\tau=-\infty}^{\infty} R_{m1m2m3}(t)\right) \quad 3\text{-}18$$

For clusters with more than 2 microphones $M = \{m_1, m_2, \dots, m_M\}$ the intra cluster coherence is calculated as:

$$C_M(t,k) = \frac{|\varphi_M(f)|^M}{\prod_{i=1}^{M} \varphi_{m_i m_i}(f)} \quad 3\text{-}19$$

where

$$\varphi_M(f) = F\left(\sum_{i=1}^{M} R_M(t)\right) \quad 3\text{-}20$$

where $R_M(f)$ is the cross correlation for all the M channels. This measurement evaluates the clustering and indicates how close the microphones are within a cluster.

Higher intra cluster coherence means the microphones of that cluster are relatively closer to each other and lower intra cluster coherence mean microphones are apart. Coherence of microphones in a compact arrays recording a single source obtains the maximum value of 1 and two microphone located far from each other and recording two uncorrelated sources obtain the minimum value of 0.

The average intra cluster distance for a cluster with M_1microphones is defined as:

$$\bar{d}_{M1} = \frac{1}{2M_1}\sum_{i=1}^{M1}\sum_{j=1}^{M1} d_{ij} \quad , \qquad i \neq j \tag{3-21}$$

where d_{ij} represents the distance between the i[th] and j[th] microphone in the cluster.

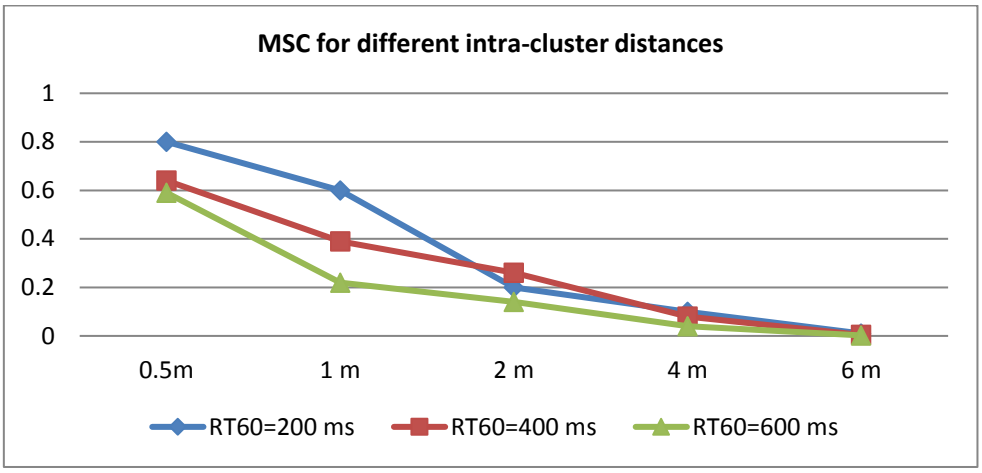


Figure 3-12: Coherence for clusters of three microphones vs. average intra cluster distance ($\bar{d}_M$)

Advantages of the coherence based clustering method are:

- Clustering microphones independently of sources energies
- Forming flexible number of clusters without any limitations
- Utilising the feature (coherence) that indicates the level of reverberation and interference explicitly with constant theoretical maximum (one) and minimum (zero).

Limitations of the coherence clustering method are:

- It might not be applicable to real time applications

3.5 Evaluation and results

Clustering can be difficult to evaluate objectively, as often there is no correct grouping that can be considered as the ground-truth [41]. Evaluation can be even more complicated when clustering is for a specific application (e.g. dereverberation) as not only clustering but the clustered dereverberation outcome should be taken into account as well. This criterion is hard to meet as usually in the meeting scenarios recorded by ad-hoc arrays the reference signal (clean anechoic source signal) is unavailable. This section proposes a systematic evaluation policy for the ad-hoc arrays to compare the clustering methods thoroughly based on the physical clusters spatially spread out within a room.

In this proposed method, the simulated room is a rectangular 8m by 4m by 3m reverberant room. All the microphones and sources are located at the same height (2m). A square grid with 0.5m step size, sweeps the room across the *X* and *Y* axes. 8 microphone clusters of size 4 (4 microphones at each cluster) are distributed on the grid in a way that the distance between the centres of two adjacent clusters is 2m and the microphones within each clusters are located on the vertices of a 0.5m square (Figure 3-13).

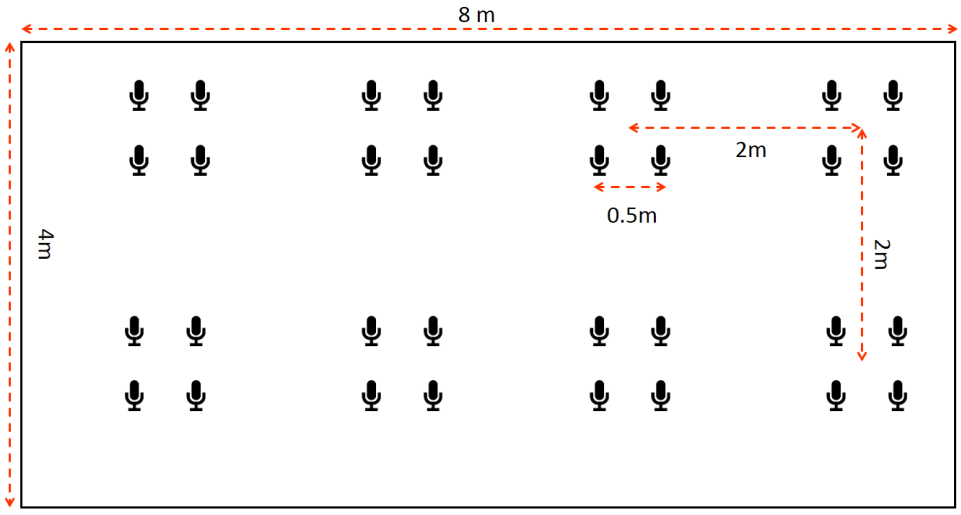


Figure 3-13: Proposed systematic clustering evaluation setup

In order to investigate the effect of the source locations and symmetry, the source is located at 5 different positions as illustrated in Figure 3-14.

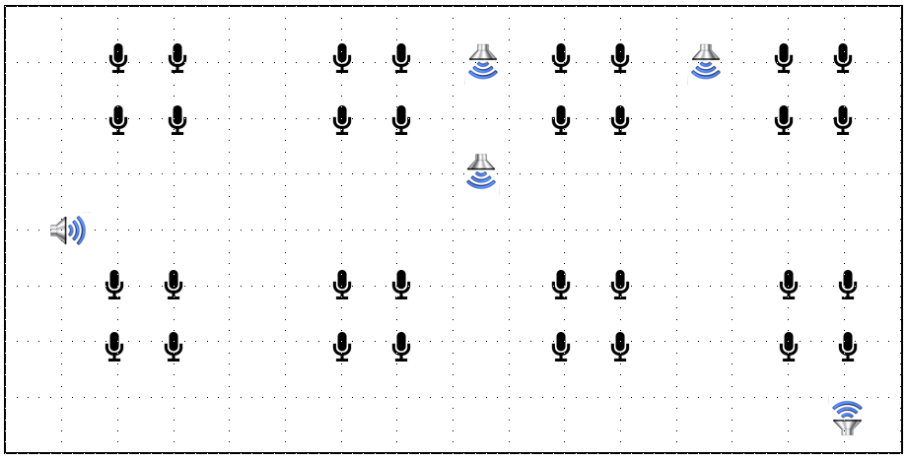
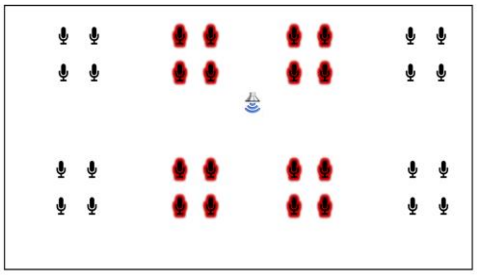


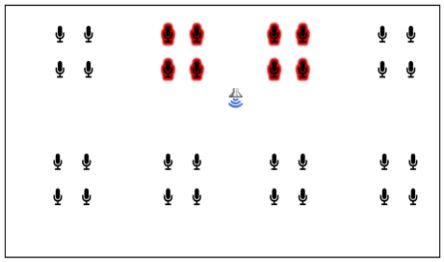
Figure 3-14: Source locations

Assuming that the physical microphone clusters are the ground truth for the acoustic clusters, it is possible to evaluate the proposed and the baseline clustering methods applied to microphone clustering. Figure 3-15 investigates the effect of the applied clustering feature on the formed clusters for one source location.

a)



b)



c)

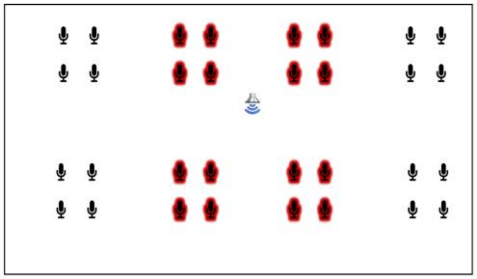


Figure 3-15: The effect of the applied discriminative feature on the formed clusters: a) Proposed time delay and attenuation RIR features b) kurtosis of the LP residual signal c) Coherence, clustered by the kmeans algorithm (k=2)

Figure 3-16 investigate the effect of the source location on the formed clusters (clusters highlighted with red are clustered together and the rest of the clusters are

also clustered together). It is observed that the source location affects the microphones that are clustered together. RIR time delay and attenuation features are applied as the discriminative features for the code-book based clustering method.

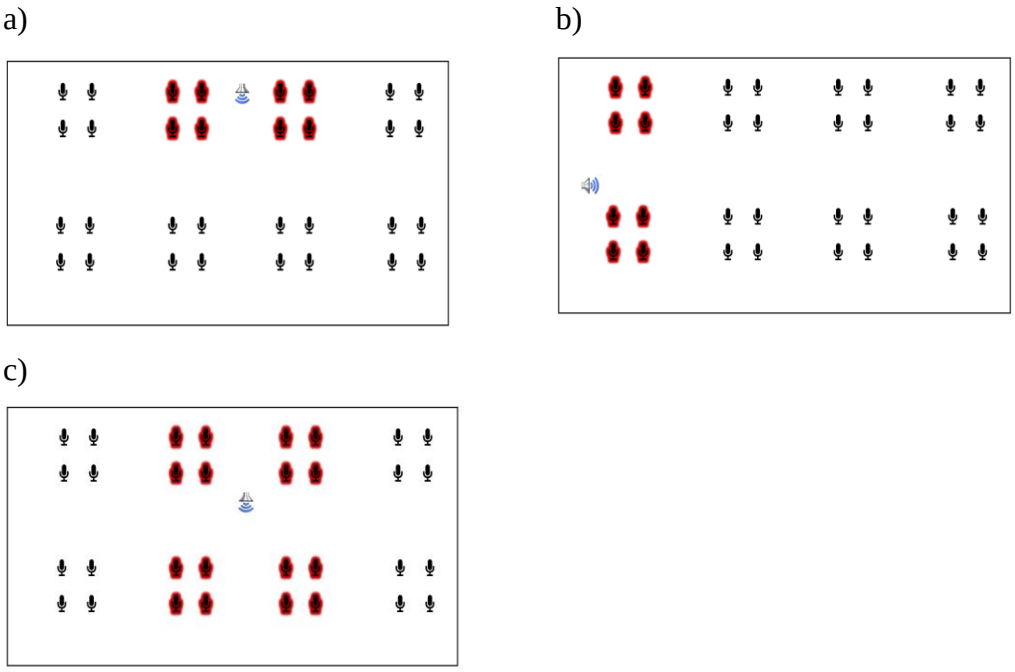


Figure 3-16: The effect of the source location on the formed clusters: coherence based algorithm

This section describes the evaluation results for the proposed code-book based clustering method, proposed coherence based clustering method and the proposed discriminative features. The results are average results for 25 different ad-hoc setups with one active source and 32 microphones. Speech sentences for the voherence features are derived from speech signal from 5 different male and female speakers. Effects of the noise, discriminative features and the applied clustering methods have been investigated. The Limitations and advantages of each method and feature are also highlighted.

The value of L (number of echoes) from (3-4) is an important factor in codebook and discriminative feature vector generation for the code-book based method. For all the experiments L=3, which means the direct path signal along with the first three echoes are utilised as discriminative features. The effect of the L value on clustering

performance and feature extraction is also investigated. *L=0* only considers the direct path signal arrival time and amplitude and does not take into account any of the echoes and therefore it cannot discriminate microphones effectively. On the other hand, when the number of echoes increases (e.g. *L*=8), first order echoes (direct path signal reflected off a reflector) and second order echoes (echoes reflected off a reflector) get mixed up and that causes error. Generally, there is one direct path signal (*L=0*) and 6 first order reflections (four walls plus the ceiling plus the floor) and considering more echoes is not helpful as some second order echoes arrive before some first order echoes at a microphone position.

For *M* randomly positioned microphones, if microphone m_j is clustered with other spatially close microphones (inter-cluster distances compared with mean intra-cluster distance), the microphone m_j clustering result is labeled *"V"* (Valid) otherwise is labelled *"I"* (Invalid). The success rate, SR [75], is applied to evaluate all methods and is calculated as:

$$SR = \frac{n(V)}{n(I+V)} \times 100 \qquad 3\text{-}22$$

where *n(V)* is the number of microphones clustered correctly and *n(I+V)*is the total number of microphones (*M*) from 3-4. The effect of different number of applied echoes (L) on the clustering SR (3-22) is investigated in Figure 3-17. The error bars roughly show the variation of the SR for each value of L. 20 random scenarios are calculated for each L value.

Figure 3-17: The effect of the number of echoes on the clustering SR

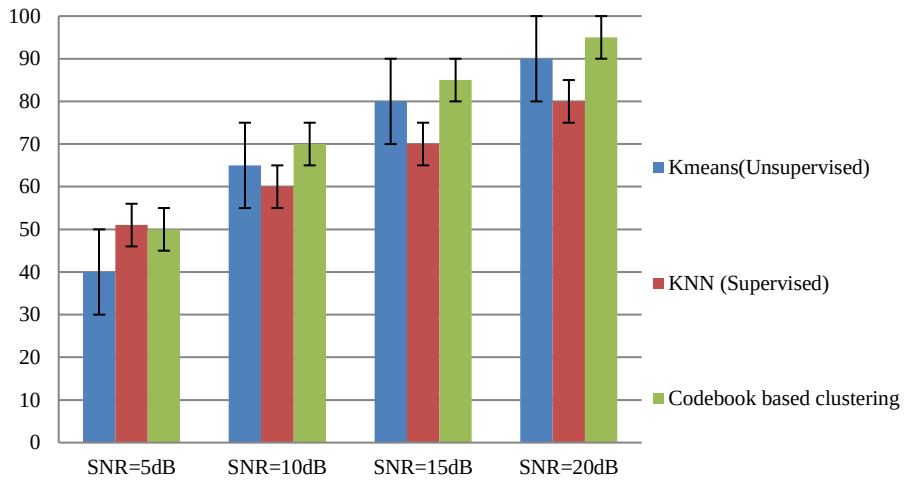


Figure 3-18: Microphone clustering Success Rate (SR) for 5 center points at different noise levels

Noisy signals from Loizou data-base [83], [84] at different SNR's are added to the simulated RIRs with an 8 kHz sampling rate and $RT_{60} = 100ms$ to $RT_{60} = 600ms$ for all experiments. It is concluded that noise affects all the methods and the highest SR is achieved by the highest SNR and the proposed RIR features (Figure 3-18).

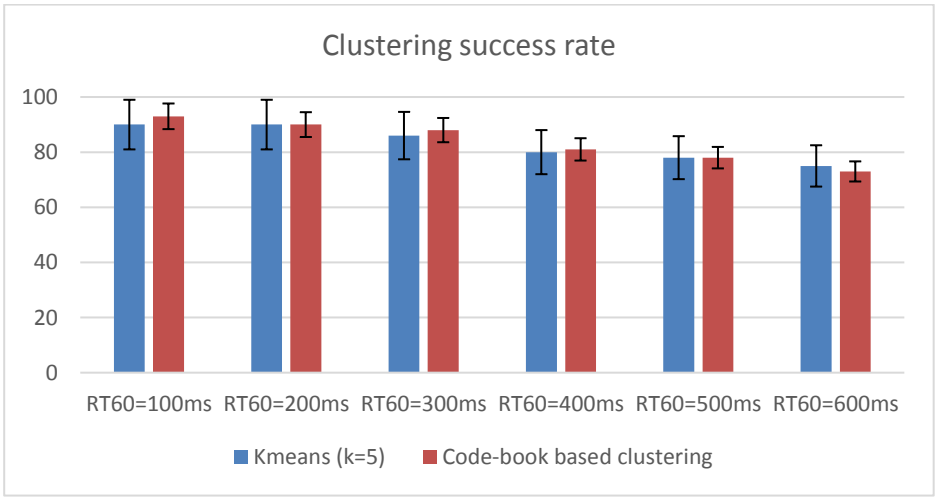


Figure 3-19: The effect of RT60 on clustering success rate.

The effect of reverberation time is also investigated on the clustering success rates for Kmeans clustering methods with k=5 and the proposed code-book based method with 5 centre points. It is concluded that the highest success rate is achieved when the reverberation time is very small (e.g. 100ms) (Figure 3-19).

A supervised K nearest Neighbour (KNN) method can also be applied for microphone segmentation but as the results suggest, mismatch between the clean training set and noisy test set affects the success rate of the supervised method (i.e. KNN) significantly (Figure 3-18).

Based on these results it is concluded that the proposed codebook-based method provides the highest success rates for all SNR conditions assuming that the RIRs at the centre points and the microphones' locations are available.

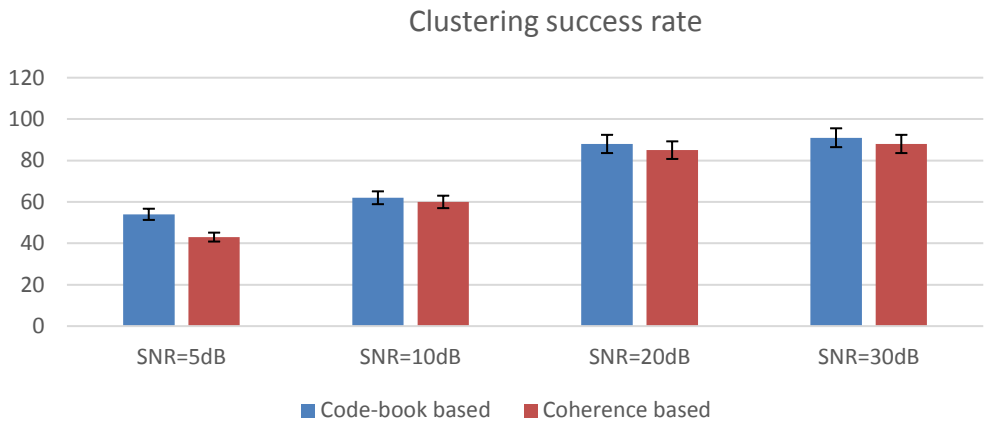


Figure 3-20: comparison of the proposed methods

The two proposed methods are compared in Figure 3-20. Although these two methods require different assumptions (e.g. dual microphone nodes for the coherence based and the knowledge of RIRs for the code-book based method) the experimental setups are similar (i.e. the source position, room geometry and the microphone locations). It is concluded that the code-book based method is more accurate however it is shown that both methods are highly affected by noise.

The overall comparison of the proposed methods and features and their limitations are presented in Table 3-2.

Table 3-2: Proposed features and clustering methods

Method	Feature	Limitations	Target application
Proposed Code-book based	*RIR time delays and amplitude*	*Requiring the cues at the centre points*	*Meeting at rooms with an available code-book*
Proposed Code-book based	*Kurtosis of LP residual signal*	*Requiring the cues at the centre points*	*Meeting at rooms with an available code-book*
Proposed Coherence based	*Coherence magnitude square*	*Dual nodes are required*	*Meetings, press conferences*
Baseline K-means	*Kurtosis, RIR cues, coherence*	*Pre-defined number of cluster*	*Meetings, press conferences*

3.6 Conclusion

This chapter described two novel approaches to clustering microphones to form ad-hoc arrays based on discriminative features derived from the RIRs and speech signals. The RIR features represent the time delays of the echoes and the peak amplitudes received by the microphones and provide a compact set of parameters for use within supervised and unsupervised learning algorithms including a proposed codebook-based approach. The coherence feature is derived from speech signals recorded by dual microphone nodes. Investigations and simulations of this research showed that by using a relatively small codebook (5 centre points), it is possible to cluster microphones in reverberant environments accurately. Effects of the number of applied echoes (L), SNR, the number of centre points and RT_{60} time on the clustering performance are also investigated. Results suggest that the proposed codebook-based clustering algorithm can outperform KNN supervised classification method and Kmeans unsupervised clustering method applied to microphone segmentation and clustering, in terms of clustering success rate and robustness to noise.

Comparison of the proposed methods and the state of the art features applied within baseline clustering algorithms show that the proposed methods can outperform the cepstral features and the standard clustering techniques. The proposed coherence based method does not require any prior knowledge of the number of clusters and flexibly choose the right number of cluster based on their spatial distance (estimated by the coherence feature).

The effect of noise is investigated and it is concluded that the increase in the noise level distorts the echo peaks and the signals and consequently decreases the accuracy of the extracted features and clustering results. It is also concluded that noise has a more destructive effect on the code-book-based method compared to the other methods investigated.

4 Source localisation with ad-hoc microphone arrays

4.1 Introduction

This chapter proposes a novel source localisation method in the context of ad-hoc microphone arrays by extracting relative source to microphone distance cues from the RIRs and the speech signals [85].

Estimating the location and the Direction of Arrival (DOA) of the sound sources from microphone recordings has various applications including informed noise cancellation [86] and speech enhancement where noise is estimated based on its angle of arrival or phase [87]. This type of approach to the informed speech enhancement, typically requires the use of a known geometry microphone array, and the resulting multichannel recordings [88] are processed to obtain information such as the Time Difference of Arrival (TDOA) [89] that can then be used for estimating the source DOA [90], [91]. An alternative is to form an ad-hoc array from randomly placed microphones. Such an approach has challenges such as not knowing the location of each microphone, the inter-channel time delays or the phase difference between the recorded signals, which makes the state of the art approaches inapplicable to such scenarios.

A novel source localisation method using ad-hoc microphone arrays, exploiting energy attenuation as discriminative cues is proposed in [31], which is independent from the microphones gains. The proposed method in [31] is only applicable to meeting scenarios where all or most sources (i.e. 4 out of 7) and microphones are collocated or distributed within a fairly small area such as a meeting table.

Recently, obtaining the TDOA of the direct and echo components of the Room Impulse Response (RIR) has been used to derive information such as microphone locations and room shape [92]. It is also shown that RIRs can accurately localise microphones and sources if some prior information (i.e. Room geometry and the location of one microphone) is available [48], [93]. However the problem with such supervised methods is their dependency on the training data, training setup, and the participating speakers.

In this chapter the proposed features derived from RIRs and speech signals are utilised for source localisation through a novel surface-fitting method applied to the features. The proposed method of this chapter overcomes the limitation of the state of the art methods such as requiring the microphones (nodes) to communicate together [94] or assuming that sources and microphones are collocated [31]. The accuracy of the proposed method is evaluated through simulations of varying numbers of microphones that are uniformly distributed throughout rooms with different acoustic transfer functions.

The objectives of this chapter are:

- Extracting relative distance cues from ad-hoc microphones at unknown locations
- Discriminating the microphones located closer to the active source from the microphone located far from the source by analysing the proposed relative distance cues.

The main contribution of this chapter is proposing a source localisation method when the RIRs are available

- Source localisation in the context of ad-hoc arrays where the distances between the microphones and the source are unknown.
- Utilising the RIRs and the speech signals for distance cue extraction.

Publications arising from the contributions of this chapter include

- S. Pasha & C. Ritz and Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, pp. 1-6.
- S. Pasha & C. H. Ritz, "Informed source location and DOA estimation using acoustic room impulse response parameters," in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015, pp. 139-144.

4.2 The proposed surface fitting method

The RIRs describe the effect of sound transmission from a source to a receiver (microphone) in a reverberant room, and includes the reflections from the walls, ceilings and the floor. Herein the parameters extracted from RIRs are exploited to fit a TDOA [95] surface and an amplitude surface across the room which can estimate the source location. Other than time and amplitude features, Magnitude Square Coherence (MSC) and the clarity feature (C_{50}) carry relative location estimation as well. MSC can be derived from dual microphone nodes and the clarity feature is derived from RIR recordings which make them applicable to certain scenarios.

Figure 4-1 shows the block diagram of the proposed method.

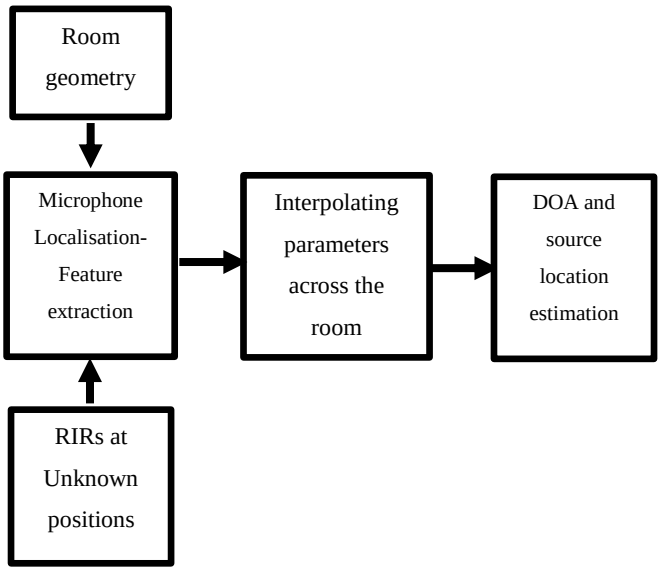


Figure 4-1: The proposed source location estimation method

Other than time and amplitude features, Magnitude Square Coherence (MSC) and the clarity feature (C_{50}) carry relative location estimation as well. MSC can be derived from dual microphone nodes and the clarity feature is derived from RIR recordings which make them applicable to certain scenarios.

4.3 Relative distance cues

The applied distance cues in this chapter are categorised into two different categories: 1) cues derived from recorded, simulated or estimated RIRs at each unknown microphone location 2) cues derived from speech signals recorded by dual microphone nodes.

RIR cues include time delays, attenuation and the clarity feature whereas MSC is derived from speech or noise signals by dual microphone nodes [69].

4.3.1 RIR time delay and attenuation cues

This section describes how source localisation is performed by deriving TDOA and relative amplitude attenuation information from recordings of the RIR obtained using an ad-hoc microphone array. It is shown that in sensor array processing, applying all the microphones in an array is not necessarily the optimised approach for applications such as signal classification [5] equalisation [63] and beamforming [43] and also it is shown that ad-hoc microphone arrays can localise sources more accurately than compact arrays due to their spatial coverage [96]. Based on these two observations, a clustered ad-hoc approach is proposed in this chapter as a modified scheme for source localisation. The justification for this hypothesis is that microphones located far from the source are highly distorted by undesired components such as noise, interference and reverberation and they usually have a lower Direct to Reverberation Ratio (DRR), so excluding these distorted microphones from the array leads to a smoother RIR cues surface fitting and consequently a more accurate source localisation process. In this chapter an attempt is also made to define a practical threshold for which applying microphones within that threshold yields the highest localisation accuracy.

In a scenario of *M* synchronised microphones and one active source at each time frame the j^{th} microphone position in the 3D Cartesian coordinates is $r_{mj} = [x_j, y_j, z_j]$ and the source is located at $r_s = [x_s, y_s, z_s]$. It is assumed that r_s and r_{mj} for *j=1* to *M* are not available however as the room geometry is known the tested microphone localisation approaches such as in [64], [97] could be applied to localise the microphone in 2D coordinates.

For RIR recording an exponential sine sweep method with a starting frequency of 22Hz and ending frequency of 22 kHz gives an accurate linear room impulse response. The method of [98] can be applied to record the RIRs of all microphones from *j=1* to *M*:

$$h(n) = \sum_{l} a(l)\delta\big(n - d(l)\big) \qquad 4\text{-}1$$

Each RIR can be represented as a train of impulses where $a(l)$ is the amplitude of the l^{th} sample and $d(l)$ is the relative time delay with respect to the direct path impulse.

The time delays and amplitude cues extracted from the RIRs have been applied to microphone clustering applications [75] as these cues reflect the distances between the microphones, the active source and reflectors (e.g. walls). Mathematically, the TOA can be calculated only if the distance between the source and the microphone *j* is known under the assumption that the microphone recordings are synchronized. However the TDOA can be measured for the microphones and the sources at unknown positions if the RIRs are available.

$$TOA_{j,s} = \frac{\|r_s - r_j\|}{c} + \tau_j \quad 4\text{-}2$$

$$h = \{h_1, \dots, h_M\} \quad 4\text{-}3$$

where r_s, r_j are the source and the microphone *j* coordinates. c is the speed of sound. As it is suggested by 4-2, r_s, r_j and the onset delay for each microphone (τ_j) is required for $TOA_{j,s}$ calculation. However if the unsynchronised recordings at two microphones locations are available the $TDOA_{i,j}$ between these two microphones can be obtained without the knowledge of r_s and r_j.

It is observed that the $TOA_{j,s}$ and TDOA have a direct relationship to the spatial distances between the source and the microphone position (Figure 4-2). This relationship can be exploited for source localisation applications.

$$TDOA_{i,j} = \|TOA_{i,s} - TOA_{j,s}\| \quad 4\text{-}4$$

$$TDOA_{i,j} = \frac{\|r_i - r_j\|}{c} \quad 4\text{-}5$$

$TDOA_{i,j}$ is a function of r_s and r_j and 4-5 suggests that the knowledge of source and microphone locations is required for $TDOA_{i,j}$ calculation. However the $TDOA_{i,j}$ can be estimated by the cross correlation method [99]. Assuming that the RIRs recordings for microphone i and j are available (h_i and h_j) the cross-correlation between these two RIRs is defined as:

$$h_i(n) * h_j(n) = \sum_{m=-\infty}^{m=+\infty} h_i(m) h_j(m+n) \quad 4\text{-}6$$

$$TDOA_{i,j} = \arg\max_n (h_i(n) * h_j(n)) \quad 4\text{-}7$$

$$k_{TOA} = \{TOA_1 \dots, \dots TOA_M\} \quad 4\text{-}8$$

$$k_{TDOA} = \{TDOA_{1,ref} \dots TDOA_{j,ref} \dots TDOA_{M,ref}\} \quad 4\text{-}9$$

k_{TOA} is the vector of the TOA features [75], k_{TDOA} is the vector of the TDOA features and $TDOA_{j,ref}$ is the TDOA between microphone j and an arbitrary reference signal from 4-4.

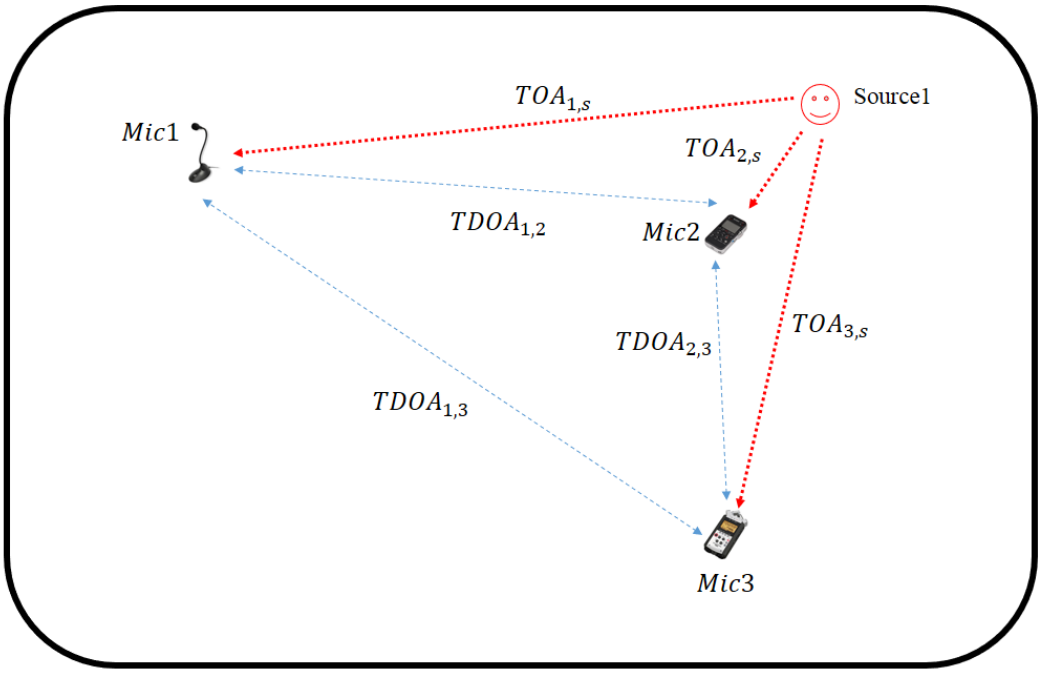


Figure 4-2: TDOA and TOA

It is also suggested that the source to microphone distance and the signal energy attenuation (A_j) are directly related [31]:

$$A_j = \frac{1}{g} \times \|r_s - r_j\| \quad 4\text{-}10$$

where g is the microphone gain. Assuming that all the microphones have the same gain, the attenuation feature 4-10 contains source to microphone relative distance information. Although 4-10 suggests that for calculation of A_j (RIR energy at microphone j location), the source location (r_s), microphone j location (r_j) and the microphone gain (g) are required, if the RIRs are available (4-6), A_j can be calculated:

$$A_j = \frac{1}{\|h_j\|} \quad 4\text{-}11$$

where

$$\|h_j\| = \sum_l |a(l)|^2 \quad 4\text{-}12$$

Assuming

$$g_1 = \cdots = g_j = \cdots = g_M \quad 4\text{-}13$$

The vector of the attenuation discriminative features is

$$k_A = \{A_1 \ldots A_j \ldots A_M\}. \quad 4\text{-}14$$

These two sets of cues (time delays and attenuation) (4-9 and 4-14)and their relationships with the spatial distances are exploited to fit two surfaces which can be utilised for source location and DOA estimation in a room of known geometry [85].

4.3.2 C50 or clarity measurement

The C_{50} or the Clarity measurement is the ratio of early to late reverberation expressed in dB. This measure is higher when the microphone to sources distance is relatively small and the recorded RIR by the microphone is dominated by the direct path impulse [100] . In contrast The C_{50} is lower when the microphone to source distance is relatively large and the second and third order reverberations are no longer negligible. It is shown that the C_{50} has an inverse relationship with the microphone to source distances and its calculation does not require the clean signal (in contrast to the Direct to Reverberation ratio (DRR)) [60]. The C_{50} is defined as the energy of the direct path impulse and the early reverberations divided by the energy of the late echoes:

$$C_{50} = 10 \times \log\left(\frac{E_{direct} + E_{early}}{E_{late}}\right) \quad 4\text{-}15$$

With

$$E_{Direct} = E(a_1\delta(n)) = a_1, \quad 4\text{-}16$$

$E_{early} = \sum_0^{N_{early}} h(n)$, and $E_{late} = \sum_{N_{early}}^{\infty} h(n)$ from (4-1) and *n* is the sample index. C_{50} can also be calculated for each RIR independently without synchronisation by:

$$C_{50} = 10 \times \log\left(\frac{\sum_0^{N_{early}} h(n)}{\sum_{N_{early}}^{\infty} h(n)}\right) \quad 4\text{-}17$$

In this chapter the hypothesis is that estimated C_{50} values across the room obtain local maxima at source locations and they fade as the microphones move away from source locations.

The advantage of using C_{50} is that nodes can be of any structure and there is no constraint on the number of microphones in each node however full knowledge of the RIRs is required.

$$k_{C_{50}} = \{C_{50_1} \dots C_{50_j} \dots C_{50_M}\} \quad 4\text{-}18$$

4.3.3 Magnitude Square Coherence (MSC)

Reverberation and interference recorded by each microphone are functions of its location in the room [61], [64]. When the microphone's signals are distorted by reverberation and interference they become statistically more independent and they will have lower intra MSC values calculated by:

$$C_{12}(f) = \frac{|\varphi_{m1m2}(f)|^2}{\varphi_{m1m1}(f)\,\varphi_{m2m2}(f)} \quad 4\text{-}19$$

where $\varphi_{m1m1}(f)$ and $\varphi_{m1m2}(f)$ are auto and cross power spectral densities between microphone m_1 and m_2. If the nodes in the ad-hoc array contain dual-channel microphone systems, it is possible to discriminate highly distorted nodes (located far from the active sources) and the node's signals predominated by the speech signals (located closer to one of the sources). This fact about MSC is utilised here as a distance cue to estimate the distances between the active sources and the nodes [62].

By applying the general equation of MSC to two microphones in the m^{th} node the signals can be modelled as:

$$\mathrm{y_{m,1}}(t,f) = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{s_n}(t,f) * \mathrm{h_{m,1,n}}(t,f) + \mathrm{v}(t,f) + \mathrm{w_{m1}}(t,f) \quad 4\text{-}20$$

$$\mathrm{y_{m,2}}(t,f) = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{s_n}(t,f) * \mathrm{h_{m,2,n}}(t,f) + \mathrm{v}(t,f) + \mathrm{w_{m2}}(t,f) \quad 4\text{-}21$$

and the MSC between these two microphones can be calculated by:

$$C_{y,m_1y_{m2}}(f) = \frac{|\varphi_{y,m_1y_{m2}}(f)|^2}{\varphi_{y,m_1y_{m1}}(f)\,\varphi_{y,m_2y_{m2}}(f)}. \quad 4\text{-}22$$

By moving away from an active source the microphones in the node will have lower $\varphi_{y,m_1y_{m2}}(f)$ values as the direct path signals attenuate and $v(t,f)$, $w_m(t,f)$ will become stronger (in terms of the signal power) therefore $\varphi_{y,m_1y_{m2}}(f)$ will

decrease whereas $\varphi_{y,m1y_{m1}}(f)\,\varphi_{y,m2y_{m2}}(f)$ do not change with distance significantly.

Table 4-1: The relationship between the MSC and the source to microphone distance

	Distance to the active source	MSC	RT_{60}	Number of microphones
Node1	10cm	0.963	600ms	2
Node2	0.5m	0.898	600ms	2
Node3	3m	0.819	600ms	2
Node1	10cm	0.999	200ms	2
Node2	0.5m	0.908	200ms	2
Node3	3m	0.876	200ms	2

The effect of the dual-microphone node to the active source distance on the MSC values in a reverberant room is presented in Table 4-1. It is clear as there is only one active source (no interference from other sources) in the room MSC values are very close to 1 and they only change with the distance.

Table 4-2 MSC and distance to two simultaneously active sources

	Distance to source1	Distance to source2	MSC	RT_{60}	Number of microphones
Node1	10cm	3 m	0.78	600ms	2
Node2	0.5m	2.6m	0.43	600ms	2
Node3	3m	10cm	0.82	600ms	2
Node1	10cm	3m	0.78	200ms	2
Node2	0.5m	2.6	0.30	200ms	2
Node3	3m	10cm	0.81	200ms	2

In Table 4-2 however the effect of the interference on the MSC values are highlighted and it is interestingly observed that the interference decreases the coherence between the channels.

Table 4-3: Noise effect on MSC					
	Distance to the active source	SNR	MSC	RT_{60}	Number of microphones
Node1	10cm	10dB	0.78	600ms	2
Node2	0.5m	10dB	0.61	600ms	2
Node3	3m	10dB	0.40	600ms	2
Node1	10cm	20dB	0.85	200ms	2
Node2	0.5m	20dB	0.71	200ms	2
Node3	3m	20dB	0.65	200ms	2

The effect of noise on the MSC values is investigated in this Table **4-3** and it is concluded that noise also affects the coherence of the microphones in one node.

The disadvantage of applying the MSC is that all nodes should have the same structure as the MSC is a function of intra node microphone distances and there should be at least two microphones at each node. On the other hand, MSC can be applied to any type of recorded signals and the recorded RIRs are not required.

The MSC (4-22) is a vector as it is a function of frequency. In order to obtain one value for each microphone during the time frames the averaged MSC across the frequencies is calculated as

$$MSC_m = \sum_f \frac{\left|\varphi_{y,m_1 y_{m2}}(f)\right|^2}{\varphi_{y,m_1 y_{m1}}(f)\,\varphi_{y,m_2 y_{m2}}(f)} \qquad 4\text{-}23$$

where f_e is the upper frequency and f_s is the lower frequency limit. By calculating (4-23) for all the dual nodes the vector of the features is obtained as:

$$k_{MSC} = \{MSC_1 \dots MSC_j \dots MSC_M\} \qquad 4\text{-}24$$

4.4 Microphone positions and the extracted Cues

The proposed method in this chapter compares the features described in the previous section as alternatives for source localisation. By extracting these features, a surface is fitted to the area of the room illustrating the interpolated feature's values at any point in the 2D room (Figure 4-3). It is important to mention that having the knowledge of the room geometry and the RIRs it is possible to localise the

microphones [64] and if the RIRs are not available (the case that MSC is applied) the microphone locations are required.

In this research the area of the surface with the following criteria is highlighted as the source area.

- Lowest TOA (Estimated by RIRs)
- lowest RIR energy attenuation
- Highest MSC
- Highest C50

The center of these areas is calculated and considered as the estimated source location ($\widehat{x_s}$, $\widehat{y_s}$).

The positions of the *M* microphones in 2D coordinates in the room are represented in a matrix as:

$$P = \begin{bmatrix} x_{mic\,1} & \dots & x_{mic\,M} \\ y_{mic\,1} & \dots & y_{mic\,M} \end{bmatrix} \quad 4\text{-}25$$

This matrix can be calculated if the RIR at the microphones locations are available [64] and by also having the derived cues and assuming that $M > 3$, it is possible to fit two surfaces in order to interpolate the cues values at all points in the room.

The extracted feature values for *M* ad-hoc microphones are presented in a vector as:

$$k = [k_1 \quad \dots \quad k_M] \quad 4\text{-}26$$

The available data points are

$$f(x_{mic\,j}, y_{mic\,j}) = k_j \quad 4\text{-}27$$

where $x_{mic\,j}, y_{mic\,j} \in P$ and $k_j \in k$. The objective is to find (interpolate) the function *f* such that $f(x_{mic\,m}, y_{mic\,m}) = k_m$, where $x_{mic\,m}, y_{mic\,m} \notin P$ and $k_m \notin k$ [101]. The surface *f(x,y)* has a general form of

$$(ax_m + by_m)^n = k_m \quad 4\text{-}28$$

As a contribution in this chapter the clustered approach to multi-channel source localisation is proposed and tested. It has previously been shown that an ad-hoc array can localise a source more accurately than compact arrays due to their spatial coverage [77] and clustered approaches are shown to be more effective than blind use of all microphones in the array for certain applications such as beamforming [76] and speech recording and classification [102]. However the boundaries of the formed

clusters/subsets are usually specified by the applied clustering algorithms, which are not necessarily forming the optimised clusters for each application. This research is trying to address this issue for source localisation by defining an outcome-based threshold for source location estimation accuracy in noisy reverberant environments.

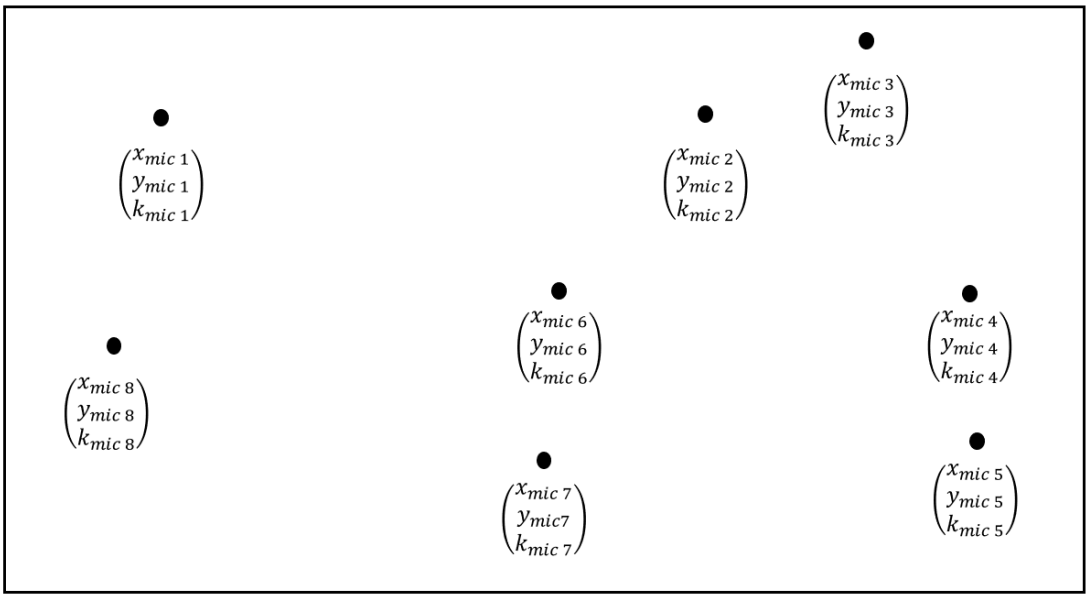


Figure 4-3: Microphone locations and features

4.5 Clustered surface fitting approach

In this section the focus is on extracting features from RIRs and speech signals in order to estimate the source location and DOA where the only primary information is the room geometry. Ad hoc microphones and sources in a room can be localised accurately by the cues derived from speech signals and RIRs [64]. However, for some applications accurate localisation of the source and perfect reconstruction of the acoustic scene are not necessary and simply discriminating distant and close sources/speakers is helpful enough. In other applications such as noise estimation/cancellation, DOA estimation is informative enough to discriminate the noise source and the target source and accurate localisation is not required [103]. This chapter does not focus on microphone localisation as they are investigated in the literature [7]. The surface fitting approach to the source localization is depicted in Table 4-4.

Table 4-4

Surface fitting source localization method for ad-hoc scenarios

a) Start with the RIRs at the microphone locations (4-1)
b) Extract the relative distance cues for each microphone RIR(e.g. 4-18)
c) Obtain the locations and the features pairs Figure 4-3
d) Fit the surface to the room based on the feature values (4-28)
e) Detect the source area based on the fitted surface (Figure 4-4, Figure 4-5)
f) If the clustered approach is applied use a subset of microphones located closer to the source (estimated by the extracted features)(Figure 4-6)

The active source is located in the region with the minimum arrival time value and the highest direct path amplitudes on the interpolated values. TDOA and attenuation cues of a subset of microphones within or close to this area can be exploited to achieve a more accurate source localisation.

Associating the feature values and the microphones location the following nodes can be obtained on the room 2-D plane: it is observed that fitted surface to the TOA values estimated by the RIRs can accurately localise the source. The red dot in the yellow area (Figure 4-4) is the source and the red dots in the other areas are the microphones.

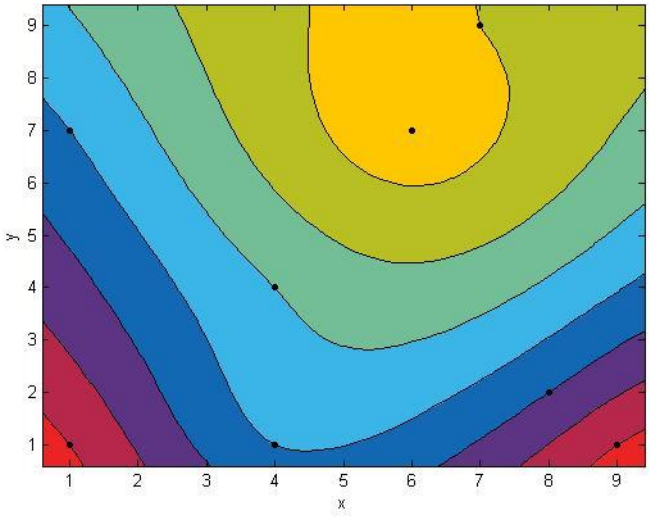


Figure 4-4: fitted surface to the time delays

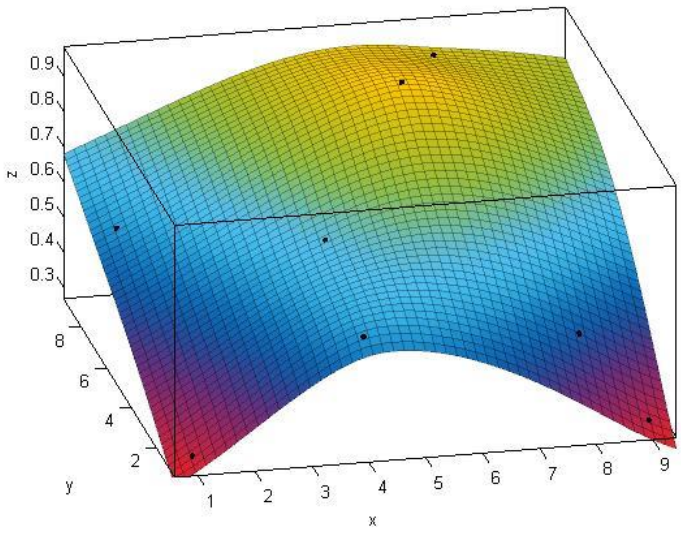


Figure 4-5: Fitted surface to the amplitude cues derived from RIRs of 8 microphones

The fitted surface localizes (Figure 4-4, Figure 4-5) the source and the Direction of Arrival of the source to each node but it is possible to go further and choose a subset of nodes (microphones) which are located close to the source (Figure 4-6) and exclusively utilise them for the surface fitting approach. This clustered approach has two main benefits; firstly, it removes the highly distorted nodes (due to reverberation) from the array; and secondly in large arrays it simplifies the surface fitting process.

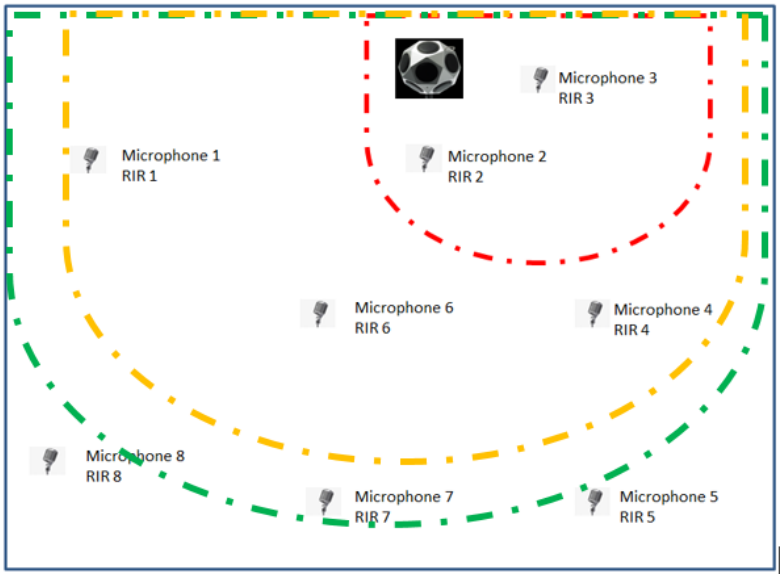


Figure 4-6: The clusters obtained by using 2, 5 and 6 closest microphones to the source

4.6 Results

The following table shows the experimental configuration of the evaluation process. The source and the microphones are randomly positioned within the room and the only available prior information is the room geometry.

Table 4-5: Experimental configuration

f_s	16kHz
RT_{60}	200ms,400ms,600ms,800ms
Room size	10m,8m,3m
Number of microphones	5 to 20
Noise	White noise, Babble noise
SNR	10,20,30dB and clean signals

The next graph shows the average results for 30 different random scenarios with 10 microphones randomly spread out in the room. The microphones relative distances to the source is estimated by the extracted features (RIR time delays) and the starting point is utilising only half of the microphones which are located closer to the source (I/M=0.5). It is shown that using I/M=0.7 or I/M=0.8 yields better results compared with the use of all the microphones in the room (I/M=1). It is shown that

consciously chosen subset of 6 microphones (out of 8) yield a more accurate source localisation (0.5m error in a 10m by 10m room) (Figure 4-7) whereas blind use of all the microphones increases the error to 0.6m. the applied features are the time delays and the attenuation features [85]. Assuming that the highest source localisation accuracy is achieved by exploiting I closest microphones to the source, in this research $\frac{I}{M}$ (the number of applied microphones divided by the total number of microphones in the ad-hoc array) is calculated as the ratio of the applied microphones (4-7).

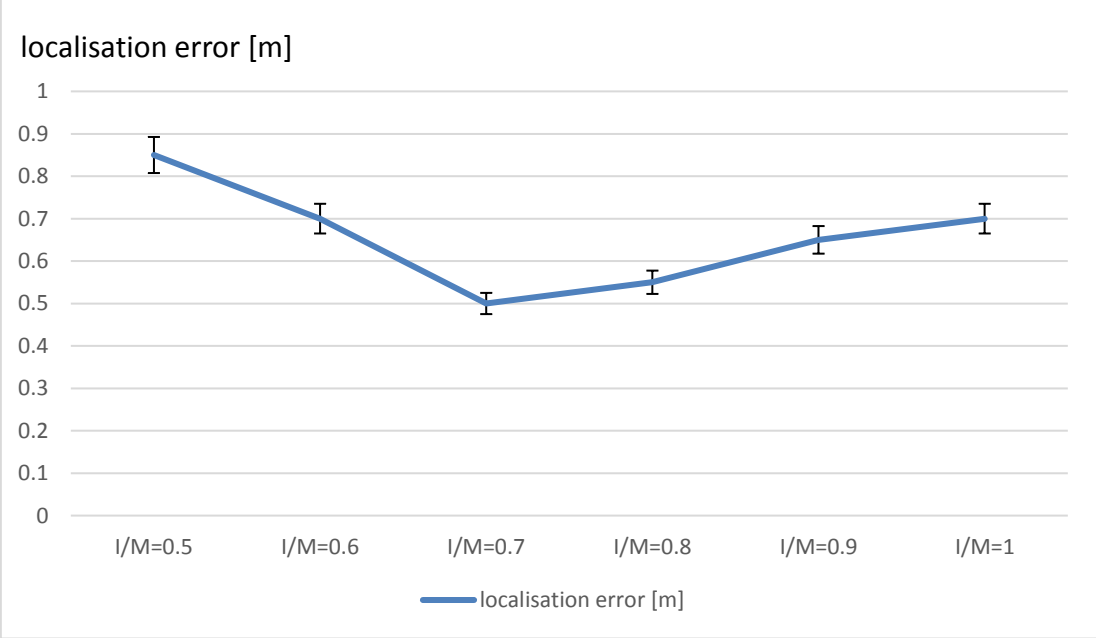


Figure 4-7: Localisation error for clustered surface fitting

The comparison of the proposed distance features derived from the speech signals, is presented in Table 4-6 and Table 4-7 for different numbers of microphones available in the room. Again it is concluded that a higher number of microphones does not necessarily lead to a more accurate source localisation and utilising a subset of microphones located closer to the source can estimate the source location more accurately.

Table 4-6: comparison of the applied features

Applied feature	SNR [dB]	RT_{60} [ms]	Number of channels	Localisation error [m]
$\boldsymbol{C_{50}}$	10	200	10	0.5
MSC	10	200	10	0.8
$\boldsymbol{C_{50}}$	10	200	15	0.4
MSC	10	200	15	0.5
$\boldsymbol{C_{50}}$	10	200	20	0.4
MSC	10	200	20	0.8

Table 4-7: comparison of the applied features

Applied feature	SNR [dB]	RT_{60} [ms]	Number of channels	Localisation error [m]
$\boldsymbol{C_{50}}$	20	200	10	0.7
MSC	20	200	10	1.2
$\boldsymbol{C_{50}}$	20	200	15	0.9
MSC	20	200	15	1.1
$\boldsymbol{C_{50}}$	20	200	20	0.7
MSC	20	200	20	1.2

As explained before the MSC can be estimated for the dual microphone nodes but C_{50} only requires one microphone per node to be calculated. It is concluded that applying C_{50} yields better results compared with the MSC.

Assuming that the room dimensions are $X(m) \times Y(m) \times Z(m)$ the step size *u* is set to 1m, 2m, 3m,4m in order to investigate the effect of the microphone grid resolution (Figure 4-8). In a 10m,8m,3m room *u=1,2,3,4* translate to 80, 20, 6, 4 microphones respectively.

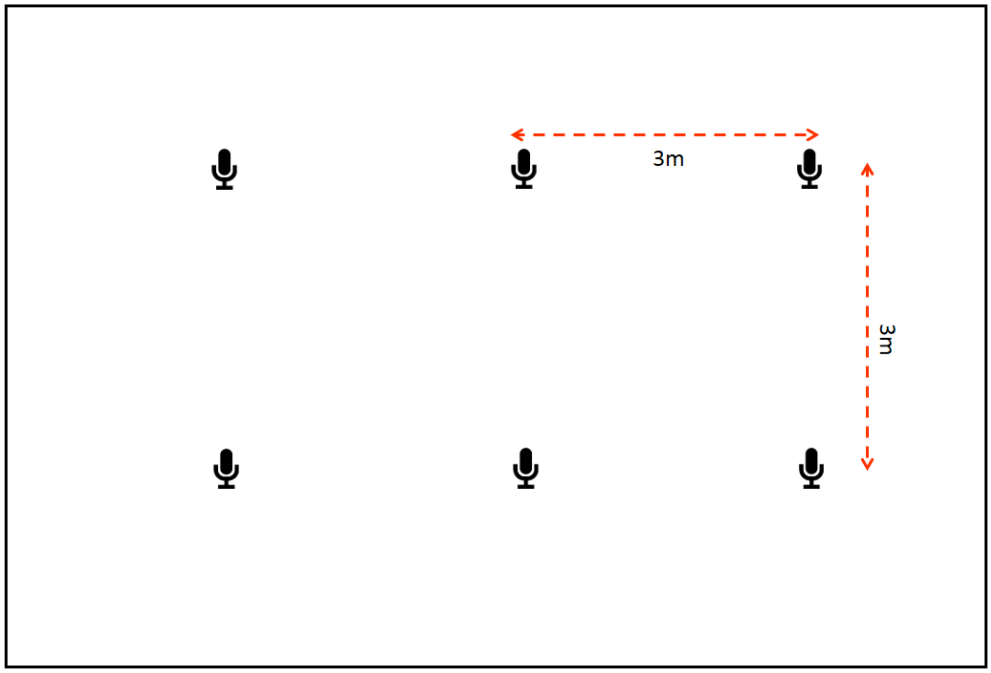


Figure 4-8: u=3m in a 10m by 8m by 3m room.

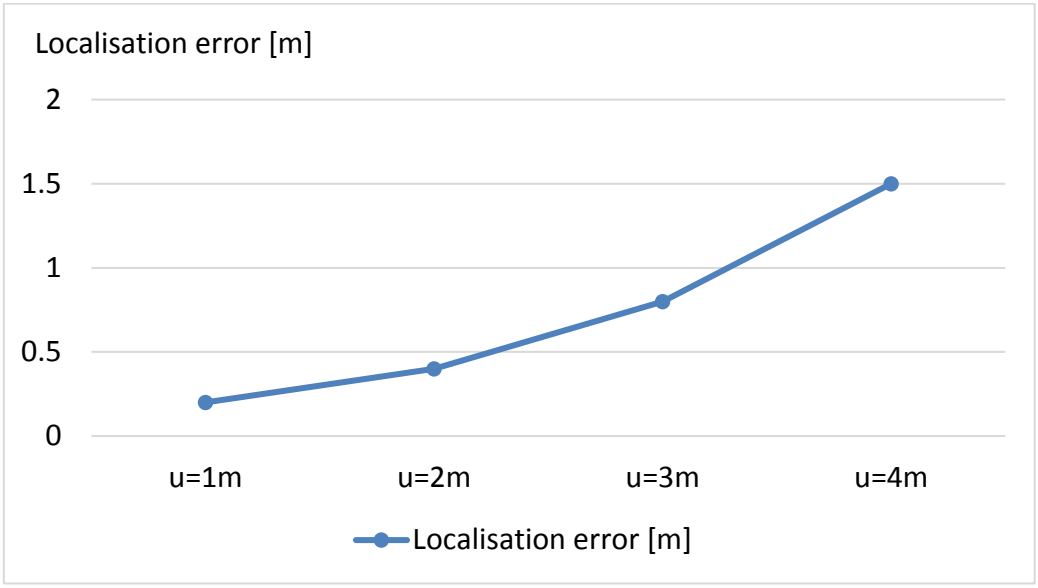


Figure 4-9: Average localisation error for different microphone distributions

It is observed that the source localisation error increases drastically with the microphone grid resolution (Figure 4-9) and the highest accuracy is obtained by *u=1m* (minimum grid step size). Although the applied setup in this experiment does not qualify as an ad-hoc array (as it is not random) but it is necessary to test the proposed method in a non-random manner in order to cancel the effect of array topology on the source localisation accuracy.

4.7 Chapter summary

In this chapter, the relative distance cues including relative RIR time delays, attenuation, the clarity feature and the averaged MSC values are extracted from the recorded speech signals and the RIRs at unknown microphone locations across a room of a known geometry. The extracted distance cues are then applied within a surface fitting algorithm to localise the source in the room. It is concluded that the clarity feature can localise the source accurately with no time alignment or synchronisation required and it only requires one microphone at each location. The time delay cues can be applied when the microphones are synchronised and the attenuation cues work accurately where all the microphones have the same gain. The clarity feature and the MSC feature do not require the assumption of microphone having the same gain or being synchronised but they have other limitations as discussed. In this chapter, it was also shown that 2D source localisation can be applied for multi-talk detection which is investigated further in Chapter 6. The proposed clustered surface fitting source localisation method is shown to yield better results compared with blind use of all microphones in the array.

5 Clustered early and late dereverberation

5.1 INTRODUCTION

Multi-channel dereverberation is a well-studied topic in the signal and speech processing research field as it is an important block in applications such as speech diarisation, video-conferencing and meetings [46]. State of the art multi-channel dereverberation methods are usually targeting scenarios with some prior information about the microphone array structure [104], [105], [106] or source signal [107] and require available training data [108] and these are therefore not directly applicable to ad-hoc scenarios where the array topology is unknown or potentially changeable and hence the training data scenarios might not match the application scenario..

Some recent research has proposed a dereverberation frameworks for ad-hoc arrays but the experimental setups are confined to ad-hoc placement of arrays of known geometry and a limited number of microphones [22] and the applied methods are basic beamforming techniques. Although it is claimed that the state of the art speech enhancement methods can be applied to ad-hoc arrays [1] the clear instruction for modifying and adapting these methods to the ad-hoc arrays such as obtaining the steering vector is not straightforward or even possible.

More advanced multi-channel dereverberation methods such as Linear Prediction (LP)-based methods rely on the fact that in reverberant environments the LP residual contains the original excitation source signals containing period peaks during voiced speech, as generated by the talker, followed by several echoed versions of the excitation (echoed peaks) due to the reverberation. In [109], it is shown that spatially averaged LP coefficients derived from microphone array recordings of reverberant speech are much closer to the clean speech signal LP coefficients than the LP coefficients derived for reverberant speech signal recorded at a single point in space for a given room. It is not clarified how far microphones can be located or what happens if the microphone array is an ad-hoc array, therefore this method is not applicable to the arrays of unknown geometry potentially distributed within a large room without required modifications.

This chapter introduces and experimentally evaluates a two-stage early and late dereverberation method for ad-hoc arrays inspired by a leading known geometry microphone array dereverberation method [110] (WPE and MVDR), reviewed and examined in the REVERB challenge [111] and other recent single channel speech enhancement methods [112] that utilise delayed linear prediction. Finding the issues with the context mismatch and unknown information about the array (e.g. relative time delays and phase differences) and overcoming them in a feasible and reasonable manner is the goal of this paper. The main limitation of the existing methods (e.g. Weighted Prediction Error and MVDR beamformers) is requiring the knowledge of the microphone array structure [22] and the recording setup (i.e. Angle of Arrival) [113]. This chapter focuses on the dereverberation of ad-hoc omni-directional microphones similar to the scenarios investigated in [114].

The main contributions of this chapter include

- Proposing a novel multi-channel dereverberation method for the ad-hoc arrays where the microphones can be located meters away from each other and the geometrical configuration of the array is unknown.
- Proposing a clustered multi-channel dereverberationa and speech enhancement approach.
- Introducing the spatial multi-channel linear prediction for ad-hoc microphones
- Introducing the kurtosis of the LP residual signal for microphone clustering

Publications arising from the contributions of this chapter include

- S. Pasha & C. H. Ritz, "Clustered multi-channel dereverberation for ad hoc microphone arrays," in Proceedings of APSIPA Annual Summit and Conference 2015, 2015, pp. 274-278.
- S. Pasha & C. H. Ritz, Y. X. Zou “Spatial multi-channel linear prediction analysis for dereverberation of ad-hoc microphone arrays”, APSIAP 2017 [Under revision]

5.2 Clustered dereverberation for Ad-hoc recording

The general target scenario, depicted in Figure 5-1, shows a few recording devices (nodes) such as laptops, iPads and smartphones, with different number of channels and arbitrary structures, randomly distributed in an unknown reverberant environment. (e.g. a lecture room). In his paper, a node refers to any recording device of any structure and number of channels at an unknown location.

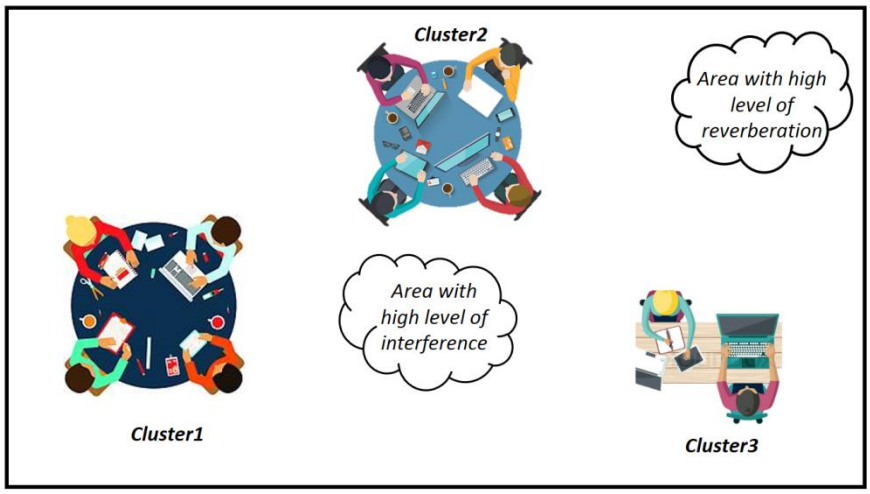


Figure 5-1: Recording by an ad-hoc microphone array

The reverberant signal recorded by microphone *m* is represented as:

$$x_m(n) = h_m(n) * s(n) + v(n) \ , \ 1 < m < M \tag{5-1}$$

where $h_m(n)$ is the Room Impulse Response (RIR) at microphone *m* location and *M* is the total number of the microphones in the room

It is assumed that the position of the microphones and the sources remains fixed

$$h_m(n) = [h_0, h_1, \dots, h_{L-1}]^T, \tag{5-2}$$

during an utterance of speech so $h(n)$ does not change. Therefore, the recorded time domain signals can be represented in a vector form as:

$$\begin{bmatrix} x_1(n) \\ \dots \\ x_M(n) \end{bmatrix} = \begin{bmatrix} h_1(n) \\ \dots \\ h_M(n) \end{bmatrix} * s(n) + \begin{bmatrix} v(n) \\ \dots \\ v(n) \end{bmatrix} \tag{5-3}$$

Although the equation above allows more than one talker in the room however, it is assumed that there is only one active speaker during each time frame. The goal of the clustered dereverberation is to find a subset of channels, $\mathrm{C}(n) = [x_1(n), \dots, x_c(n)]^T$, where $c < M$ and T represent the matrix transpose, such that the output obtained by applying the multichannel dereverberation on C, has less reverberation than is achieved when blindly using all the channels in the array. In order to achieve this, it is necessary to cluster the microphones based on some extracted discriminative feature [5] that reflects the signal reverberation level [25] and pick the cluster with less reverberation level for the multi-channel dereverberation process.

The vector of the recorded signals from 5-3 is

$$\boldsymbol{y}(n) = \begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix} \qquad 5\text{-}4$$

The objective of speech enhancement with ad-hoc arrays is retrieving the best estimate of $s(n)$ [115] by utilising the reverberant recordings ($\boldsymbol{y}(n)$). This can be done blindly through utilising all the microphones ($y(n)$) regardless of their relative distances to the sourceor by utilising a sub-set of microphones ($\boldsymbol{C}$, a subset of $y(n)$ located closer to the source) [22] such that:

$$MR(\boldsymbol{C}(n)) > MR(y(n)) \qquad 5\text{-}5$$

where MR is some dereverberation performance measurement.

5.3 The base-line Spatio-Temporal averaging method

The Spatiotemporal Averaging method for Enhancement of Reverberant Speech (SMERSH), based on the Auto-regressive modelling of the reverberant speech signal [104] is adapted to the ad-hoc array in this section as the base-line method.

5.3.1 Spatial averaging and the AR coefficients

The goal of Auto-Regressive (AR) dereverberation is to estimate $\mathbf{a} = \{a_1, \dots, a_p\}$, and $e_s(n)$ by utilising $\mathbf{x} = \{x_1, \dots, x_M\}$, from (5-3) where p is the LP order and M is the number of microphones in the array.

It is suggested that in the context of compact microphone arrays, spatial averaging of the Auto-Regressive (AR) coefficients such as short term LPC over reverberant channels converge to the LP coefficients of the clean source signal [105], [109]. Although this idea is only proposed for the compact microphone arrays of known geometries, herein it is modified (in terms of time alignment) and adapted to the ad-hoc arrays of arbitrary-random geometries [68].

The time delays between the channels can be found by the cross-correlation method:

$$del_m = argmax\ (\sum_{d=-\infty}^{+\infty} x_m(d) * x_{ref}(n-d)). \quad 5\text{-}6$$

where $*$ denotes the autocorrelation in the time domain. Having obtained the time delays between the channels, the time-aligned signals according to some arbitrary reference channel (x_{ref}) is $y_{time-aligned}(n) = [x_1(n-del_1), \dots, x_M(n-del_M)]^T$. The delay-and-sum (DSB)beamformed signal is then calculated as

$$\hat{x}_{DSB}(n) = \frac{\sum_{m=1}^{M} x_m(n-d_m))}{M}. \quad 5\text{-}7$$

In order to calculate the LPC coefficients the auto correlation method is applied in this research [116], [117] and the coefficients are represented as

$$\mathrm{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_P \end{bmatrix} \quad 5\text{-}8$$

where *P* is the LP order. Utilising b from 5-8 the residual signal $\hat{e}(n)$ is obtained by

$$\hat{e}_{DSB}(n) = \hat{x}_{DSB}(n) - \sum_{k=1}^{p} b_k \hat{x}_{DSB}(n-k). \quad 5\text{-}9$$

Although this residual signal is obtained by analysing the beamformed signals, it still contains reverberation which is further suppressed by temporal averaging between consecutive larynx cycles [104].

5.3.2 Temporal averaging for residual dereverberation

It is observed that for the reverberant speech signals modelled by the LP filter, reverberation distorts the residual signal [118]. In this research the dereverberation of the residual signals is obtained by temporal averaging of the recorded residuals

between Glottal Closure Instants (GCI) by the proposed weighted filter proposed in [104].

The residual signal from 5-9 contains reverberation [25] which should be removed before being utilised for the signal reconstruction. In order to dereverberate the residuals, it is important to detect the original peaks (GCIs) generated by the excitation signal and suppress the other echoed peaks (generated by reverberation). The following filter is applied to temporally average the residual signal and cancel the residual reverberation:

$$\hat{e}(n) = (I - T)\hat{e}_{DSB}(n) + \frac{1}{2\tau}\sum_{i=-\tau}^{\tau} T\hat{e}_{DSB}(n+i) \qquad 5\text{-}10$$

where *I* is the identity matrix and W is the time-domain Tukey window defined as:

$$T = \begin{cases} 0.5 + 0.5\cos\left(\dfrac{2\pi u}{\beta(l-1)} - \pi\right) & u < \dfrac{\beta l}{2} \\ 0.5 + 0.5\cos\left(\dfrac{2\pi}{\beta} - \dfrac{2\pi u}{\beta(l-1)} - \pi\right) & u > l - \dfrac{\beta l}{2} - 1 \\ 1 & otherwise \end{cases} \qquad 5\text{-}11$$

l is the length of one larynx-cycle (the number of samples between two consecutive glottal closure instances) and β is the taper ratio of the window ($\beta = 0.3$ in this research). The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) as in [119] is applied in order to detect the GCIs and *l*. assuming that G=$\{gci_1, .., gci_L\}$ where gci_1 is the first GCI and gci_L is the last GCI, *l* for each filter (the length of the filter changes throughout the speech signal as the distance between GCIs changes) is

$$l = gci_{i+1} - gci_i. \qquad 5\text{-}12$$

Figure 5-2 investigates the effect of β from 5-11 on the residual dereverberation performance measured by the kurtosis of the LP residual signal. For *β=0* the designed filter is a rectangular window of length *L* and for *β=1* the designed filter is a Hann window of length l.

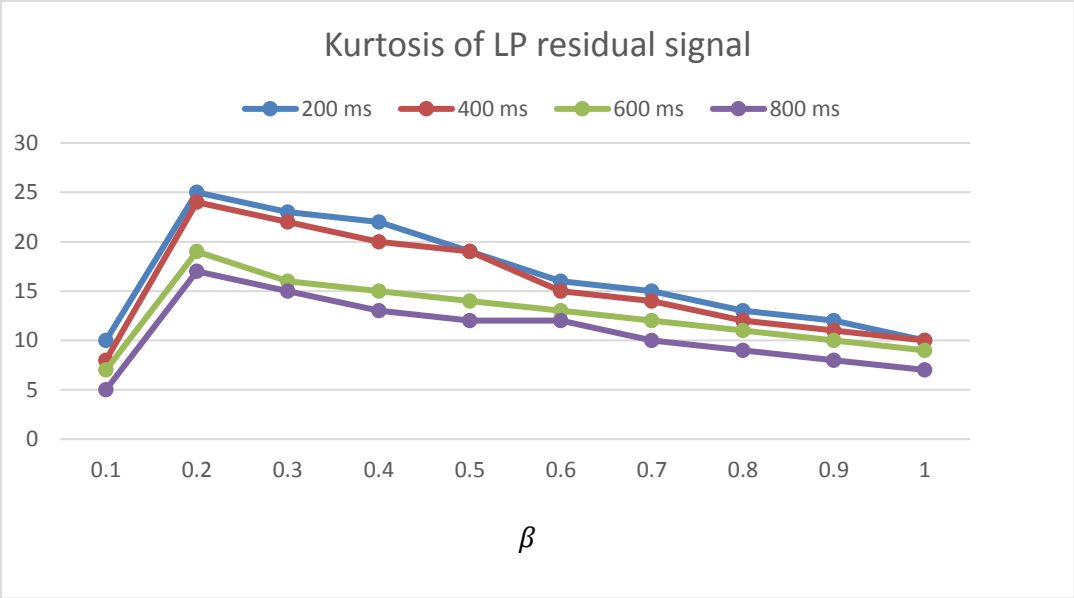


Figure 5-2: Effect of β on the residual dereverberation performance for different reverberation times

It is concluded that β= *0.2* and *0.3* yield the highest residual dereverberation performance. While low values of β or very large values (i.e. 1) cannot suppress the peaks between the GCIs effectively.

5.4 The proposed short and long-term LP residual dereverberation

The proposed dereverberation method consists of the prediction and the removal of the short-term and the long-term reverberation components from the residual signals. Depending on the reverberation time, which is a function of the room geometry and acoustics, reverberation can be categorised into two main categories [20]: Short time reverberation (early echoes) and long term reverberation (late echoes). Breaking $h_m(n)$ into two parts (early and late), the recorded signals from (5-1) can be presented as

$$x_m(n) = \sum_{n=1}^{D-1} s_j(n) * h_{j,m}(n) + \sum_{n=D}^{L-1} s_j(n) * h_{j,m}(n). \qquad 5\text{-}13$$

The long-term effect of reverberation causes the long-term time correlation of the reverberant speech that is exploited to estimate the late reverberation components in

long term dereverberation methods such as the Weighted Prediction Error (WPE) algorithm [120], [121].

There is no clear definition for the short-term and the long-term reverberation but typically echoes received within 80ms after the direct path signal arrival, are labelled as short time echoes [122] and the rest up to a certain delay are the long term reverberation. For a sample room impulse response $h(n)$ the early and late echoes are generated by convolving the source clean signal with a train of pulses:

$$h(n) = \sum_{d} a_d \delta(n-d) \tag{5-14}$$

For small values of d (e.g. smaller than$80\text{ms} \times \text{f}_s\text{kHz}$) [123], [124] the reverberation is considered early and it can be modelled as

$$h_{early}(n) = \sum_{d=0}^{80ms\times fs} a_d \delta(n-d) \tag{5-15}$$

And for higher values of d, the echo is considered long term reverberation or late echoes

$$h_{late}(n) = \sum_{d=80ms\times fs}^{\infty} a_d \delta(n-d) \tag{5-16}$$

however, based on the setups (reverberation time and the room dimensions) these boundaries might vary (e.g. 96ms for short term and up to 1280ms for long-term [113]).

5.4.1 Short-term dereverberation through spatial multi-channel LP

The short term reverberation is the set of echoes that occur within a short delay (e.g. 80ms) after the direct path signal and removing this type of echoes might lead to the loss of some original speech components. In this chapter the spatial LP is proposed as the modified LP analysis tailored for the ad-hoc scenarios. The spatial LP is proposed for the pre-whitening task [20], [125].assuming that channel m recording is represented by $x_m(n)$

$$x_m(n) = \sum_{k=1}^{P_{short}} b_{m,k} x_m(n-k) + e_m(n). \tag{5-17}$$

The spatial multi-channel LP coefficients are obtained by calculating the autocorrelation function and the kurtosis of the standard single channel LP residual signals (β_m) for each channel separately

$$r_m(c) = E\big(e_m(n)e_m(n+c)\big), \qquad c = 0,1,2,\dots \tag{5-18}$$

$$\begin{bmatrix} b_1 \\ \vdots \\ b_{P_{short}} \end{bmatrix} = \begin{bmatrix} \bar{r}_m(0) & \cdots & \bar{r}_m(P_{short}-1) \\ \vdots & \ddots & \vdots \\ \bar{r}_m(P_{short}-1) & \cdots & \bar{r}_m(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_m(1) \\ \vdots \\ \bar{r}_m(P_{short}) \end{bmatrix} \tag{5-19}$$

where r_m is the autocorrelation function and β_m is the kurtosis of the LP residual signal applied as the distance cue [25]. The residual and the reconstructed signal ,$\bar{x}(n)$, are obtained by

$$e(n) = x(n) - \sum_{k=1}^{P_{short}} b_k x(n-k) \tag{5-20}$$

and

$$\bar{x}(n) = \sum_{k=1}^{P_{short}} b_k x(n-k) \tag{5-21}$$

5.4.1.1 Spatial Multi channel Linear prediction

The weighted average auto-correlation function $\bar{r}(c)$ is obtained for M channels as

$$\bar{r}(c) = \frac{1}{M} \times \sum_{m=1}^{M} r_m(c). \tag{5-22}$$

As it is inferred from (5-22) all the M autocorrelation functions are equally weighted in the averaging process. The averaged autocorrelation function can be written in a more general form of a weighted average autocorrelation ($\bar{r}_w(c)$), in order to take into account the source to microphone distances for each microphone. Assuming that the applied weights are $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_M\}$ the weighted average autocorrelation function is calculated as

$$\bar{r}_w(c) = \frac{1}{\sum_{m=1}^{M} \beta_m} \times \sum_{m=1}^{M} \beta_m r_m(c) \tag{5-23}$$

where β_m is the weights to $r_m(c)$. And the filter coefficients are obtained by the Yule–Walker method.

$$\mathbf{w}_s = \begin{bmatrix} \bar{r}_w(0) & \cdots & \bar{r}_w(P_{short}-1) \\ \vdots & \ddots & \vdots \\ \bar{r}_w(P_{short}-1) & \cdots & \bar{r}_w(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_w(1) \\ \vdots \\ \bar{r}_w(P_{short}) \end{bmatrix} \quad 5\text{-}24$$

and the pre-whitened signal is

$$\tilde{e}_m(n) = x_m(n) - \sum_{k=1}^{P_{short}} w_{s,k} x_m(n-k). \quad 5\text{-}25$$

where $w_s = \{w_{s,1}, \dots, w_{s,P_{short}}\}$.

Assuming that the source to microphone distances for all the M microphones are $\{q_{1,s}, \dots, q_{M,s}\}$, the ideal distance weights are $\mathbf{q} = \{\frac{1}{q_{1,s}}, \dots, \frac{1}{q_{M,s}}\}$. It is observed that using $\mathbf{q}$ as the weights significantly improves the autocorrelation function estimation compared with the proposed method in [125]. In other words applying the inverse of the source to microphone distances as the weight estimates the clean source signal autocorrelation function more accurately than (5-22). However the knowledge of the source to microphone distances ($\mathbf{q}$) is not usually available or retrievable and using $\mathbf{q}$ is not practical for the ad-hoc scenarios.

Figure 5-3 illustrates the improvement made by the spatial multi-channel LP in the estimation of the clean LP coefficients for 250 random ad-hoc scenarios with 2 to 6 microphones. Itakura error [105] is applied as the measurement. Er_{w_s,a_s} is the Itakura distance between the clean LP coefficiens ($a_{s,k}$) and the estimated coefficients ($w_{s,k}$).

$$Er_{w_s,a_s} = \left| \sum_{k=1}^{P_{short}} \left(\frac{w_{s,k}}{a_{s,k}} - \log \frac{w_{s,k}}{a_{s,k}} - 1 \right) \right| \quad 5\text{-}26$$

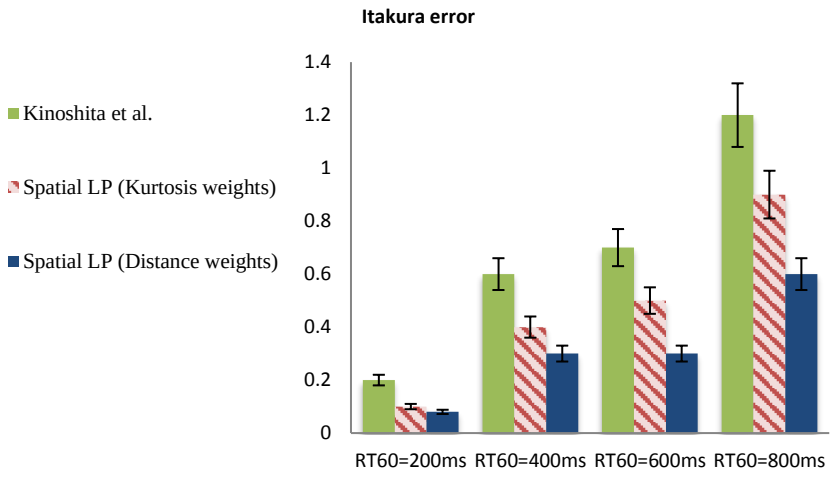


Figure 5-3: Effect of the spatial multi-channel linear prediction on the Itakura error

5.4.2 Long term dereverberation thorugh delayed LP

The main part of the long term linear prediction method consists of robust blind deconvolution based on long-term linear prediction, which tries to estimate the late echoes. The long-term effect of reflections caused by reverberation generates the long-term time correlation of the reverberated speech that can be exploited to estimate the late reverberation components using the long term linear prediction algorithm [20]. As opposed to multi-channel late dereverberation methods such as [125] in this research pre-whitening based on averaging the autocorrelation functions and obtaining the LP coefficients is not applied. It is suggested that pre-whitening before the dereverberation is required as a primary step however the applied pre-whitening method is proposed for short term dereverberation which is performed by early dereverberation.

Long term dereverberation is achieved using a delayed long term linear prediction filter [125] as described by:

$$\bar{x}(n) = \sum_{i=1}^{P_{long}} w_{long_i}\left(n - i - D_{long}\right) + \bar{e}(n) \qquad 5\text{-}27$$

where D_{long} represents the delay of LP filtering which for long term dereverberation application is considered between 224ms to 1280ms [113] in the literature as it needs to deal with late echoes and P_{long} is the long term dereverberation filter length. In this contribution, D_{long} from (5-27) is estimated as

$$D = \underset{\tau}{\operatorname{argmax}} \left(\sum_{\tau=\tau_{min}}^{\tau=\tau_{max}} \tilde{e}_m(n) * \tilde{e}_m(n+\tau) \right) \quad 5\text{-}28$$

Similar to standard linear prediction, the prediction coefficients ($\boldsymbol{W_{long}}$=$[w_1, w_2, \dots, w_{P_{long}}]^T$) are obtained by:

$$\left(E\{\bar{\boldsymbol{e}}(n - D_{long})\bar{\boldsymbol{x}}^{\boldsymbol{T}}(n - D_{long})\}\right)\boldsymbol{W_{long}} = E\{\bar{\boldsymbol{e}}(n - D_{long})\bar{x}(n)\} \quad 5\text{-}29$$

$$\boldsymbol{W_{long}} = (E\{\bar{\boldsymbol{e}}(n - D_{long})\bar{\boldsymbol{x}}^{\boldsymbol{T}}(n - D_{long})\})^{-1} E\{\bar{\boldsymbol{e}}(n - D_{long})\bar{e}(n)\} \quad 5\text{-}30$$

The dereverberated signal can be obtained by filtering the reverberant residuals

$$\tilde{e}(n) = \bar{e}(n) - \sum_{i=1}^{P_{long}} w_{long_i} * \bar{e}(n) \quad 5\text{-}31$$

(5-31) can be rewritten for an ad-hoc array of *M* randomly located microphones in a reverberant environment as

$$\begin{bmatrix} \tilde{e}_1(n) \\ \vdots \\ \tilde{e}_M(n) \end{bmatrix} = \begin{bmatrix} \bar{e_1}(n) \\ \vdots \\ \overline{e_M}(n) \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{P_{long}} w_{long_i} * \bar{e_1}(n) \\ \vdots \\ \sum_{i=1}^{P_{long}} w_{long_i} * \overline{e_M}(n) \end{bmatrix} \quad 5\text{-}32$$

The reconstructed speech signals are obtained by applying the synthesis LP filter on the dereverberated residuals as:

$$\tilde{S} = \begin{bmatrix} \tilde{s}_1(n) \\ \vdots \\ \tilde{s}_M(n) \end{bmatrix} \quad 5\text{-}33$$

The single channel beamformed signal, if required, is obtained by applying (5-6) and (5-7) to $\tilde{S} = \{\tilde{s}_1(n), \dots, \tilde{s}_M(n)\}$. The effect of the filter length (P_{long} from 5-27) on the residual dereverberation performance is investigated in Figure 5-4 and it is concluded that longer filter can remove the late echoes more successfully.

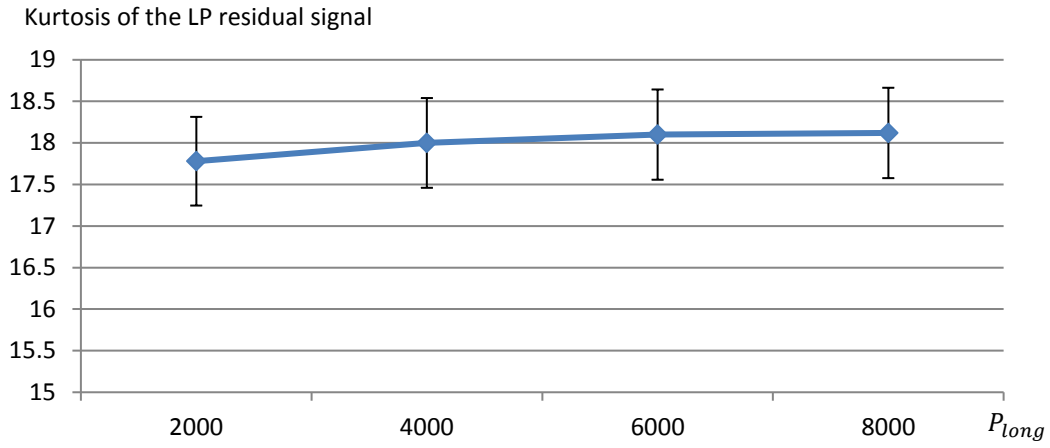


Figure 5-4: Effect of the delayed LP filter length on the late residual dereverberation

5.5 Clustered multi-channel dereverberation

In order to improve the dereverberation process this clustered method excludes the highly reverberant microphone signals from the array and only applies the dereverberation process to a smaller, less reverberant subset of microphones [68] (Figure 5-5). Similar to [22] where it is suggested to pick the best node based on a predefined criteria and apply the dereverberation method only on 3 channels with the highest input quality, herein this idea is extended to choose a flexible (in terms of the number of the channels) cluster of microphones that yield the highest dereverberation performance.

Figure 5-5: Proposed Combined method

As discussed in [5] working with raw audio and speech signals is inefficient and computationally intensive, therefore the first step of any clustering algorithm is extracting discriminative features. The extracted features from the microphones in the ad-hoc array are represented by vector $K = [k_1, \dots, k_M]^T$, where there is one feature (value) derived for each microphone. As the kurtosis of the LP residual signal

is an indicator of the source to microphone distance and reverberation level and also is independent of the source energy level (Kurtosis advantage over amplitude attenuation), herein the kurtosis of the LP residual signal is introduced for microphone clustering for dereverberation applications. As the proposed method of this research is based on linear prediction and obtaining residual signals, calculation of the discriminative feature (Kurtosis of the LP residuals) does not add any extra computation cost to the overall system. The following proves that the kurtosis of the LP residual signal calculated over s short time frame of length T_f samples is independent of the source energy level and microphone gains:

$$\beta_j = \frac{E\{e_j^4(n)\}}{E^2\{e_j^2(n)\}} - 3 \qquad 5\text{-}34$$

$$= \frac{\frac{1}{T_f}\sum_{n=1}^{T_f}(e_j(n) - \bar{e}_j(n))^4}{(\frac{1}{T_f}\sum_{n=1}^{T_f}(e_j(n) - \bar{e}_j(n)^2))^2} - 3 \qquad 5\text{-}35$$

where $\bar{e}_j(n)$ denotes the average value of $e_j(n)$ across T_f. Assuming that for another speech source such that $S_j(n) = \propto S_i(n)$ (or equally a different microphone gain), consequently $e_j(n) = \propto e_i(n)$, therefore:

$$\beta_j = \frac{E\{e_j^4(n)\}}{E^2\{e_j^2(n)\}} - 3 \qquad 5\text{-}36$$

$$= \frac{\frac{1}{T_f}\sum_{n=1}^{T_f}(e_j(n) - \bar{e}_j(n))^4}{(\frac{1}{T_f}\sum_{n=1}^{n}(e_j(n) - \bar{e}_j(n)^2))^2} - 3 \qquad 5\text{-}37$$

$$= \frac{\frac{1}{T_f}\sum_{n=1}^{T_f} \propto^4 (e_i(n) - \bar{e}_i(n))^4}{(\frac{1}{T_f}\sum_{n=1}^{T_f}(\propto^2 (e_i(n) - \bar{e}_i(n)^2))^2} - 3 \qquad 5\text{-}38$$

$$\frac{\frac{\propto^4}{T_f}\sum_{n=1}^{n}(e_i(n) - \bar{e}_i(n))^4}{(\frac{\propto^4}{T_f}\sum_{n=1}^{n}((e_i(n) - \bar{e}_i(n)^2))^2} - 3 = \beta_j(n) \qquad 5\text{-}39$$

The kurtosis values are calculated for 10 different microphone gains (which is equivalent to different source energy levels)(Figure 5-6). It is observed that the kurtosis of the LP residual signal is robust against source energy levels and different microphones gains. These characteristics are especially important in the context of the ad-hoc array where the talkers might use their own recording devices and the microphone gains are not the same for all the recording devices. In this research the kurtosis of the LP residual signal is utilized as the microphone clustering feature in order to cluster the microphones into two (k=2) clusters by the Kmeans method.

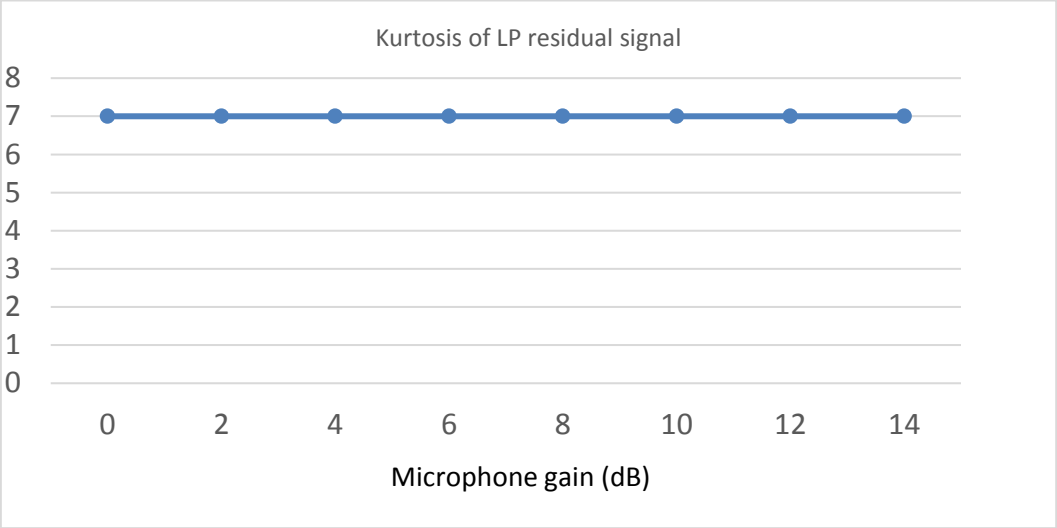


Figure 5-6: Kurtosis versus microphone gains (dB) calculated for 500ms frames

Table 5-1 compares the kurtosis of the LP residual signal with other distance cues such as signal power, TOA and TDOA.

Table 5-1: Advantages of the kurtosis feature

Gain independent	Limitation of the signal power
Not affected by the time delay between the microphones	Limitation of the TDOA and TOA features
Does not require binaural recording	Limitation of the coherence features

5.6 Results and Evaluation

In this section the proposed method is compared with the other dereverberation methods including the Weighted Prediction Error (WPE) [126] and Minimum Variance Distortionless Response (MVDR) beamformer, the SMERSH algorithm [104], [105]and the kurtosis maximisation method [20]. Results are obtained for different reverberation times and noise types to achieve a reliable conclusion. In this section two experiments have been implemented to evaluate the objectives of the proposed approach. The first experiment evaluates the proposed method's effectiveness in speech enhancement and compares it with the multichannel dereverberation methods from the Reverb challenge [111]. The second experiment compares the clustered dereverberation approach with the blind use of all the microphones and investigates the effect of the clustered dereverberation where highly distorted channels, estimated by kurtosis of LP residual signal are excluded from the dereverberation process.

Table 5-2: Experimental configuration

Parameter	Applied values
f_s	16kHz
RT_{60}	200ms,400ms,600ms,800ms
Room size	10m,8m,3m
Number of microphones	2 to 8
Noise	White noise, Babble noise
SNR	10,20,30dB and clean signals
Discriminative microphone clustering feature	Kurtosis of LP residual signal
P_{short}	20
P_{long}	6000
D_{long}	$RT_{60}\ (s) \times f_s$ (proposed)
Kurtosis maximization filter order	100
τ_{min}	200 (samples)
τ_{max}	1600 (samples)

5.6.1 Experiment1: Dereverberation performance

The configuration of Table 5-2 is applied for the experiments in order to evaluate the performance of the proposed method and to compare its results with the state of the art multi-channel speech enhancement methods. Clean signals from IEEE corpus and noisy signals from the NOIZEUS database [84] are utilised to generate reverberant noisy speech signals at recorded arbitrary locations (5-1) by simulating the RIRs.

The comparison of the proposed method and the state of the art methods is presented in Figure 5-7 and Figure 5-8, for ten sentences read by male and female talkers in a $10m \times 8m \times 3m$ room with RT_{60} of 200ms, 400ms, 600ms and 800ms. The Perceptual Evaluation of Speech Quality (PESQ) (Minimum=1, annoying and Maximum=5, clean), Direct to Reverberant Ratio (DRR) and the Cepstral Distance (CD) [127] are calculated as the quality measurement and dereverberation performance measurements. The results represent the averaged measurements over 250 experiments (5 set of speech files at 50 random setups) for each reverberation time and the applied method. The reverberant speech files are randomly chosen from different SNR values and noise types as in Table 5-2. It is concluded that the proposed method outperforms the state of the art WPE+MVDR method in short reverberation times but for reverberation times longer than 400ms the WPE+MVDR is more successful. The kurtosis maximisation method outperforms the proposed method in terms of kurtosis of the LP residual values but distorts the signal quality significantly.

$$DRR\ (dB) = 20 \times \log_{10}\left(\frac{|s(n)|}{|\tilde{s}(n) - s(n)|}\right) \qquad 5\text{-}40$$

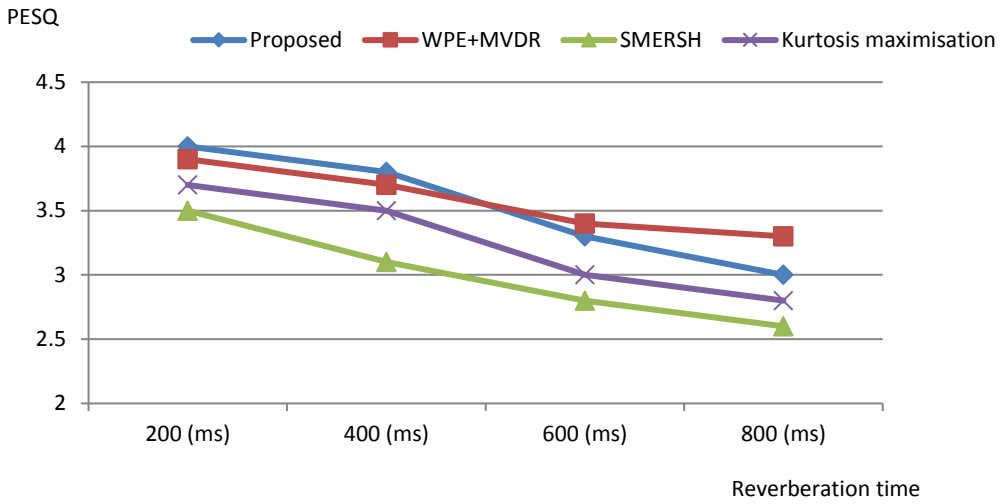


Figure 5-7: PESQ for different reverberation times

Figure 5-8 compares the proposed two stage dereverberation method with the baseline SMERSH and the state of the art WPE+MVDR and the kurtosis maximisation method. It is concluded that for short reverberation times (i.e. less than 400ms) the proposed method outperforms the WPE+MVDR method.

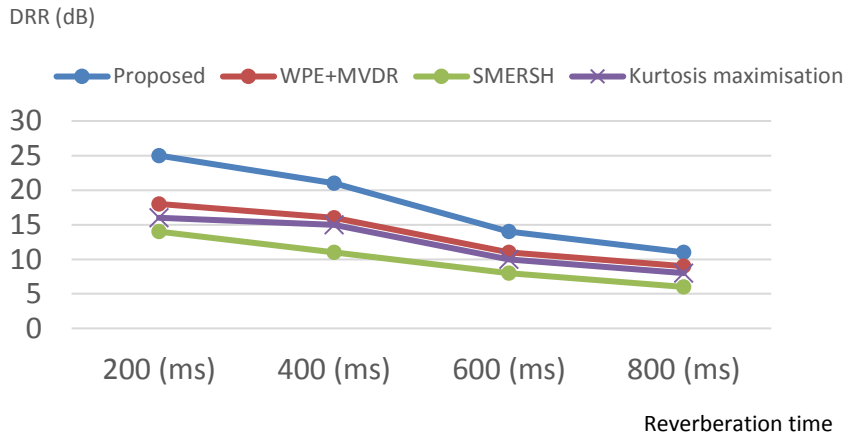


Figure 5-8: Dereverberation performance (SNR=10dB)

It is concluded that the proposed method outperforms the state of the art multi-channel-dereverberation methods when applied to the ad-hoc arrays. The results obtained by the experiment s of this chapter are compatible by similar experimental studies of ad-hoc microphones which show MVDR beamformer cannot be applied to ad-hoc microphone arrays of unknown structures [115].

Figure 5-9 investigates the effect of the proposed adaptive D_{long} on the dereverberation performance. It is suggested that adapting the long term dereverberation delay value, proportional to the reverberation time outperforms the fixed delays including the values suggested in [20].

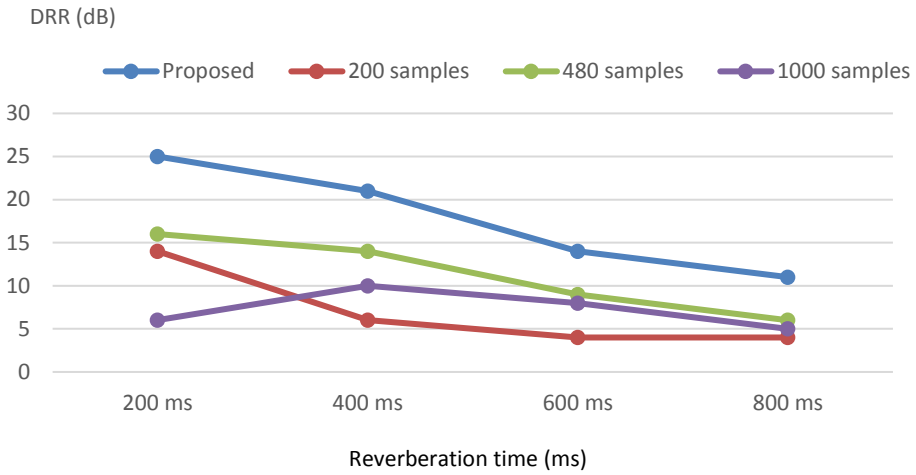


Figure 5-9: Reverberation performance for different D_{long} values and reverberation times

5.6.2 Experiment2: Clustered dereverberation

Figure 5-10 shows a formed cluster located closer to the source (estimated by the kurtosis values), four microphones are labelled as close and four microphones are labelled as far [68] and the improved dereverberation performance is obtained by exclusively applying the proposed method to the chosen subset (Figure 5-10). Figure 5-11 shows the comparison between the blind use of all microphones and the proposed clustered method. It is concluded that for long reverberation times (i.e. longer than 400ms) choosing a subset of microphones closer to the source can significantly improve the dereverberation performance. The size of the chosen cluster depends on the distribution of the microphones around the source location and can vary from 1 to *M* from (5-1).

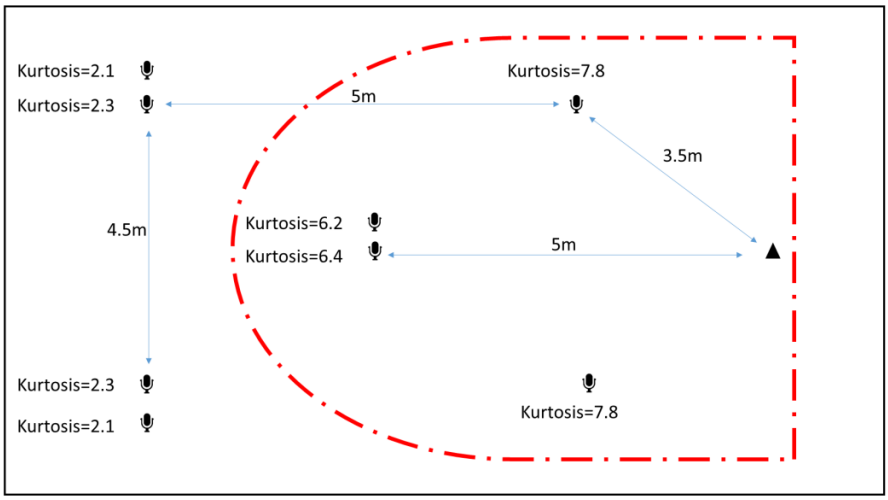


Figure 5-10: Sample clustered ad-hoc microphones, the black triangle represents the source location

Figure 5-11 investigates the effect of the clustered approach on the dereverberation performance of the base-line and the proposed method of this paper. It is observed that excluding microphones located far from the source, which are usually highly distorted improves the dereverberation performance.

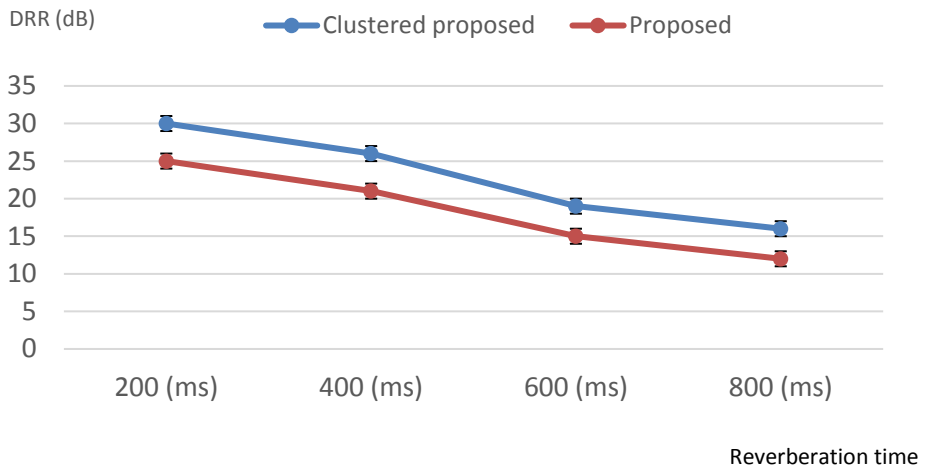


Figure 5-11: Effect of clustering on the dereverberation performance for different reverberation times

5.7 Chapter summary and conclusion:

In this chapter a novel clustered dereverberation method for ad-hoc arrays, where the microphone array geometry is unknown is proposed and successfully tested. The proposed spatial multi-channel LP which takes into account the spatial distances between the microphones and the source is applied for the pre-whitening phase. The delayed LPC analysis is applied to remove the long term reverberation. The standard delayed LPC analysis is modified by choosing the delay value adaptively based on pre-whitened residuals. The overall performance of the system is improved by removing the highly distorted microphones from the array. Results suggest that adaptively choosing the LP analysis delay improves the dereverberation performance.

6 Source counting by ad-hoc microphone arrays

6.1 Introduction

Speaker overlap or multi-talk during the meetings is a significant contributor to error in speaker diarisation [128], source localisation, word counting [129], source separation [130] and speech enhancement [131] applications. Overlaps are problematic for speech and microphone clustering as the overlapped frames (time-segments) contain components that belong to more than one source speaker. Detecting segments of speech that contain more than one source signal and considering them for source separation is one approach to address the issues caused by overlap [130].

Errors caused by speech overlaps and the baseline features for the overlap detection are discussed in [132], [133] and it is suggested that conversational features such as speaker change statistics, can help the speaker diarisation methods over long-term segments with short durations, such as 5 seconds. It has been also previously shown that the detection of the overlapping segment can improve the speech diarisation accuracy for clustering based methods by 15% [134].

Various approaches have been proposed to enhance the recording in the presence of overlapping source(s) [135] but they suffer from limiting requirements such as the prior knowledge of the number of sources [5], the predefined threshold and the clean training data [25].

In this chapter diffuseness estimates are proposed as a robust feature in reverberant environments for overlap and speech activity detection [136] over short time frames (i.e. 20ms to 300ms) when using ad-hoc arrays of unknown arbitrary geometries. Diffuseness and the level of reverberation contain source to microphone distance cues and can be utilised to discriminate sources based on their distances to the microphones. It is also suggested in this chapter that this feature can be applied as an interfering talker detection feature.

In order to estimate the Coherent to Diffuse Ratio (CDR) feature from noisy speech signals, a novel method is proposed in [29]. This method extracts the CDR features from short (20ms-30ms) noisy reverberant speech frames and does not require a training phase. The proposed method is designed for dual-microphone systems and frame-wise processing. The advantage of the CDR features compared with other location and speech activity cues such as signal power [137] is that the CDR values as the ratio of the direct path signal to the reflected signal are independent of the source energy level and do not require time alignment and synchronisation of the signals.

The proposed multi-talk detection approach described in this chapter utilises the estimated CDR features for real time interfering talker detection and source counting using ad-hoc dual microphone nodes where the distance between the microphones is unknown. This contribution also overcomes the limitations of similar real-time methods such as requiring the knowledge of the microphone array structure [138]. Similar to the state-of-the-art source counting methods, herein it is assumed that the sources may overlap in some time-frequency zones however, the proposed method does not require conversational features, long time-frequency frames of overlaps and the statistical parametrisation of the speech sources.

Counting the active participants in a meeting based on the coherence features is also investigated in this chapter. An offline method robust to reverberation and the microphone spacing is proposed and successfully tested.

The main contributions of this chapter include

- Extracting relative distance and interference cues independent of the microphone gains, sampling frequencies and microphones internal time delays for interfering talkers over short time segments.
- Pseudo real time source counting over short time segments for overlapping talkers with no prior information about the microphone arrays structure and the source locations.
- Detecting speakers overlap in the context of ad-hoc arrays

Publications arising from contributions of this chapter are

- S. Pasha, C. Ritz and Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, 2016, pp. 1-6.
- S.Pasha, C. Ritz, Y. X. Zou "Towards real-time source counting by estimation of coherent-to- diffuse ratio estimates from ad-hoc microphone array recordings" Fifth Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA) 2017
- S. Pasha, Jacob Donley and C. Ritz, "Speaker counting and diarisation through analysis of the magnitude squared coherence frequency response for highly reverberant signals" APSIPA 2017 [Under revision]

6.2 CDR calculated for dual channel ad-hoc nodes

In array signal processing, environmental noise [139] is often modelled by the superposition of an infinite number of uncorrelated, spatially distributed noise sources. In applications such as underwater acoustics or radio communication, this model is motivated by the presence of many independent noise and interfering sources around the receiver which create a diffuse noise filed. The most common assumption for the spatial distribution is a sphere centered on the receiver, which corresponds to what is known as a diffuse or spherically isotropic noise field. The spatial coherence function between two omnidirectional sensors in a diffuse noise field is real-valued and given by:

$$C_{coherence} = \frac{\mathrm{SIN}(Kd)}{Kd} = \frac{\mathrm{SIN}(2\pi fd/c)}{2\pi fd/c} \qquad 6\text{-}1$$

For 320 samples (20ms at 16kHz sampling rate) of white Gaussian noise the MSC is calculated and it is shown that the MSC between the two signals are not coherent and the value of MSC is very low (less than 0.4) for the majority of the frequencies.

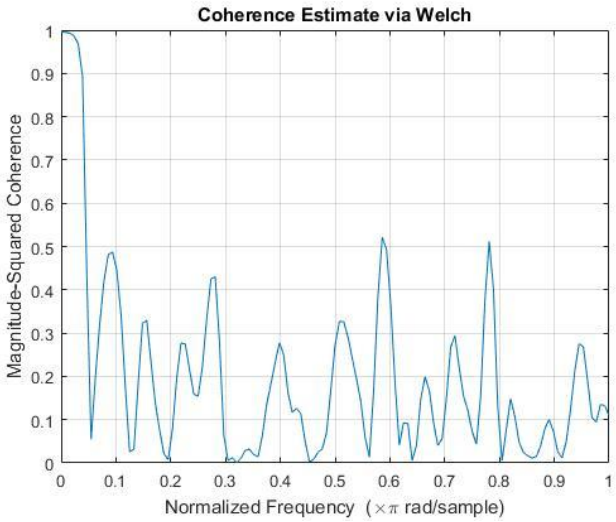


Figure 6-1: MSC calculated for two white Gaussian noise signals recorded by dual channel microphones

where *K* is the wavenumber, *d* is the inter-channel distance and *c* represents the speed of sound. Assuming that *d* is identical for all the nodes, the coherence or the

Coherence to Diffuse Ratio (CDR) can be utilised to detect the presence of a source. This can be done more accurately if the nodes are spread out within the room and very close to the sources.

Coherence is a function of frequency and if two signals are highly coherent the average coherence across all the frequencies (6-1) is a higher value and if two signals are uncorrelated the average coherence across all the frequencies is a lower value.

Assuming there are *N* nodes of dual omini-directional microphones with identical inter-channel distances, *d*, each channel at each node receives a unique reverberant version of the source signal due to its spatial location and Room Impulse Response (RIR):

$$\mathbf{X_n}(t) = S(t) * \mathbf{H_n}(t) + \mathbf{N}(t) \qquad 6\text{-}2$$

where $\mathbf{X_n}(t) = (X_{n1}(t), X_{n2}(t))$ is the recorded signals by the two channels at node *n*, *1<n<N*, $S(t)$ is the clean, anechoic source signal (assuming there is only one active source) and $\mathbf{H_n}(t) = (\, h_{n1}(t), h_{n2}(t))$ is the RIR matrix at the n^{th} node location. $\mathbf{N}(t)$ represents the diffuse noise and the reverberation is modelled by $S(t) * \mathbf{H_n}(t)$.

Coherence is calculated for a dual microphone with a 10cm inter-channel spacing in a clean anechoic environment over 160 samples (20ms at 8kHz sampling rate) and it is shown that in the majority of the frequencies the MSC is equal to 1 (the maximum) (Figure 6-2) as the signals are very similar. The frequencies where the MSC is low are the frequencies that the speaker does not have significant energy (Figure 6-3). The effect of noise and reverberation on the MSC is shown in Figure 6-4.

It is concluded that in an anechoic room with no noise or interference the received signal by the two channels are very similar (one signal is the delayed version of the other channel) and therefore the coherence obtain it maximum value.

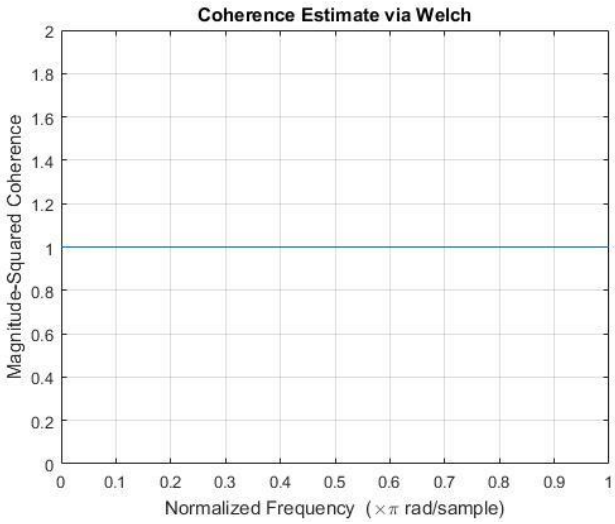


Figure 6-2:MSC between two clean anechoic speech frames (20ms) recorded by a dual channel node (*d*=15cm)

Noise as a non-coherent component distorts the MSC graph and it is observed that some frequencies which are most likely dominated by noise have lower MSC values.

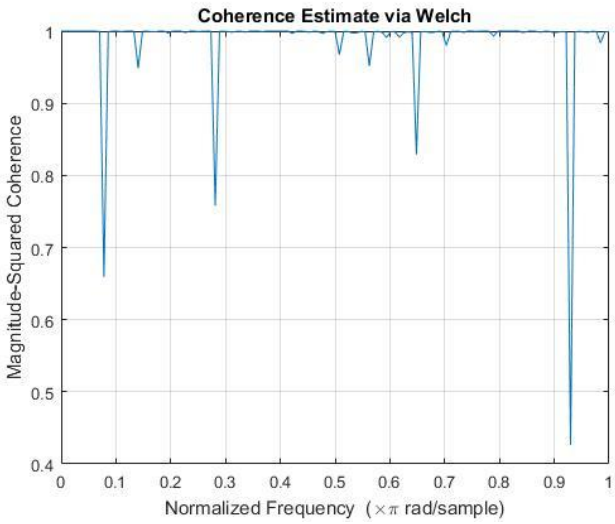


Figure 6-3: MSC between two noisy channels signals of a dual node in an anechoic room (*d*=15cm, SNR=10dB)

Figure 6-4 and Figure 6-5 depict the effect of reverberation and noise as non-coherent components on the coherent MSC values.

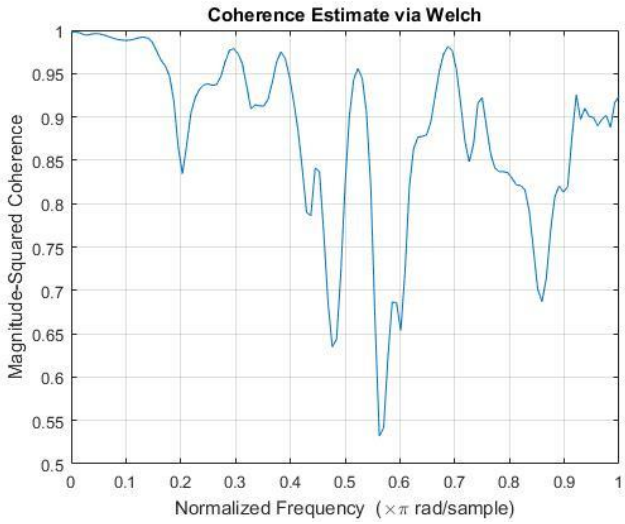


Figure 6-4: MSC between two noisy channels of a dual node in a reverberant room (*d*=15cm, SNR=10dB, RT60=400ms)

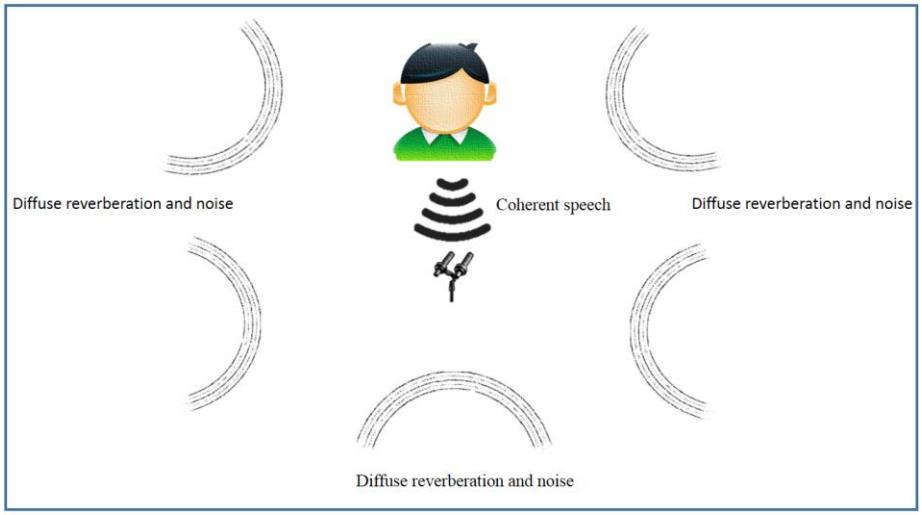


Figure 6-5: Effect of reverberation and noise on CDR values

It is concluded that when the reverberation and the noise are present only the speaker speech frequencies have relatively high MSC values and the frequencies dominated by the non-coherent component obtain low values.

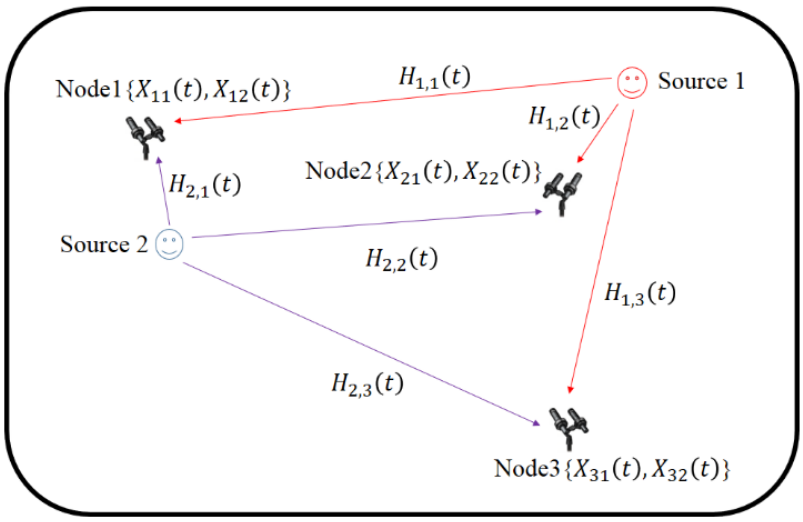


Figure 6-6: Dual node ad-hoc arrays

The RIR between each node and the active source, $\mathbf{H}_n(t)$, (Figure 6-6) is a function of the source to node distances and room geometry and characteristics such as the RT_{60}. If there is more than one simultaneously active source in the room (cross talk) the recorded signals can be represented as:

$$\mathbf{X_n}(t) = \sum_{k=1}^{S} (S_k(t) * \mathbf{H_{k,n}}(t)) + \mathbf{N}(t) \qquad 6\text{-}3$$

where S is the number of simultaneously active sources at time t, and $1<k<S$ is the source index. It is shown that the coherence between the two channels signals, $X_{n,1}(t)$, $X_{n,2}(t)$, at each node is a function of source to node distance, frequency, noise, interference and reverberation level [140]. The calculated coherence cues have been previously used for speech activity detection and it was shown that in dual microphone systems, the inter-channel coherence value is a function of interference level and the distance between the active source and the node [81].It is also shown that there is no need to calculate this measurement using the full length signals and they can be accurately estimated utilising 20ms frames of the noisy speech signals.

Estimated coherence features are applied as distance features to discriminate the microphone nodes located close to an active source. For dual microphone systems and two active sources (Figure 6-6) each channel's signal can be represented as:

$$X_{n1}(t) = \sum_{k=1}^{2}(S_k(t) * h_{k,n1}(t)) + N(t) \quad \text{6-4}$$

$$X_{n2}(t) = \sum_{k=1}^{2}(S_k(t) * h_{k,n2}(t)) + N(t) \quad \text{6-5}$$

The coherence between these two noisy and reverberated signals at node *n* is higher when the active source is closer to the node and is lower when the active source is located far from the node. For instance, in Figure 6-6, node1 is dominated by source2 and node2 is dominated by source1 hence these two nodes have higher coherence features even in cross talk situations whereas node3 is not close to any source and in case of cross talk it receives a mixture of source1 and source2 signals equally which has a low Signal to Interference Ratio (SIR) and coherence feature. MSC is defined as:

$$C_x(t) = \frac{\left|\varphi_{X_{n1}X_{n2}}(t)\right|^2}{\varphi_{X_{n1}X_{n1}}(t)\,\varphi_{X_{n2}X_{n2}}(t)} \quad \text{6-6}$$

where $\varphi_{X_{n1}X_{n2}}(t)$ is the cross power spectra function.

$$CDR_n(l,f) = \frac{C_{u_n}(f) - C_{x_n}(l,f)}{C_{x_n}(l,f) - C_s(l,f)}, \quad \text{6-7}$$

from which we propose the use of the average CDR over the entire frequency band and *L* frames, given by

$$\overline{CDR}_n = \frac{1}{L(f_B - f_0)} \int_{f=f_0}^{f_B} \sum_{l=1}^{L} CDR_n(l,f)\ \mathrm{d}f, \quad \text{6-8}$$

6.3 Estimated CDR as a distance cue

Assuming that *S* simultaneously active sources (6-3) have different angles of arrival at each node, the vector of the angle of arrivals at node *n* from all *S* sources can be represented as:

$$\boldsymbol{\theta_n} = \{\theta_1, \dots, \theta_S\} \quad \text{6-9}$$

Similarly, the distances between the n^{th} node and all *S* simultaneously active sources can be represented in a vector as:

$$\boldsymbol{D_n} = \{D_1, \dots, D_S\} \quad \text{6-10}$$

Both these two vectors are unknown in this research and it is noteworthy that CDR is a function of θ_n, D_n and S. [141]

The long term reverberation caused by $\mathbf{H}_{k,n}(t)$ and $N(t)$ are diffuse as they do not have any specific angle of arrival and they arrive at each node from all directions under the assumption that reverberant sound can be modelled as a mixture of a direct component and a perfectly diffuse reverberation component which are mutually uncorrelated. The only coherent component of (6-3) is the direct path signal from the dominant source to the node which can be modelled mathematically as:

$$X_{n|coherent}(t) = S_k(t) * H_{k,n}(\tau_{nk}) \tag{6-11}$$

where τ_{nk} is the time delay between the source *k* and node *n*. (6-6) can be rewritten in the time-frequency domain as:

$$C_{x_n}(l,f) = \frac{\left|\varphi_{X_{n1}X_{n2}}(l,f)\right|^2}{\varphi_{X_{n1}X_{n1}}(l,f)\,\varphi_{X_{n2}X_{n2}}(l,f)} \tag{6-13}$$

$$C_N(l,f) = \frac{\left|\varphi_{N_1N_2}(l,f)\right|^2}{\varphi_{N_1N_1}(l,f)\,\varphi_{N_2N_2}(l,f)} \tag{6-14}$$

Assuming that the source coherence ($C_s(l,f)$) is equal to 1 the CDR can be calculated as:

$$CDR = \frac{C_N(l,f) - C_x(l,f)}{C_x(l,f) - C_s(l,f)} \tag{6-15}$$

$$\overline{CDR}_n = \frac{1}{L(f_B - f_0)} \int_{f=f_0}^{f_B} \sum_{l=1}^{L} CDR_n(l,f)\ \mathrm{d}f \tag{6-16}$$

The proposed scheme can give an estimate of the Coherent to Diffuse Ratio (CDR) and Direct-to-Reverberant energy Ratio (DRR) since the dominant direct speech can be considered as the coherent signal whereas the diffuse noise and the reverberant-interfering speech forms the diffuse or non-coherent component. This fact is utilised in this research to distinguish the nodes with higher CDR values (more likely located close to an active source e.g. node 1 and node 2 in Figure 6-6) from nodes with lower CDR values (likely located far from active sources e.g. node3 in Figure 6-6).

The relationship between the estimated CDRs and the source to node distance when there is one, two, three and four simultaneously active sources, *S=1,2,3,4* in

(6-3). A scenario with two active sources is depicted in Figure 6-7. The inverse relationship between the source to node distance and the CDR feature is evident from the results shown in Figure 6-8 and Figure 6-10. A 20ms frame speech recording is utilised to calculate the CDR estimate.

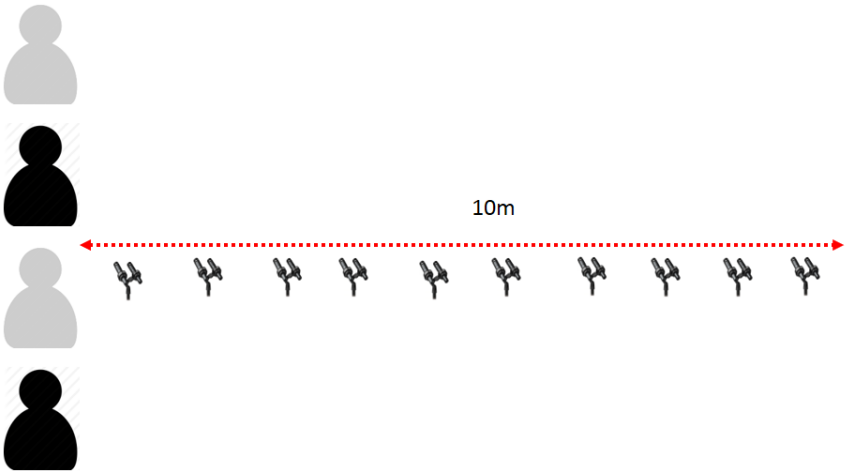


Figure 6-7: Two active sources and a dual node at different distances

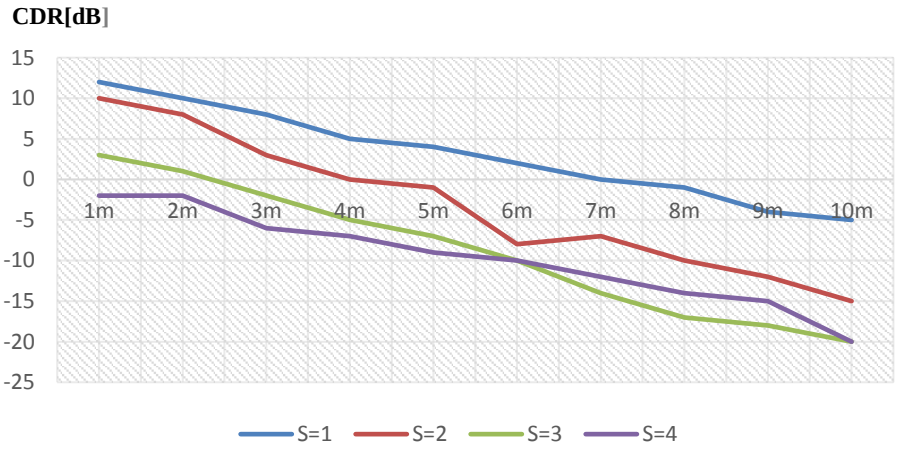


Figure 6-8: The effect of source to node distance on CDR for different number of simultaneously active sources

For a scenario with one node located at the equal distance (2m) form all the four participants in a meeting (Figure 6-9) the CDR values are estimated when *S*, varies from 1 to 4, which, respectively means one, two, three or all four participants are talking simultaneously. The effect of interference on the estimated CDR is shown in Figure 6-8. As the CDR is a ratio of the coherent source signal to the diffuse source

signal it does not vary with the source energy level and is robust against the inconsistency between the sources energy levels [142]. It is observed that the CDR estimate drops with the interference and source to microphone distance. These two observations are exploited for real time source counting and cross talk applications in this chapter.

6.4 Estimated CDR as an interference cue

The following setup is considered to investigate the effect of *S* (the number of simultaneously active sources) on the CDR estimates over 20ms frames.

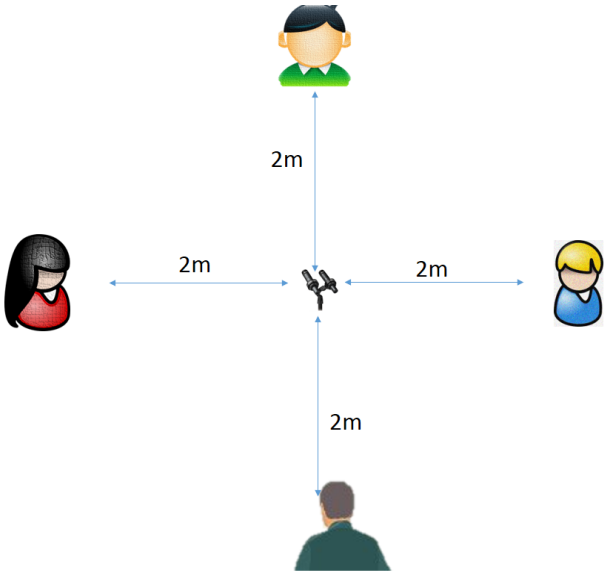


Figure 6-9: Experimental setup

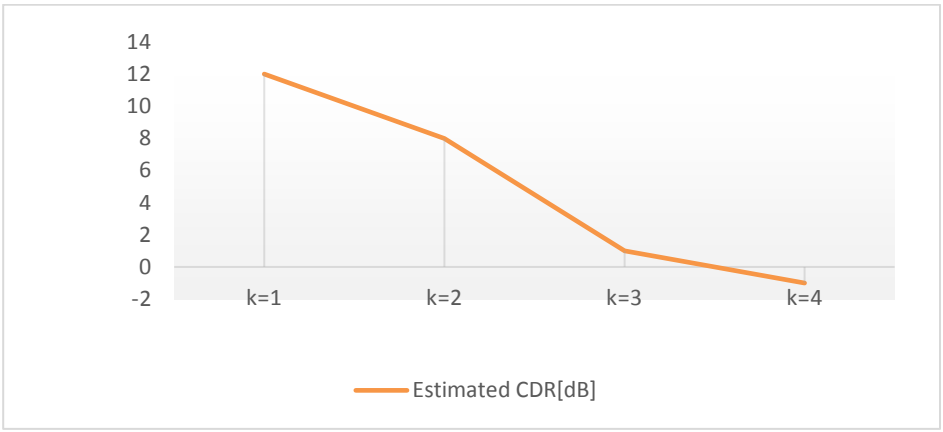


Figure 6-10: Effect of interference on CDR estimates

The effect of the frequency band-width and the reverberation time on the calculated CDR values over 20ms time-frames and averaged for a 3 second long sentence is investigated in the following graphs.

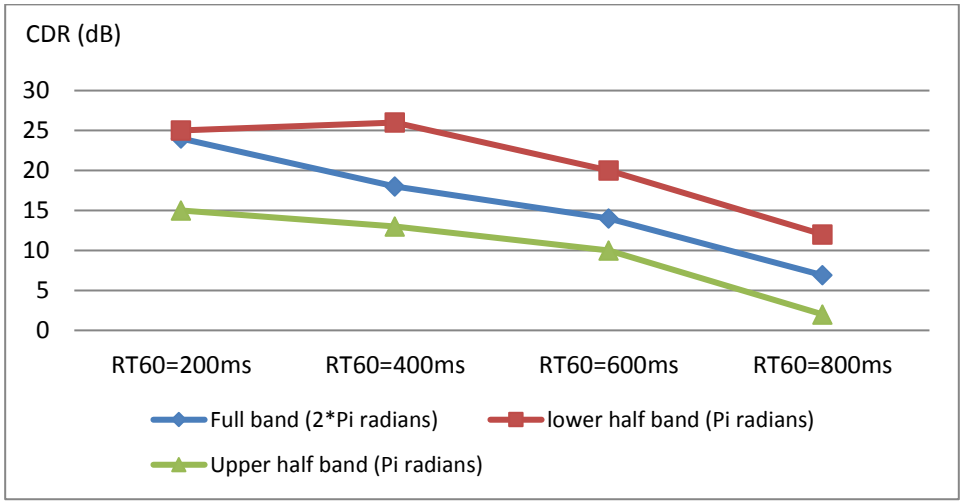


Figure 6-11: The effect of reverberation time and the frequency band on the estimated CDR values

It is interesting that the CDR values calculated for the lower band (0 – 4kHz) yields higher CDR values compared with the full band and the upper band (4kHz-8kHz) when $f_s = 8$ kHz. This is probably because most of the speech signal energy belongs to the lower band. The upper band signal contains less coherent speech and consequently it has a lower CDR value.

6.5 CDR for multi-talk detection and source counting

One of the desired characteristics of any detector is that its features are sufficiently simple, easy to calculate, have discriminatory power and work well under changing noise conditions [143]. The CDR is independent of the sources energy levels and can be applied where loud and quiet sources are simultaneously active. The methods here assume that all nodes are of the same structure because.

The target scenario of the active source counting method is a spontaneous meeting where each participant is located close (less than 30cm) to a recording device and the distance between two adjacent nodes is not less than one meter. By the observations made in this chapter and the setup assumptions, nodes with higher CDR values are more likely located close to an active source, and hence it is possible to count the nodes with relatively high CDR values in order to find the number of

simultaneously active sources. The proposed algorithm is summarised in Figure 6-12.

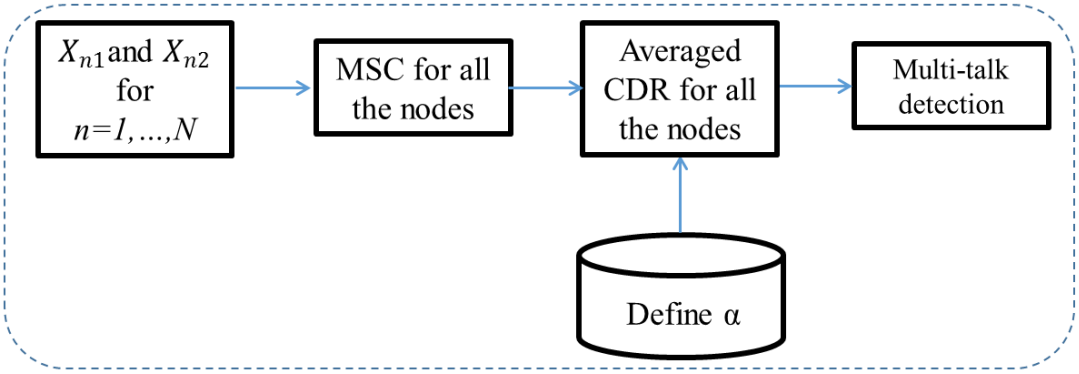


Figure 6-12: The proposed multi-talk detection method diagram

6.5.1 Proposed multi-talk detection method

As it was observed in the previous section the interference affects the nodes CDR values (calculated or estimated). The following method is proposed for the overlapping frames detected for the ad-hoc scenarios, where there is only one node within a 30cm distance from each speaker and not any two speakers are closer than 100cm. This assumption is necessary to guarantee that one source is not counted twice (i.e. two nodes with high CDR located close to one source).

Table 6-1:The proposed Multi-talk detection method

➢ Obtain $\boldsymbol{X_{n1}}$ and $\boldsymbol{X_{n2}}$ for all *N* ad-hoc nodes (6-2)
➢ Calculate $\boldsymbol{C_N(l,f)}$ and $\boldsymbol{C_x(l,f)}$ for all nodes (6-10,6-11) and obtain the CDR value for the time-frequency bins (6-15)
➢ Average the CDR estimates over *P* adjacent frames and all the frequencies.
➢ Having the CDR values at all nodes count the number of nodes within $[\boldsymbol{CDR_{max}}, \boldsymbol{\alpha \times CDR_{max}}]$ interval (6-18).
➢ If the number of nodes within the interval is greater than one multi-talk (overlap) has occurred.

6.5.2 Proposed Source counting by CDR values at each node

The proposed algorithm is summarised in Table 6-2. The CDR values are estimated for all the nodes.

$$\mathcal{A} = \{\overline{CDR}_n\}_{n\in\{1,\ldots,N\}} \quad 6\text{-}17$$

and $\overline{CDR}_n$ is kept in the set $\mathcal{A}$ if

$$\overline{CDR}_n \geq \overline{CDR}_{\max} - \alpha(\overline{CDR}_{\max} - \overline{CDR}_{\min}), \quad 6\text{-}18$$

where $\overline{CDR}_{\min} = \min(\mathcal{A})$, $\overline{CDR}_{\max} = \max(\mathcal{A})$ and α is a parameter to set the threshold of maintained CDR values.

The number of the remaining nodes in $\mathcal{A}$ after applying (6-18) is counted as the number of simultaneously active sources.

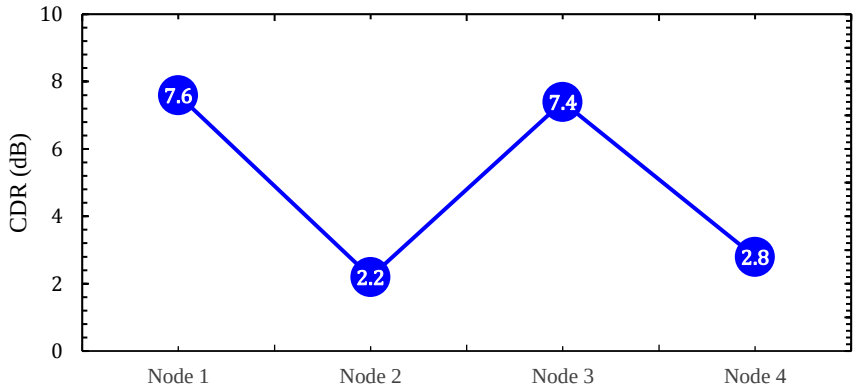


Figure 6-13: CDR values at each node when two sources are active simultaneously.

Table 6-2 explains the proposed source counting method by analysing and comparing the CDR values at the ad-hoc nodes.

Table 6-2: Proposed source counting by CDR at each node

➢ Start with $\boldsymbol{x_{n,1}}$and $\boldsymbol{x_{n,2}}$ for all the $\boldsymbol{N}$ ad-hoc nodes (6-2)
➢ Calculate $\boldsymbol{C_{u_n}(f)}$ and $\boldsymbol{C_{x_n}(l,f)}$ for all nodes (6-13),(6-14).
➢ Average the CDR estimates over $\boldsymbol{L}$ adjacent frames and across the frequency band of interest.
➢ Having the CDR values at all nodes, $\boldsymbol{\mathcal{A}}$, find the global minimum ($\overline{\boldsymbol{CDR}}_{\mathbf{min}}$) and global maximum ($\overline{\boldsymbol{CDR}}_{\mathbf{max}}$).
➢ Count the number of nodes in the top $\boldsymbol{\alpha \times 100}$% of CDR values. i.e. within $[\overline{\boldsymbol{CDR}}_{\mathbf{max}} - \boldsymbol{\alpha}(\overline{\boldsymbol{CDR}}_{\mathbf{max}} - \overline{\boldsymbol{CDR}}_{\mathbf{min}}), \overline{\boldsymbol{CDR}}_{\mathbf{max}}]$ interval.
➢ The number of maxima (nodes with highly coherent speech signals) represents the number of simultaneously active sources for the time frame. If more than one, cross talk would have happened.

6.6 Offline speaker counting in highly reverberant environment through clustering the coherence features

In a meeting scenario with *M* participants located at fixed locations the objective of the offline speaker counting is to estimate *M* based on the dual-channel recording with unknown inter-channel distance *d*. The dual recordings from 6-4 and 6-5 contain coherent speech (direct path signal) and diffuse noise and reverberation. The frequencies with high MSC values are the frequencies generated by each speaker vocal tracts [144] and the frequencies with lower MSC values are the diffuse noise and the reverberation.

$$x_p(n) = \tilde{x}_p^{co}(n) + \tilde{x}_p^{di}(n), \qquad 6\text{-}19$$

The MSC feature is calculated for each frequency bin (k) by

$$\boldsymbol{c}(k) = \frac{\left|\varphi_{x_1|x_2}(k)\right|^2}{\varphi_{x_1|x_1}(k)\,\varphi_{x_2|x_2}(k)}, \qquad 6\text{-}20$$

for a dual channel (6-2) ad-hoc frequency domain recordings $(x_1(k), x_2(k))$ at unknown locations. It is observed that different participants of a meeting have different coherence frequency responses $(\boldsymbol{c}(k))$ from (6-20) due to their different locations and speech characteristics.

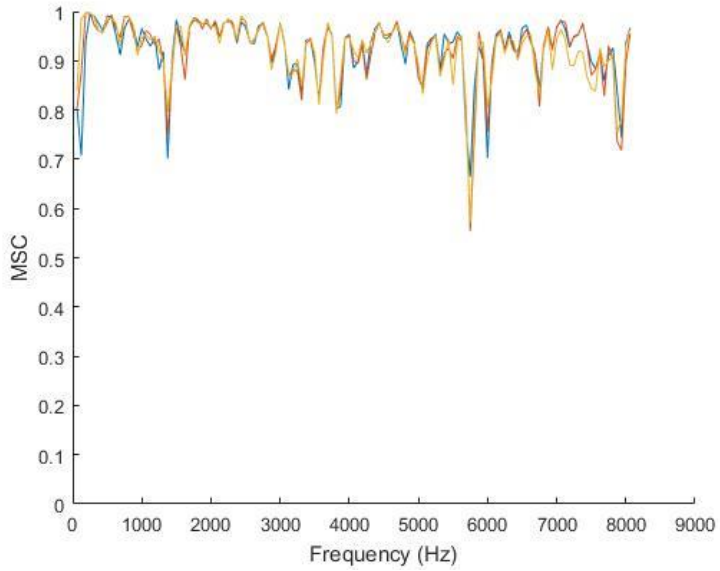


Figure 6-14: Three different sentences (2 seconds long) read by the same speaker at the same location.

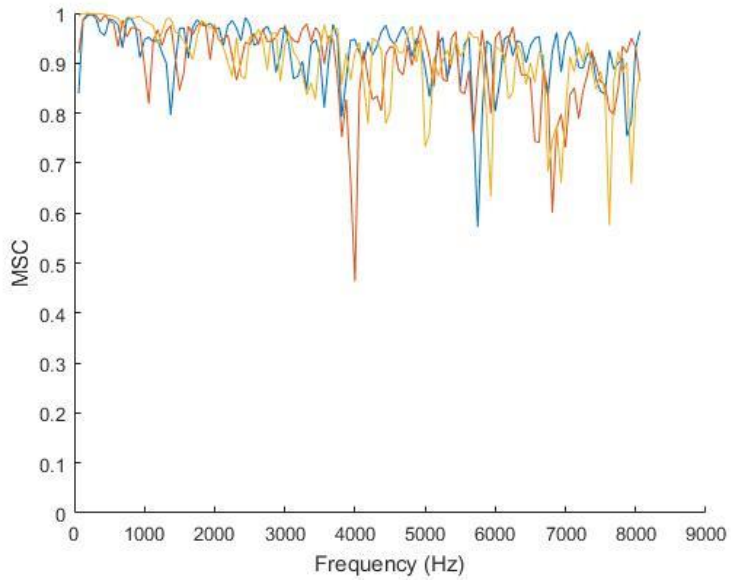


Figure 6-15: Three different speakers read the same sentence (2 seconds long) at the three different locations.

As it is shown in Figure 6-14, MSC curves are very similar for the same speaker [145] regardless of the pronounced words as long as the speaker does not move. This observation suggests that the MSC features derived from the same speaker cluster together. Figure 6-15 indicates that the different speakers at different locations have distinctly different MSC features even when they read the same sentence.

These observations made by analysis several speakers and locations is utilised in this section to form clusters (from 2 clusters to arbitrary $\widehat{\boldsymbol{M}}_{\mathbf{max}}$) for the speakers speech segments and count the optimal number of clusters as the estimate of the number of sources ($\widehat{\boldsymbol{M}}$). Table 6-3 summarises the proposed offline source counting method based on the MSC values.

TABLE 6-3
THE PROPOSED OFFLINE SPEAKER COUNTING METHOD

1)	Start with the recorded mixture $\boldsymbol{x_p(n)}$ from.
2)	Obtain the speech signal for each time segment in the frequency domain.
3)	Extract the MSC features for each time segment of the speech signal and obtain $\mathbf{C}_k$
4)	Cluster the extracted features ($\mathrm{c}(k)$) into $K = 2$ to $\widehat{M}_{\max}$ clusters and choose the optimal K (based on the Calinski Harabasz (CH) [146] clustering evaluation metric) as the number of clusters.
5)	The optimal number of the clusters ($\widehat{M}$) is the estimate for the number of sources.

The K-means clustering method [33] is applied to cluster the extracted MSC features (6-20) for 2 second segments into $K = 2$ to $\widehat{M}_{\max}$. The optimal clustering results (i.e. the optimal number of K) is then chosen based on the Calinski Harabasz (CH) [146] clustering evaluation criteria. The optimal number of the clusters ($\widehat{M}$) is compared with the real number of the participants. For the experimental studies, 256 frequency bins are applied in order to calculate $\boldsymbol{c}(k)$. Having $\boldsymbol{c}(k)$ for each segment the matrix of the MSC features are obtained as

$$\mathbf{H}_{l,k} = \begin{bmatrix} c(0,0) & \cdots & c(L-1,0) \\ \vdots & \ddots & \vdots \\ c(0,N-1) & \cdots & c(L-1,N-1) \end{bmatrix} \qquad 6\text{-}21$$

For all the time segments and the frequency bins.

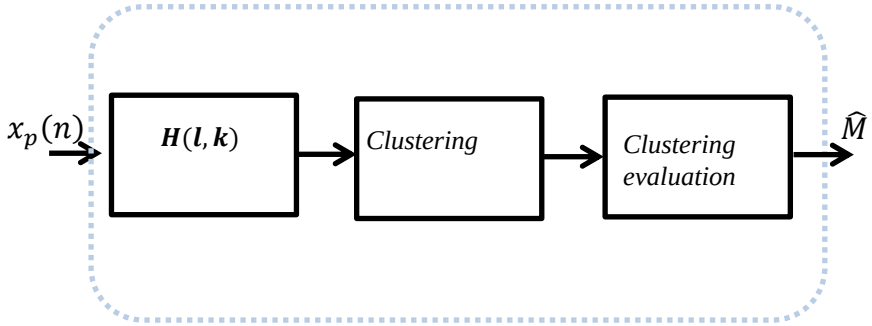


Figure 6-16: The proposed offline source counting method based on MSC features

The Success Rate (SR) (6-22) is applied as the performance measurement. Assuming that T_c is the number of scenarios that the number of sources is estimated correctly (i.e. $\widehat{M} = M$) and T_t is the total number of test scenarios, the Success Rate (SR) evaluation measurement is defined as

$$\text{SR} = \frac{T_c}{T_t} \times 100. \qquad 6\text{-}22$$

This method is evaluated in the results section and is compared with the baseline TDOA method.

6.7 Experimental evaluation and results

The baseline speaker diarisation and cross talk detection systems are based on assigning each speech segment to a unique cluster (speaker) in the output and the overall system is evaluated using the metric known as the Diarisation Error Rate (DER) [147], [148] which is the sum of speech/non-speech error and speaker detection error. A slightly different evaluation approach is proposed in this section as the objective is not speaker diarisation but overlap detection and source counting (6-23), (6-24).

CDR values at each node locations are calculated over short time frames of 20ms, which corresponds to 320 samples at 16 kHz sampling frequency and are averaged across all the frequencies (6-8). This is the typical time duration for which a speech segment is assumed to be stationary. However, better performance can be obtained when a larger value is chosen for the frame length [28] or the averaged CDR value

across consecutive time frames are applied as the discriminative feature (e.g. 15 frames which translates to 300ms).

The Experimental configuration is summarised in Table 6-4.

Table 6-4: The Experimental configuration

Parameter	***Setting***
Sampling frequency	*16kHz*
Frame length	*20ms*
FFT length	*160*
Frame shift	*160 samples*
Intra-channel distances	*15cm*
SNR	*10dB*

The experimental setups with one and three active speakers are depicted in Figure 6-17 and Figure 6-18.

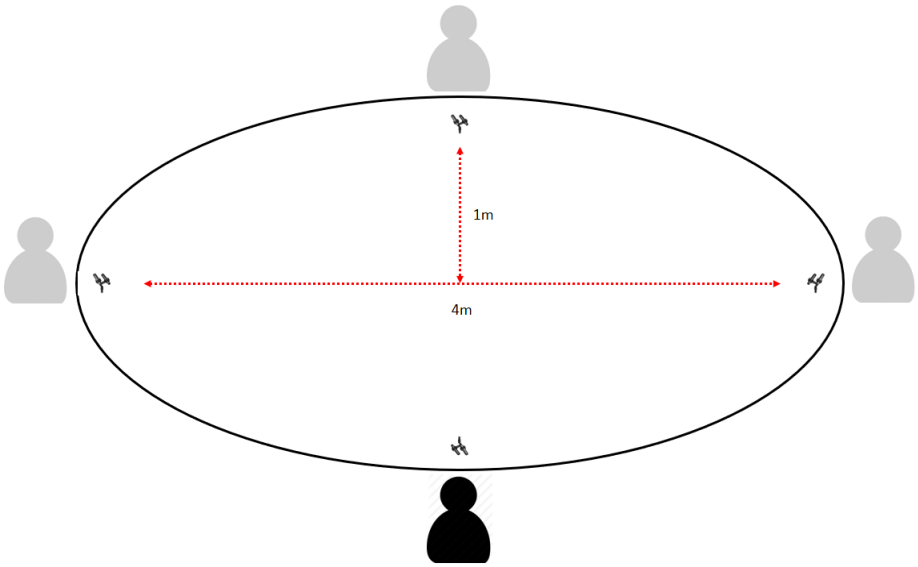


Figure 6-17: The experimental setup with 4 nodes and 4 participants when there is only one active source

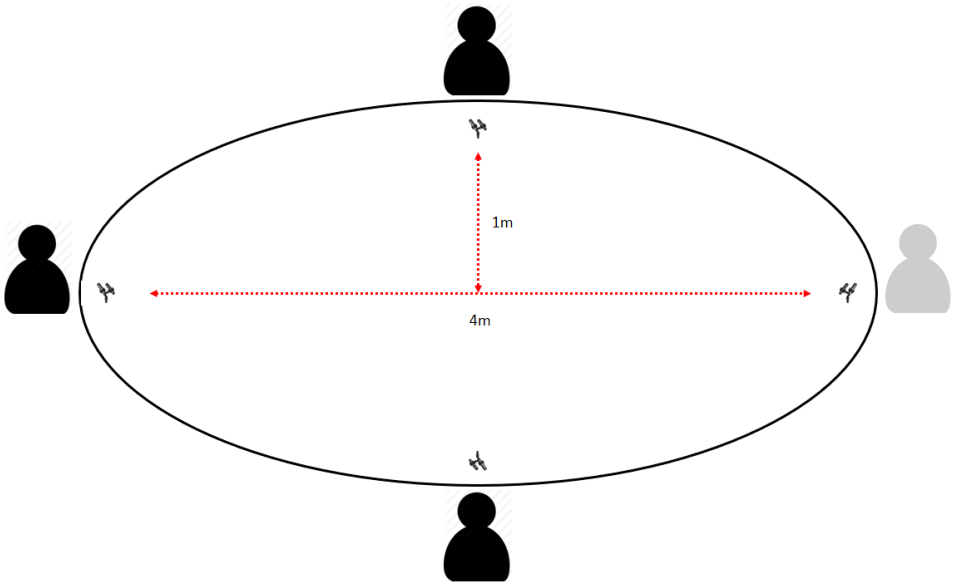


Figure 6-18: The experimental setup with 4 nodes and 4 participants when there is three active sources

6.7.1 Multi-talk detection

Overlap detection aims to flag the time-frequency bins with more than one active source without attempting to count the number of simultaneously active talkers. The True Positive rate (TPR) for cross talk detection without focusing on the number of simultaneously active sources is defined as [149]:

$$TPR_{multi-talk} = \frac{T_{cc}}{T_{cc} + T_{c1} + T_{1c} + T_{11}} \quad 6\text{-}23$$

100 time segments are applied for each value of P and overall 700 time segments are randomly generated as single-talk and multi-talk to test the proposed multi-talk detection method (Figure 6-19).

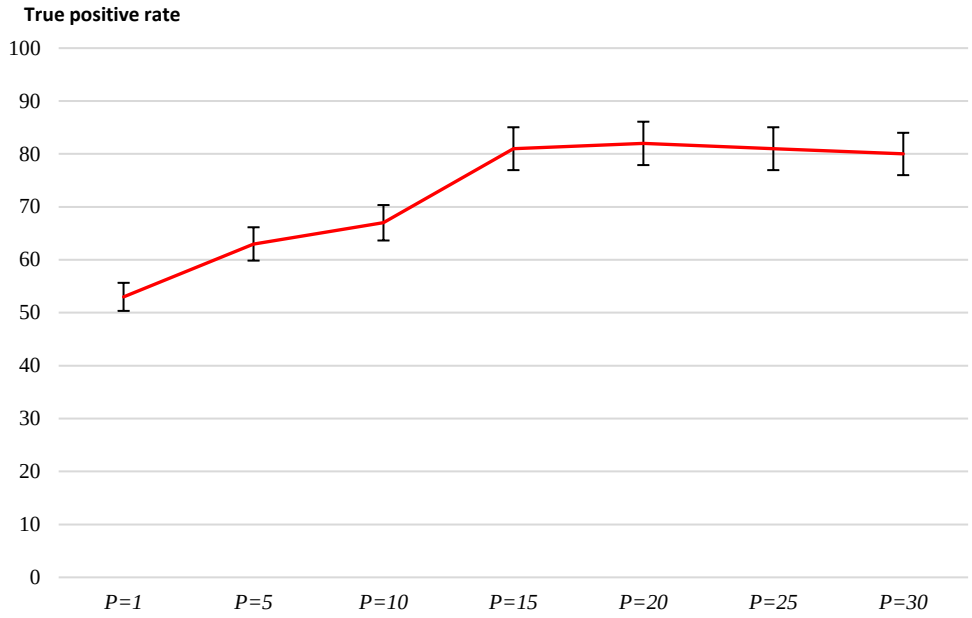


Figure 6-19: Interfering talker(s) detection success rate

where T_{cc} is the number of frames with more than one active source labeled as cross talk correctly, T_{c1}, T_{1c} are incorrectly labeled frames (cross talk labelled as single source or vice versa) and $T_{11}$is the single talk frames labelled correctly as single talks.

The CDR estimation method of [69] is applied here for all the experiments as it does not require the coherent signal direction of arrival ($\boldsymbol{\theta_n}$ from (6)), it is shown that Direction of Arrival (DOA) based methods do not yield successful source counting results (48.6% accuracy).

The results are presented for different values of *P* (the number of applied adjacent time frames) and it is concluded that *P* values equal to or greater than 15 (which translates to 300ms frames or longer) yield higher interference detection success rate compared with shorter frames. This can partly be a result of the inaccurate CDR calculation/estimation over shorter frames and partly because of the speech characteristics over short frames.

6.7.2 Simultaneous Source counting results

In this section the CDR values are utilised for counting the number of simultaneously active sources with making use of the spatial coverage of ad-hoc

arrays. This is done by implementing the proposed method in section 6.4.2 for 25 different ad-hoc scenarios in terms of the room dimensions, reverberation times, the number of sources (1 to 4) and the number of the dual nodes (4 to 10) and averaging the results. A more detailed source counting evaluation is presented in Figure 6-20 and summarises the source counting confusion matrix for 100 ad-hoc scenarios. The True Positive Ratio (TPR) for source counting is defined as:

$$TPR_{Source\ counting} = \frac{T_{kk}}{T_{k1} + T_{k2} + \cdots + T_{kM}} \qquad 6\text{-}24$$

where T_{kk}, $k \neq 1$ is the number of frames with *k* active sources correctly labelled as having *k* active sources and T_{kj} is the number of frames with *k* active sources which are incorrectly labelled as having $j \neq k$ active sources. $T_{k1} + T_{k2} + \cdots + T_{kM}$ is the overall number of the frames in the test set.

100 time segments with 1 to 4 active sources are applied for *P=15* and overall 400 segments are randomly generated to evaluate the proposed source counting algorithm (Figure 6-20).

	One detected active source	***Two detected active source***	***Three detected active source***	***Four detected active source***
k=1	*87%*	*10%*	*3%*	*0%*
*k=**2***	*12%*	*81%*	*7%*	*0%*
K=3	*0%*	*12%*	*78%*	*10%*
k=4	*0%*	*20%*	*22%*	*58%*

Figure 6-20: TPR confusion matrix for simultaneously active sources, *P=15*

It is concluded that for a small number of sources (i.e. 1 and 2) the proposed source counting is able to detect the number of sources with an accuracy of 81% minimum and the increase in the number of the simultaneously active sources decreases the source counting accuracy.

6.7.3 Offline source counting results

Offline Source counting results for the proposed participant counting method for each meeting is illustrated in Figure 6-21.

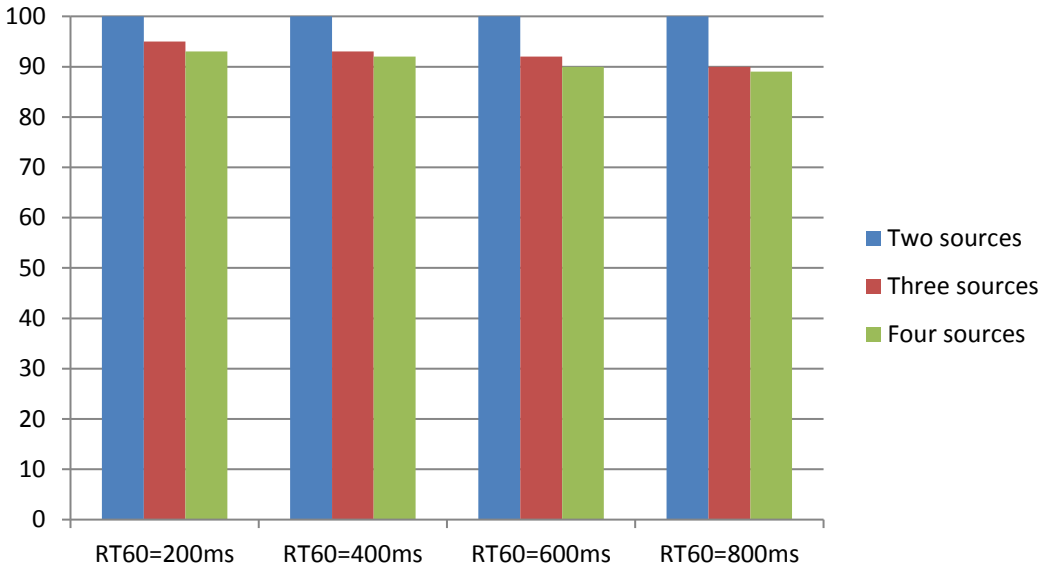


Figure 6-21: Meeting participant counting results, SNR=40dB

It is shown that the proposed method is robust to reverberation (Figure 6-21). The proposed method is also robust to inter-channel spacing (d) and it is shown that the distance between the microphones at high SNR and low reverberation times does not affect the participant counting results (Figure 6-22).

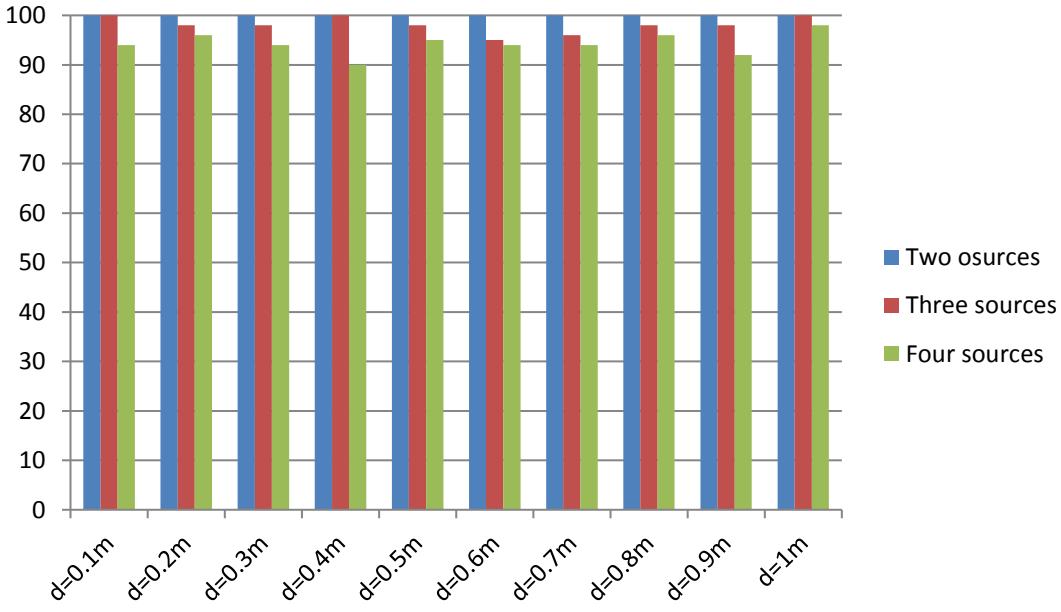


Figure 6-22: Meeting participant counting results, SNR=40dB, Reverberation time=200ms

The proposed method is shown to be more accurate than the base-line TDOA estimates from Generalised Cross-Correlation with Phase Transform (GCC-PHAT) [150].

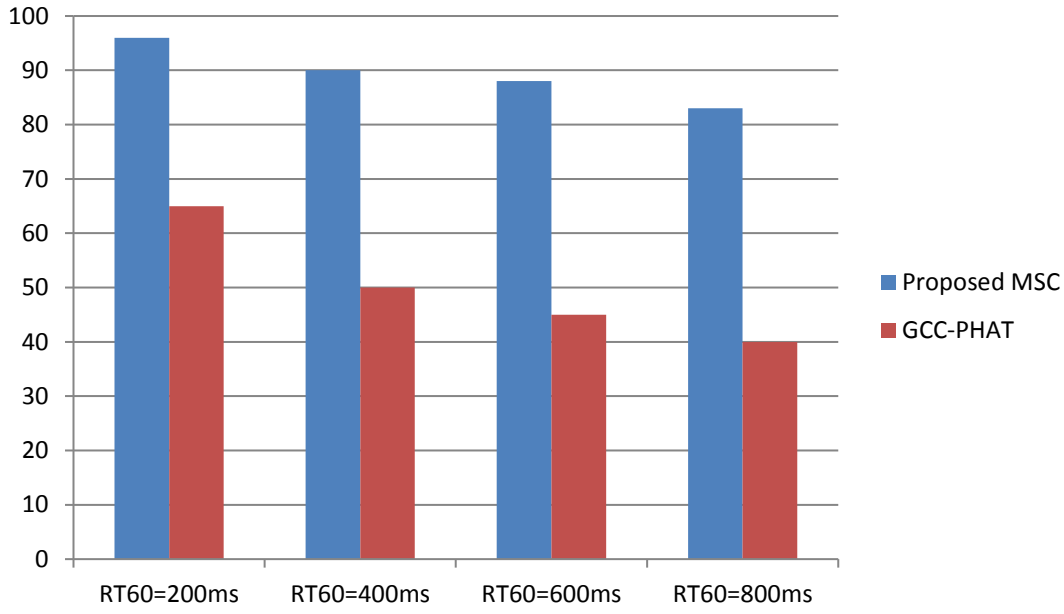


Figure 6-23: Average results for 2 to 6 sources for different reverberation times.

Figure 6-23 investigates the effect of the number of sources on the source counting accuracy (6-24) and it is concluded that the proposed method can outperform the existing feature (GCC-PHAT) in reverberant environments. It is also concluded that the proposed method is robust to reverberation.

6.8 Conclusion

This chapter proposed a new feature for cross talk (overlap) detection during multi-party meeting scenarios based on real-time and pseudo real-time estimated CDR cues. It is shown that by estimating CDR features or calculating the MSC and the CDR features over short time segments, it is possible to detect interfering sources and the cross talk, independent of the sources energy level in the context of ad-hoc arrays. The proposed feature can be extracted without the time alignment of the ad-hoc channels and the proposed method does not require the prior knowledge of the room geometry, microphone and source locations, room impulse responses or microphone array structure. The proposed feature is also applied for source counting and it is concluded that under justifiable and acceptable distance conditions, it is practically possible to count the number of simultaneously active sources utilising the spatial coverage of the ad-hoc arrays. The proposed methods of this chapter are applicable to real time scenarios and yields 80% successful multi-talk detection rate and average 75% success in source counting.

Proposing a new cross-talk detection feature and applying it to the ad-hoc arrays is the novelty of this approach which does not require statistical modelling of the speech sources or a training phase. The proposed method in this chapter can accurately detect overlaps shorter than 500ms.

For the offline source counting the novel MSC feature and clustering based method is proposed and successfully tested. It is concluded that the proposed method is robust to reverberation. Very accurate source counting results (minimum 80% success rate) are obtained that outperforms the baseline GCC_PHAT methods in moderately and highly reverberant environments.

7 Conclusion and future works

7.1 Conclusion

In this thesis applications of the ad-hoc microphone arrays as emerging recording tools for press conferences, lecture halls and meetings are investigated and novel methods and features are proposed or modified for microphone clustering, source localisation, multi-channel speech enhancement, source counting and multi-talk detection. The proposed methods are specifically tailored to the context of the ad-hoc arrays considering the specifications of such arrays. As the target scenarios of this research is broad and not confined to any certain microphone structure or number of the channels, for each application the most suitable and general feature which can be applied to any ad-hoc scenario is chosen and applied. The proposed features are based on the RIR amplitude attenuation and time delay features for microphone clustering and source localisation, kurtosis of the LP residual signal for microphone discrimination and informed dereverberation and coherent to diffuse ratio for multi-talk detection and source counting.

The proposed clustering and source localisation methods benefit from the wide and flexible spatial coverage of the ad-hoc arrays and overcome the missing knowledge of the microphone arrays geometry and the relative distances. The derived side information such as the relative source to microphone distances is also utilised to propose an informed multichannel dereverberation method in the context of ad-hoc arrays.

In this thesis the code-book based microphone clustering is proposed for microphone clustering, the surface fitting approach is proposed for the source localisation, two-stage short and long-term dereverberation based on the spatially modified linear prediction is applied to the ad-hoc scenarios and a coherence based approach is proposed for source counting and multi-talk detection.

7.2 Recommendations for future research

According to the literature (reviewed in chapter 2 and chapter 4) it is possible to reconstruct the room geometry and localise the microphones and the sources in the room. By deriving such information it is possible to estimate the RIRs at microphones locations and exploit the estimated RIRs for some informed speech dereverberation method (Chapter 5). Although it is not possible to obtain the accurate RIRs by reconstructing the acoustic scene, deriving this information and having a rough estimate of the RIRs at each microphone location, helps guide the equalisation process. In addition to dereverberation, the full reconstruction of the acoustic scene can be applied for informed noise removal and interference suppression by detecting the closest microphone (cluster of microphones) to the non-diffuse noise source and using it to estimate the noise at other microphones locations. The noise estimate knowledge along with the estimated RIRs can be applied for informed noise cancellation.

The proposed spatial linear prediction method also needs to be further investigated in terms of finding optimised values for weights. This may be done through proposing a relative distance feature that maximises the LP coefficients estimation accuracy or by proposing a clustered approach to LP estimation.

References

[1] Tavakoli, Jensen, Chrsitensen and Benesty, “A Framework for Speech Enhancement With Ad Hoc Microphone Arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 6, pp. 1038-1051, June 2016.

[2] M. Souden, K. Kinoshita and T. Nakatani, “An integration of source location cues for speech clustering in distributed microphone arrays,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013.

[3] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal Acoustic Society of America,* p. p 943, April 1979,.

[4] W. S. Woods, E. Hadad, I. Meks, B. Xu, S. Gannot and T. Zhang, “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.

[5] 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*, Vancouver, BC, 2013.

[6] L. Wang, K. Hon, J. Reiss and A. Cavallaro, “Self-Localization of Ad-Hoc Arrays Using Time Difference of Arrivals,” *Transactions on Signal Processing,* vol. 64, no. 4, pp. 1018-1033, Feb.15, 2016.

[7] 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.

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

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

[10] 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.

[11] S. Miyabe, N. Ono and S. Makino, “Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain,” in *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013.

[12] P. Pertila, M. S. Hamalainen and M. Mieskolainen, “Passive Temporal Offset Estimation of Multichannel Recordings of an Ad-Hoc Microphone Array,” *Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11 , pp. 2393-2402, 2013.

[13] 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.

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

[15] R. Lienhart, I. Kozintsev, S. Wher and M. Yeung, “On the importance of exact synchronization for distributed audio signal processing,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[16] P. C. Loziuo, Speech enhancement, Theory and practice, Boca raton, FL: CRC press, 2013.

[17] M. Delcroix, Hikichi, Takafumi and M. Miyoshi, “Dereverberation and Denoising Using Multichannel Linear Prediction,” *in Transactions on Audio, Speech, and Language Processing, IEEE,*, vol. 15, no. 6, pp. 1791-1801, Aug. 2007.

[18] A. Warzybok, I. Kordasi and J. Jungmann, “Subjective speech quality and

speech intelligibility evaluation of single-channel dereverberation algorithms," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014.

[19] J. Bensty and Y. Huang, A Perspective on Single-channel Frequency-domain Speech Enhancement, M & C, 2011.

[20] A. Baghaki, M. O. Ahmad and M. Swamy, "A new two-stage method for single-microphone speech dereverberation," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, QC, 2016.

[21] B. Kim, Y. Hwang and H. M. Park, "Speech enhancement based on softmasking exploiting both output SNR and selectivity of spatial filtering," *Electronics Letters*, vol. 50, no. 12, pp. 889-891, June 2014.

[22] V. Tavakoli, J. Jensen, M. Christensen and J. Benesty, "Pseudo-Coherence-based MVDR beamforming for speech enhancement with ad-hoc microphone arrays," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, South Brisbane, 2015.

[23] 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.

[24] J.-C. Junqua, Robust Speech Recognition in Embedded Systems and PC Applications, Kluwer, 2002.

[25] T. K. T. Horiuchi, Hayashida, Nakayam, Nishiura and Yamashita, "Close/distant talker discrimination based on kurtosis of linear prediction residual signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[26] W. Ong and A. Tan, "Robust voice activity detection using gammatone filtering and entropy," in *International Conference on Robotics, Automation and Sciences (ICORAS)*, Melaka, 2016.

[27] P. Giannoulis, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *2015*, Nice, 23rd European Signal Processing Conference (EUSIPCO).

[28] 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.

[29] M. Jeub, C. Nelke, C. Beaugeant and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *19th European Signal Processing Conference*, Barcelona, 2011.

[30] A. Popper and R. Fay, Sound Source Localization, Springer, 2005.

[31] Z. Liu, Z. Zhang, L. Wei He and C. Phil, "Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays," in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Honolulu, HI, USA, 2007.

[32] E. Vincent, N. Bertin, R. Gribonval and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107,115, May 2014.

[33] G. Paliouras, "Machine Learning and Its Applications," 2001.

[34] Y. Jiamg and R. Liu, "Binaural deep neural network for robust speech enhancement," in *International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Guilin, 2014.

[35] M. Kolbaek, Z. Tan and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153-167, 2016.

[36] T. Vu, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.

[37] R. Talmon, I. Cohen and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," in *2011, pp. 5084-5087.*, Prague, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[38] P. Lin, Y. Jui, Y. Ying, Y. Chen and M. Wu, "Unsupervised Speaker

Clustering Using SVM Training Missclassification Rate for Short-Duration Speech Signals," in *Fourth International Conference on Genetic and Evolutionary Computing, Shenzhen*, 2010.

[39] H. Almogotir kadhim, L. Woo and S. Dlay, "Novel algorithm for speech segregation by optimized k-means of statistical properties of clustered features," in *IEEE International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, 2015, pp. 28.

[40] S. Y. Kung, Kernel Methods and Machine Learning, Cambridge university press, 2014.

[41] Himawan, McCowan and Sirdharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, 2010.

[42] Y. Jia, L. Yu and I. Kozintsev, "Distributed Microphone Arrays for Digital Home and Office," in *International Conference on Acoustics Speech and Signal Processing Proceedings,*, Toulouse, 2006.

[43] M. Taghizadeh, A. Asaei, P. Garner and H. Bourlard, "Ad-hoc microphone array calibration from partial distance measurements," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Villers-les-Nancy, 2014 .

[44] V. M. Tavakoli, J. R. Jensen, J. Benesty and M. G. Christensen, "A partitioned approach to signal separation with microphone ad hoc arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, Shanghai, 2016.

[45] R. Arnuncio and B. Juang, "Blind Source Separation of Acoustic Mixtures with Distributed Microphones," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Hawaii, USA, 2007.

[46] I. Himawan, Speech recognition using ad-hoc microphone arrays, PhD Dissertation, Queensland university of technology, 2010.

[47] 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.

[48] K. Youssef, S. Aregentier and L. Zarader, "A binaural sound source

localization method using auditive cues and vision," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012.

[49] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi and T. Yamada, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014.

[50] S. Galgali, S. Priyanka, R. Shashank and A. Patil, "Speaker profiling by extracting paralinguistic parameters using mel frequency cepstral coefficients," in *International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),*, Davangere, 2015, pp. 486-489., 2015.

[51] Takashima, R. Takiguchi and Y. Ariki, "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function,," in *IEEE International Conference on Acoustics, Speech and Signal Processin*.

[52] S. Sadjadi and L. Hansen, "Blind reverberation mitigation for robust speaker identification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012.

[53] S. Ganapathy, J. Pelecanos and K. Omar, "Feature normalization for speaker verification in room reverberation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011.

[54] T. Le, T. Nowakowski, L. Daudet and J. De rosny, "Experimental validation of TOA-based methods for microphones array positions calibration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.

[55] M. Yang, D. Jackson, J. Chen, Z. Xiong and J. Williams, "A TDOA localization method based on de-embedding the propagation background," in *Texas Symposium on Wireless and Microwave Circuits and Systems (WMCS)*, Waco, TX, 2016.

[56] Z. Dong and M. Yu, "Research on TDOA based microphone array acoustic localization," in *12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Qingdao, 2015.

[57] Z. Huang, D. Zhan, D. Ying and Y. Yan, “Robust multiple speech source localization using time delay histogram,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Shanghai, 2016.

[58] Y. Chan, R. Hattin and J. Plant, “The least squares estimation of time delay and its use in signal detection,” *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 26, no. 3, pp. 217-222, Jun 1978.

[59] R. Berken and I. Cohen, “Microphone array power ratio for quality assessment of reverberated speech,” in *EURASIP journal on advances in signal processing.* , December 2015.

[60] P. Parad, D. Sharma and P. Naylor, “Non-intrusive estimation of the level of reverberation in speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Florence, 2014.

[61] Y. Ji, Y. Baek and Y. Park, “A priori SAP estimator based on the magnitude square coherence for dual-channel microphone system,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* South Brisbane, QLD, 2015.

[62] N. Yousefian and C. Loizou, “A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 2, pp. 599-609, Feb. 2012.

[63] S. Bharitkar and C. Kyriakakis, “A cluster centroid method for room response equalization at multiple locations,” in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* New Platz, NY, 2001.

[64] I. Dokmanic, L. Vetteli and Martin, “How to Localize Ten Microphones in One Fingersnap,” in *22nd European Signal Processing Conference,* Lisbon, Portugal, 2014.

[65] I. Dokmanic, Y. M. Lu and M. Vetterli, “Can one hear the shape of a room: The 2-D polygonal case,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2012 .

[66] D. Frey, A. Coelho and R. M. Rangayyan, “The loudspeaker as a measurement sweep generator for the derivation of the acoustical impulse response of a concert hall,” in *Canadian Conference on Electrical and Computer Engineering,* Niagara Falls, ON, 2008.

[67] J. Eaton, H. Moore, P. A. Naylor and J. Skoglund, “Direct-to-Reverberant Ratio estimation using a null-steered beamformer,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015.

[68] 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.

[69] A. Schwartz and W. Kellerman, “Coherent-to-Diffuse Power Ratio Estimation for Dereverberation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006-1018, June 2015.

[70] Rockah and P. Schultheiss, “Array shape calibration using sources in unknown location—Part I: Far-field sources,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, p. 286–299, 1987.

[71] Schultheiss and Y. Rockah, “Array shape calibration using sources in unknown locations—Part II: Near-field sources and estimatorimplementation,”,” *IEEE Trans. Acoust., Speech, Signal Process.,* , Vols. vol.ASSP-35, no. 6, , pp. pp. 724–735, , Jun. 1987..

[72] M. Hennecke and G. Fink, “Towards acoustic self-localization of ad hoc smartphone arrays,” in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, 2011.

[73] N. Patwari, J. Ash, S. Kyperountas, A. Hero and R. Moses, “Locating the nodes: cooperative localization in wireless sensor networks,” *Signal Processing Magazine* , vol. 22, no. 4, p. 54–69, 2005.

[74] V. Raykar, V. Kozintsev and R. Lienhart, “Position calibration of microphones and loudspeakers in distributed computing platforms,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70-83, Jan. 2005.

[75] S. Pasha, Ritz and Zou, “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.

[76] I. Himawan, I. McCowan and S. Sridharan, “Clustered Blind Beamforming From Ad-Hoc Microphone Arrays,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661-676, May 2011.

[77] A. Asaei, M. Davis, H. Bourlard and V. Cevher, “Computational methods for structured sparse component analysis of convolutive speech mixtures,” in *International conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012.

[78] P. Pertila, M. Mieskolainen and S. Hamalanen, “Passive self-localization of microphones using ambient sounds,” in *20th European Signal Processing Conference (EUSIPCO)*, Bucharest, 2012.

[79] T. Ajdler and M. Vetterli, “The plenacoustic function, sampling and reconstruction,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.Proceedings. (ICASSP '03).,*, 2003.

[80] R. Heusdens and D. Gaubitch, “Time-delay estimation for TOA-based localization of multiple sensors,” in *International Conference on Acoustics, Speech, Signal Processing*, Florence, Italy, 2014.

[81] S. Vesa, “Binaural Sound Source Distance Learning in Rooms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498-1507, 2009.

[82] D. Ramirez, J. Via and I. Santamaria, “A generalization of the magnitude squared coherence spectrum for more than two signals: definition, properties and estimation,” in *International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, 2008.

[83] Y. Hu and P. Loizou, “Subjective evaluation and comparison of speech enhancement algorithms,” *Speech Communication*, vol. 49, no. 7, pp. 588-601, 2007.

[84] J. Ma, Y. Hu and P. Loziou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387-3405, 2009.

[85] 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.

[86] B. Liao, L. Huang, C. Guo and H. C. SO, “New Approaches to Direction-of-Arrival Estimation With Sensor Arrays in Unknown Nonuniform Noise,” *IEEE*

Sensors Journal, vol. 16, no. 24, pp. 8982-8989, Dec.2016.

[87] C. Anderson, P. Teal and Poletti, "Multichannel Wiener filter estimation using source location knowledge for speech enhancement," in *Workshop on Statistical Signal Processing (SSP),* Gold Coast, 2014.

[88] V. V. Reddy, A. W. Khong and B. P. Ng, "Unambiguous Speech DOA Estimation Under Spatial Aliasing Conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 22, no. 12, pp. 2133-2145, Dec. 2014.

[89] M. Farmani, S. Pedersen, Z. Tan and J. Jensen, "Informed TDoA-based direction of arrival estimation for hearing aid applications," in *Global Conference on Signal and Information Processing (GlobalSIP),* Orlando, FL, 2015.

[90] T. Le and N. Ono, "Robust TDOA-based joint source and microphone localization in a reverberant environment using medians of acceptable recovered TOAs," in *International Workshop on Acoustic Signal Enhancement (IWAENC),* Xi'an, China, 2016.

[91] L. Lu and C. Wu, "Novel Robust Direction-of-Arrival-Based Source Localization Algorithm for Wideband Signals," *IEEE Transactions on Wireless Communications,* vol. 11, no. 11, pp. 3850-3859, November 2012.

[92] F. Peng, T. Wang and B. Chen, "Room shape reconstruction with a single mobile acoustic sensor," in *Global Conference on Signal and Information Processing (GlobalSIP),* Orlando, FL, 2015.

[93] D. Salvati, C. Drioli and L. Foresti, "Sound Source and Microphone Localization From Acoustic Impulse Responses," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1459-1463, Oct. 2016.

[94] T. Pham, D. Scherber and C. Papadopoulos, "Distributed source localization algorithms for acoustic ad-hoc sensor networks," in *Sensor Array and Multichannel Signal Processing Workshop,* Barcelona, Spain, 2004.

[95] M. Omer, A. Quadeer, M. Scharawi and Y. Al-Naffouri, "Time delay estimation in a reverberant environment by low rate sampling of impulsive acoustic sources," in *11th International Conference on Information Science, Signal Processing and their Applications*, Montreal, QC, Canada, 2012.

[96] A. Asaei, H. Taghizade, Bourlard and V. Cevher, “Model-based sparse component analysis for reverberant speech localization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), , vol., no., pp.1439,1443, 4-9*, 2014.

[97] D. Su, T. Vidal-calleja and V. Miro, “Simultaneous asynchronous microphone array calibration and sound source localisation,” in *International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, 2015.

[98] O. Meng, D. Sen, S. Wang and L. Hayes, “Impulse response measurement with sine sweeps and amplitude modulation schemes,” in *2nd International Conference on Signal Processing and Communication Systems*, Gold Coast, 2008.

[99] I. Memon, D. Ali and F. Ali Mangi, “Source Localization Wireless Sensor Network Using Time Difference of Arrivals (TDOA),” *International Journal of Scientific & Engineering Research,* vol. 4, no. 7, 2013.

[100] S. Pasha, C. Ritz and X. Y. 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.

[101] Timothy chartier, Numerical Methods,Design, Analysis, and Computer Implementation of Algorithms, Princeton University Press, 2012.

[102] M. Taseska and E. Habets, “Informed Spatial Filtering for Sound Extraction Using Distributed Microphone Arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 22, no. 7, pp. 1195-1207, July 2014.

[103] M. Togami, S. Suganuma, Y. Kawaguchi and T. Hashimoto, “Transient noise reduction controlled by DOA estimation for video conferencing system,” in *IEEE 13th International Symposium on Consumer Electronics*, Kyoto, 2009.

[104] N. D. Gaubitch and P. Naylor, “Spatiotemporal Averaging method for Enhancement of Reverberant Speech,” in *15th International Conference on Digital Signal Processing*, Cardiff, UK, July 2007.

[105] N. Gaubitch, B. Ward and A. Naylor, “Statistical analysis of the AR modeling of reverberant speech,” *Acoustical Society of America,* p. 4031–4039, 2006.

[106] Y. Huang, J. Luebs, J. Skoglund and B. Kleijn, “Globally optimized least-squares post-filtering for microphone array speech enhancement,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.

[107] S. Braun and E. Habets, “Dereverberation in noisy environments using reference signals and a maximum likelihood estimator,” in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, 2013, Sept. 2013.

[108] B. Dufea and T. Shimamura, “Reverberated speech enhancement using neural networks,” in *Intelligent Signal Processing and Communication Systems, 2009. ISPACS 2009. International Symposium on,*, Kanazawa, , 2009, pp. 441-444..

[109] N. D. Gaubitch, P. A. Naylor, Darren and B. Ward, “ON THE USE OF LINEAR PREDICTION FOR DEREVERBERATION OF SPEECH,” in *International workshop on acoustic echo and noise control (IWAENC 2003)*, Kyoto, Japan, 2003.

[110] M. Delcroix, T. Yoshioka, A. Ogawa and M. Fujimoto, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge,” 2015.

[111] “The REVERB challenge website,” [Online]. Available: http://reverb2014.dereverberation.com/. [Accessed 2015].

[112] S. Mosayyebpour, M. Esmaeili and A. Gulliver, “Single-Microphone Early and Late Reverberation Suppression in Noisy Speech,” *in IEEE Transactions on Audio, Speech, and Language Processing,*, Vols. vol. 21, no. 2,, pp. pp. 322-335,, Feb. 2013.

[113] M. Delcroix, T. Yoshioka, A. Ogawa and M. Fujimoto, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge,” in *Reverb workshop*, 2014.

[114] 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, 2017.

[115] N. 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.

[116] P. P. Vaidyanathan, The Theory of Linear Prediction, Morgan & Claypool Publishers, 2008.

[117] Box, G. Jenkins and G. Reinsel, Time Series Analysis: Forecasting and Control. 3rd edition, Englewood Cliffs, NJ: Prentice Hall, 1994.

[118] S. Mosayyebpour, H. Sheikhzadeh, T. Gulliver and M. Esmaeilie, "Single-Microphone LP Residual Skewness-Based Inverse Filtering of the Room Impulse Response," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1617-1632, 2012.

[119] P. Naylor, A. Kounodues, J. Gudnason and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34-43, Jan. 2007.

[120] T. Nakatani, T. Yoshioka, M. Miyoshi and B. Huang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *ICASSP*, Las Vegas, 2008.

[121] M. Parchami, W. Zhu and B. Chmpagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech communication*, vol. 87, p. 49–57, January 2017.

[122] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, T. Waterschoot and P. Naylor, "A Single-Channel Non-Intrusive C50 Estimator Correlated With Speech Recognition Performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, 2016.

[123] A. Schwartz and W. Kellerman, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006-1018, June 2015.

[124] H. Kuttruff, Room acoustics, Londond: Taylor and Francis, 2000.

[125] Kinoshita, Delcroix and Miyoshi, "Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp.

534-545, 2009.

[126] "WPE implementation," NTT, [Online]. Available: http://www.kecl.ntt.co.jp/icl/signal/wpe/download.html. [Accessed December 2016].

[127] "The REVERB challenege evaluation," [Online]. Available: http://reverb2014.dereverberation.com/result_se.html. [Accessed January 2017].

[128] D. Charlet, C. Barras and S. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013.

[129] N. Sholouhi, A. Ziaei, A. Sangwan and J. Hansen, "Robust overlapped speech detection and its application in word-count estimation for Prof-Life-Log data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015.

[130] R. Mukai, S. Araki, H. Sawada and S. Makino, "Removal of residual cross-talk components in Blind Source Separation using time-delayed spectral subtraction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, 2002.

[131] P. Mowlaee, G. Christensen and H. Jensen, "New Results on Single-Channel Speech Separation Using Sinusoidal Modeling," *Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1265-1277, July 2011.

[132] O. Walter, L. Drude and R. Haeb-umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, 2015.

[133] L. Drude, A. Chinaev, D. Tran and R. Haeb, "Towards online source counting in speech mixtures applying a variational EM for complex Watson mixture models," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[134] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *IEEE Workshop on Automatic Speech Recognition &*

Understanding (ASRU), Kyoto, 2007.

[135] S. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013.

[136] P. Vicinus, R. Schulz, M. Klose and R. Orgkmeister, "Voice Activity Detection within the Nearfield of an Array of Distributed Microphones," in *Speech Communication; 10. ITG Symposium*, Braunschweig, Germany, 2012.

[137] T. Matheja, M. Buck and T. Wolff, "Enhanced speaker activity detection for distributed microphones by exploitation of signal power ratio patterns," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, 2012.

[138] D. Pavlidi, A. Griffin, M. Pigut and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Hoboken, NJ, 2012.

[139] N. Grbic, S. Nordholm and A. Johansson, "Speech enhancement for hands-free terminals," in *2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces*, Pula, Croatia, 2001.

[140] M. Rahmani, N. Yousefian, A. Akbari and B. Ayad, "A real-time architecture for dual-microphone speech enhancement systems,," in *5th IEEE GCC Conference & Exhibition, 2009*, Kuwait City, 2009, pp. 1-5..

[141] Y. C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Audio, Speech, Lang. Process,* Vols. vol. 18, no. 7,pp. 1793–1805, Sep. 2010.

[142] S. Pasha, J. Donley, C. Ritz and Y. Zou, "Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017.

[143] M. A. Iqbal, J. W. Stoke, J. C. Platt, A. A. Surendran and S. L. Grant, "Doubletalk detection using real time recurrent learning," in *International*

Workshop on Acoustic Echo and Noise Control (IWAENC), Paris, France, 2006.

[144] R. Mammone, X. Zhang and R. Ramachandran, "Robust speaker recognition: a feature-based approach," *Signal Processing Magazine,* vol. 13, no. 5, p. 58, Sept. 1996.

[145] W. Chan, N. Zheng and T. Lee, ",Discrimination Power of Vocal Source and Vocal Tract Related Features for Speaker Segmentation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 6, pp. 1884-1892, Aug. 2007.

[146] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1650-1654, Dec 2002.

[147] H. Minnee, "Segment-oriented evaluation of speaker diarisation performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Shanghai, 2016.

[148] A. Miro, "Robust Speaker Diarization for meetings," in *Ph.D. dissertation, Universitat Politecnica de Catalunya,* 2006.

[149] 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.

[150] L. Wang, T, K. Hon, J. D. Reiss and A. Cavallaro, "An Iterative Approach to Source Counting and Localization Using Two Distant Microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 6, pp. 1079-1093, June 2016.