1-1-2016

Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays

Shahab Pasha
University of Wollongong, sp900@uowmail.edu.au

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

Yue-Xian Zou
Peking University, zouyx@pkusz.edu.cn

Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays

Abstract

This paper proposes a novel approach to detecting multiple, simultaneous talkers in multi-party meetings using localisation of active speech sources recorded with an ad-hoc microphone array. Cues indicating the relative distance between sources and microphones are derived from speech signals and room impulse responses recorded by each of the microphones distributed at unknown locations within a room. Multiple active sources are localised by analysing a surface formed from these cues and derived at different locations within the room. The number of localised active sources per each frame or utterance is then counted to estimate when multiple sources are active. The proposed approach does not require prior information about the number and locations of sources or microphones. Synchronisation between microphones is also not required. A meeting scenario with competing speakers is simulated and results show that simultaneously active sources can be detected with an average accuracy of 75% and the number of active sources counted accurately 65% of the time.

Disciplines

Engineering | Science and Technology Studies

Publication Details

S. Pasha, C. Ritz & Y. X. Zou, "Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016, 2016, pp. 1-6.

Detecting multiple, simultaneous talkers through localising speech recorded by ad-hoc microphone arrays

Shahab Pasha[1],Christian Ritz[1] and Y.X Zou[2]
[1]University of Wollongong, School of Electrical, Computer and Telecommunication Engineering, Wollongong, NSW, Australia
[2]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, China

***Abstract*— This paper proposes a novel approach to detecting multiple, simultaneous talkers in multi-party meetings using localisation of active speech sources recorded with an ad-hoc microphone array. Cues indicating the relative distance between sources and microphones are derived from speech signals and room impulse responses recorded by each of the microphones distributed at unknown locations within a room. Multiple active sources are localised by analysing a surface formed from these cues and derived at different locations within the room. The number of localised active sources per each frame or utterance is then counted to estimate when multiple sources are active. The proposed approach does not require prior information about the number and locations of sources or microphones. Synchronisation between microphones is also not required. A meeting scenario with competing speakers is simulated and results show that simultaneously active sources can be detected with an average accuracy of 75% and the number of active sources counted accurately 65% of the time.**

I. INTRODUCTION

Detecting multiple, simultaneously active talkers is essential to achieving high accuracy in source separation and speech diarization algorithms applied to multichannel (microphone array) recordings. Most multichannel speech separation algorithms use Direction of Arrival (DOA) for speaker discrimination. Conventional source separation methods (e.g. Principle Component Analysis (PCA) and Non-Negative Matrix Factorisation (NMF)) require prior information, such as the number of sources [1] and they usually focus on discriminating sources through estimates of the DOA. Binaural localisation methods require intraural information such as level and time differences along with Head Related Transfer Function (HRTF) to estimate the DOA which requires mathematical modeling of the HRTF or statistical modeling of binaural signals [2]. This mathematical modelling is computationally expensive and time consuming. Some recent research utilises Room Impulse Responses (RIRs) to localise sources by a single microphone by extracting cues that reflect the source DOA. This method is shown to be accurate, however it requires training for each setup and speaker which is not feasible for all scenarios [3].

As a practical acoustic scene analysis scenario a meeting room with seven participants has been analyzed in [4] and energy cues have been applied to localise randomly distributed speakers and microphones where at least 3 sources (out of seven) and microphones are collocated. Although it is shown that the proposed normalised energy cues can overcome issues such as different microphone/laptop gains and qualities, unknown microphone positions and asynchronised signal recordings, the assumption of microphones and sources being collocated is not realistic for all meeting scenarios.

More recently, spatial cues are derived from speech signals recorded by randomly distributed microphone arrays to discriminate sources [5]. Inter node (level difference) and intra node (local normalised recording vector) cues derived from microphone arrays are utilised within Watson and Dirichlet mixture models to discriminate sources based on their spatial locations. It is concluded that the performance of the proposed source separation approach is superior to the best node (a single recording device that may have more than one microphone attached forms a node) selection and comparable to centralized processing in terms of conventional blind source separation metrics where there are at least two microphones at each node.

It is shown that microphones located relatively close to each other have similar Magnitude Square Coherence (MSC) values and these values can be exploited to from local microphone clusters [6]. In other words, MSC values contain location cues. As the MSC relates to the relative distance between the active source and a microphone, in this research it is utilised as a distance-indicating feature to localise the active sources and detect the simultaneously active sources.

It was previously proposed by the authors that information derived from RIRs (time delay and gain attenuation) can indicate the relative distances between an active source and each microphone [7]. Although these derived cues are relative rather than being absolute, it is shown that if the room geometry is known, they can be utilised to localise an active source in a 2D plane accurately (assuming there is at least five randomly distributed single microphones in the room) [8]. The advantage of ad-hoc arrays for source location estimation will be investigated more in this research and simultaneous sources with identical DOAs that cannot be discriminated by relying on DOA estimation methods [9] will be detected through pin pointing the source location on a 2D plane.

The problem of room geometry reconstruction by utilising only one RIR is solved by researchers and the theorem about the uniqueness of the solution is stated [10]. Although it is possible to estimate the room geometry from one RIR in this research we assume that the room geometry is already known (reconstructed).

The main contributions of this work are:

- Source localisation by pin pointing the source on a 2D plane with no constraint on the microphones and sources locations (limitation of [4] where it is assumed source and microphones are collocated)
- Detecting simultaneously active sources that have identical DOAs but different distance relative to a recording location.

Section II of this paper explains the data model and introduces the derived distance cues. Section III is dedicated to active source localisation by exploiting relative distance cues. Section IV utilises the active source location information of each frame for detecting multiple, simultaneously active sources and compares the proposed method with state of the art approaches. The paper is concluded in section V where proposed future work is described.

II. DISTANCE CUES AND SOURCE/MICROPHONE LOCATIONS

In this section the data model of the recorded RIRs and speech signal by nodes randomly distributed in a room is firstly described. This is followed by a description of the two proposed cues: the intra node Magnitude Square Coherence (MSC) and the C_{50} or clarity measurement. It is assumed that microphone positions can be reliably estimated with knowledge of the room geometry using methods such as [11].

A. Distributed multi-node recording of reverberant speech

In a general meeting scenario where an unknown number of competing sources (N) are being recorded by a distributed microphone array of M nodes at unknown locations, the m^{th} node recording can be represented mathematically at each frequency f and time t in the short time Fourier transform domain as:

$$y_m(t,f) = \sum_{n=1}^{N} s_n(t,f) * h_{mn}(t,f) + v(t,f) + w_m(t,f) \quad (1)$$

where $y_m(t,f) = [y_{m,1}(t,f), \dots, y_{m,N_m}(t,f)]^T$ contains the multi-channel recording of all N_m microphones in the m^{th} node and $h_m(t,f) = [h_{m_1}(t,f), \dots, h_{m_{Nm}}(t,f)]$ is the Room Impulse Response (RIR) at each microphone's location within the m^{th} node. $v(t,f)$ and $w_m(t,f)$ are the diffuse noise and the interfering sources at the m^{th} node location, respectively. The goal is to extract informative relative distance cues from $y_m(t,f)$ and $h_m(t,f)$ that reflect the distance between the m^{th} node and the active sources. In this section intra Magnitude Square Coherence (MSC) and the clarity feature, C_{50}, are introduced and justified as relative distance cues.

	Distance from the active source	MSC	RT_{60}	Number of microphones
Node1	10cm	0.9637	600ms	2
Node2	0.5m	0.8988	600ms	2
Node3	3m	0.8194	600ms	2
Node1	10cm	0.9995	200ms	2
Node2	0.5m	0.9083	200ms	2
Node3	3m	0.8765	200ms	2

Table 1: MSC values at different points of a reverberant $\mathbf{10m \times 10m \times 3m}$ room. Obtained by dual microphone nodes with 10cm inter-channel distances

The room reverberation obtained from the RIRs can reveal the microphone locations in a room [11,12]. Microphones with similar RIRs (similar time delays and amplitudes) are located close and can be grouped together as a cluster [7]. In the time domain a room impulse response from (1) can be represented mathematically as a truncated train of L (e.g. 2000) samples:

$$h(t) = a_1\delta(t) + a_2\delta(t-\tau_1) + \cdots + a_L\delta(t-\tau_L) \quad (2)$$

The RIR representation of (2) can also be modelled in the form of:

$$h(t) = h_{direct}(t) + h_{early}(t) + h_{late}(t) \quad (3)$$

where h_{direct} is the direct path component (clean anechoic signal), h_{early} represents the early echoes arriving within 50ms (or 80 ms) and h_{late} represents the late echoes arriving after 50ms. These three components relatively change with the node-active source distance and this fact can be exploited for extracting distance cues from echoic RIRs.

It is clear that the recorded signals by the m^{th} node at the source position are highly correlated as the direct path component of (3) will be higher in magnitude than the early or late reflections. In contrast, as the node to source distance increases, the direct component reduces in magnitude compared to the early and late reflection components. This change in the active source-node distance will affect both MSC and the ratio of the direct path signal to the reverberant components (3). There are various measures of the direct to reverberant ratio and here the C_{50} or clarity measures is used, which has been shown to be a reliable estimate of speech quality, where it is proposed that this also correlates to source-to-microphone distances. The main advantages of using these teo features are that they are both independent from microphone's gains and delays and do not require time alignment or synchronisation. MSC is applicable to speech signals recorded within each dual microphone node and C_{50} is applicable to RIRs recorded by single microphones.

B. Intra node Magnitude Square Coherence (MSC)

Reverberation and interference recorded by each microphone are functions of its location in the room [11,12]

and as the microphones of each node are not exactly collocated they record slightly different echoes and interferences. 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:

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

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) [13]. 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.*" [14].

By applying the general equation of (1) to two microphones in the m^{th} node the signals can be modelled as:

$$\mathrm{y_{m,1}}(t,f) = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{s_n}(t,f) * \mathrm{h_{m,1,n}}(t,f) + \mathrm{v}(t,f) + \mathrm{w_{m1}}(t,f) \quad (5)$$

$$\mathrm{y_{m,2}}(t,f) = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{s_n}(t,f) * \mathrm{h_{m,2,n}}(t,f) + \mathrm{v}(t,f) + \mathrm{w_{m2}}(t,f) \quad (6)$$

And the MSC between these two microphones can be calculated by applying (4) to (5,6):

$$C_{y,m_1y_{m2}}(f) = \frac{|\varphi_{y,m_1y_{m2}}(f)|^2}{\varphi_{y,m_1y_{m1}}(f)\,\varphi_{y,m_2y_{m2}}(f)} \quad (7)$$

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 $v(t,f)$, $w_m(t,f)$ from (1) will become stronger (in terms of signal power) whereas $\varphi_{y,m_1y_{m1}}(f)\,\varphi_{y,m_2y_{m2}}(f)$ do not change with distance significantly.

The effect of the node-active source distance on MSC values in a reverberant room with two different RT_{60} values (200ms and 600ms) is presented in table 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 by distance from the active source.

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 recorded signals and the RIRs are not required.

Figure 1 illustrates the MSC values calculated for dual microphone nodes (with 10 cm distance) across a meeting room with two simultaneous active sources on a 2D grid with

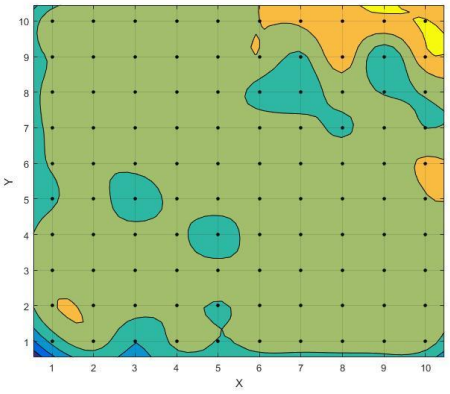

Figure 1 Source regions are detected as the regions with maximum MSC values, two simultaneous actives sources at (1m,1m,1m) and (9m,9m,1m)

one meter step sizes. This figure shows the challenge of picking the right threshold that indicates simultaneous sources are active. Three orange zones are highlighted as source areas in figure 1. All the sources and the nodes have the same height (1m). By analysing this figure Multi-talk can be detected correctly by the number of sources are counted incorrectly (3 instead of two).

C. C50 or clarity measurement

The C_{50} or Clarity measurement is the ratio of early to late reverberation expressed in dB. This measure is higher when the microphone-sources distance is relatively small and the recorded signal by the microphone is dominated by the direct path signal. In contrast it is lower when microphone-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-source distances and for calculating C_{50} the clean signal is not required (in contrast to the Direct to Reverberation ratio (DRR)) [15,16]. The C_{50} is defined in (8).

$$C_{50} = 10 \times \log\left(\frac{E_{direct}+E_{early}}{E_{late}}\right) \quad (8)$$

with $E_{Direct} = a_1\delta(n)$, $E_{early} = \sum_0^{t=50ms} h(n)$, and $E_{late} = \sum_{50ms}^{\infty} h(n)$ from (3) and n is the frame index. Using (2), C_{50} can be calculated for each RIR without synchronisation by:

$$C_{50} = 10 \times \log\left(\frac{\sum_0^{t=50ms} h(t)}{\sum_{50ms}^{\infty} h(t)}\right) \quad (9)$$

In this research 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 RIRs is required. Figure 2 shows a meeting room with two simultaneous active sources successfully detected by C_{50} values calculated across the room.

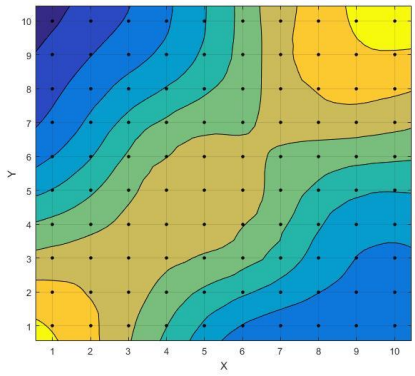

Figure 2: Clarity features calculated for 100 RIRs across a 10m,10m,3m room. Sources at (1m,1m,1m) and (9m,9m,1m)

III. 2D SOURCE LOCALISATION THROUGH SURFACE FITTING

The features explained in section II can be applied within a surface fitting method for source localisation [8] and if more than one active source (peaks of the surface) is detected and localised, multiple, simultaneous sources are assumed to be active. Source counting can be performed based on their Direction of Arrivals (DOAs) [17] however sources with identical DOAs cannot be discriminated by the proposed method of [17] and in some applications DOA estimation leads to detection of one virtual source instead of two sources at different angles [18]. In order to discriminate and count sources with identical DOAs herein active sources and their 2D locations (x and y coordinates) are determined. For the MSC feature speech frames of length 200 samples and for the C_{50} measurement RIRs of length 2000 samples (16K sampling frequency) are simulated. It is assumed in all the experiments that all the nodes and sources have the same height.

D. Multiple source localisation

Most source localisation and speech separation algorithms assume that sources are W-disjoint orthogonal [17] which means at each time-frequency component at most one source is active. The multiple source localisation algorithm proposed in this paper relates the extracted features (Section II) to spatial distances between active sources and all the nodes in a room with known geometry. Extracting features at each node's locations and the fitted surface across the room facilitate finding local maximums of MSC and C_{50}. These extremum points correspond to active sources locations estimated by utilising known node locations [11]. If the fitted surface obtains more than one local maxima, simultaneously active speech is predicted to have occurred. The local maxima zones approximately localise the active sources which can then be used within algorithms for separating spatially distributed sources.

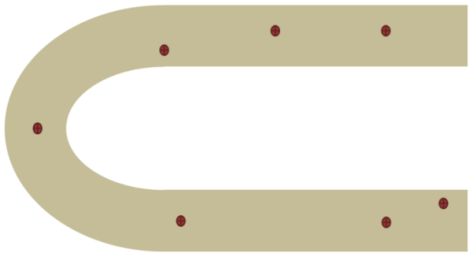

Figure 3: A conference table with 7 randomly distributed microphone nodes

IV. EVALUATION AND RESULTS

The meeting scenario of figure 3 is simulated with 3 to 7 nodes (one microphone per node for the clarity feature and nodes of two microphones with identical distances for MSC calculation are required) and 2 to 4 competing sources in a 10 m-by-10 m-by-3 m room with an RT_{60} of $600ms$. For each active source one utterance form IEEE NOIZEUS (clean data) is convolved with the RIR at its location to generate the reverberated mixture signal as (1). All the experiments are performed with the same speech database but different sources-nodes numbers and distances. Five different setups for each participants and sources numbers are simulated and average results are presented. Two types of measurements are defined to evaluate the proposed features (Section II). The first objective is to detect the frames with more than one active source (multi-talk detection) and the second objective is to count the simultaneously active sources for those frames. The largest number of competing participants (i.e. 4) with the largest number of nodes (i.e. 7) has the highest multi-talk detection rate as 7 nodes are spatially distributed in the room and collect more distance cues from sources and the extracted cues surface is fitted with a higher resolution. In addition, 4 simultaneous active sources generate more peaks (figure 1 and 2) compared with other scenarios so it is easier to detect the multi talk.

On the other hand, a higher number of simultaneous active sources yields a fitted surface with more random peaks which cannot be verified by a predefined threshold as active sources so the source counting success rate will drop with the number of simultaneous active sources (figure 5 and 7).

For R frames from S experimental setups with different number of simultaneous active sources and nodes, X represents the number of frames with more than one active speaker correctly detected as multi talk and Y represents the number of frames with correctly predicted number of active sources (for each setup). Therefore, multi-talk detection success rate (MT_{sr}) and source counting success rate (SC_{sr}) are calculated by (10,11) respectively.

$$MT_{sr} = \frac{1}{S}\sum_{i=1}^{S}\frac{X(i)}{R(i)} \quad (10)$$

$$SC_{sr} = \frac{1}{S}\sum_{S}\frac{Y(s)}{R(s)} \quad (11)$$

Figure 4 shows the success rate of multi talk detection using MSC values extracted from dual microphone nodes with 10 cm

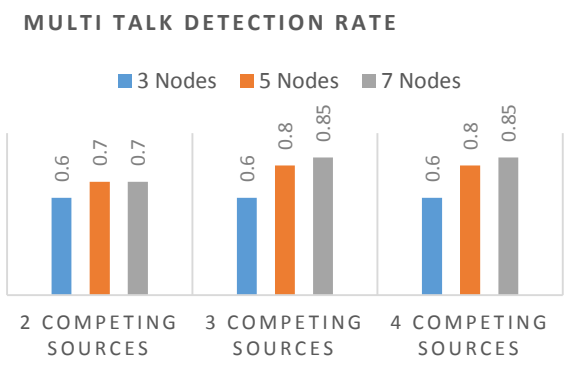

Figure 4: Multi-talk detection rate in a $\mathbf{10m \times 10m \times 3m}$ room, $\boldsymbol{RT_{60} = 600ms}$ based on the MSC features

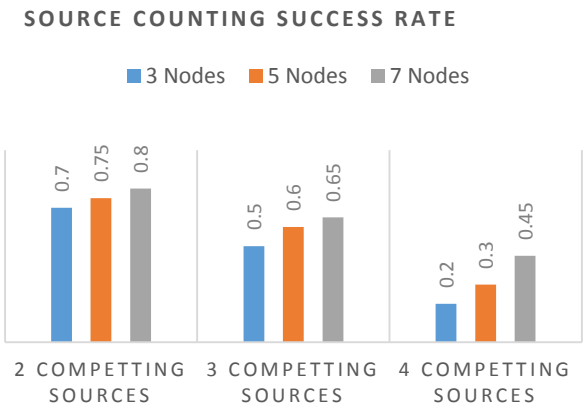

Figure 5: Source counting success rate in a $\mathbf{10m \times 10m \times 3m}$ room, $\boldsymbol{RT_{60} = 600ms}$ based on the MSC features

distance between microphones within each node and it is observed that multi-talk detection is more successful when there are more simultaneously active sources. On the other hand, as the number of simultaneously active sources increases,
the source counting accuracy decreases (Figure 5). It is noteworthy that multi-talk detection does not count the number of simultaneously active sources and determines that more than one source, two to four sources, are simultaneously active

Figure 6 and figure 7 show the same experiments with the C_{50}(clarity) feature. It is concluded that in most setups the clarity feature outperforms the MSC value except for the source counting with 4 competing sources.

The comparison between the MSC feature and the C_{50}feature show that the C_{50} feature is a more reliable feature for multi-talk detection and source location estimation feature for ad-hoc arrays when only 2 or 3 sources are simultaneously active. Although it is shown that the C_{50} feature can be estimated from speech signals [16] in this research RIRs are available at each microphone location. For calculating the C_{50} features at each microphone position (8), the RIRs can be recorded or extracted from the reverberant speech signals [19].

The length of the applied RIRs (2000 samples in this research) is determined by the RT_{60} time.

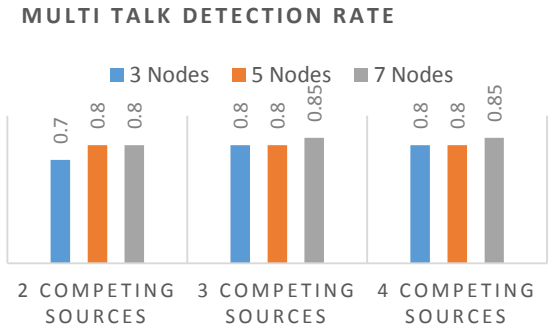

Figure 6: Multi-talk detection rate in a $\mathbf{10m \times 10m \times 3m}$ room, $\boldsymbol{RT_{60} = 600ms}$ based on the clarity features

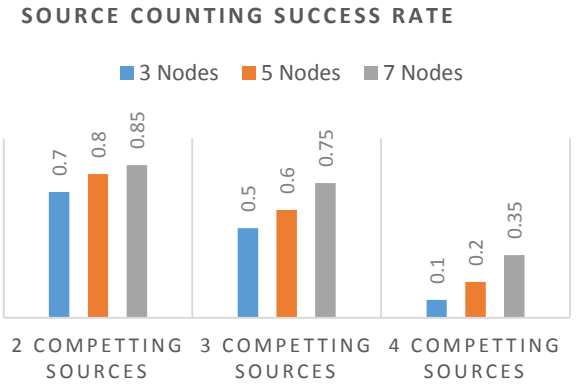

Figure 7: Source counting success rate in a $\mathbf{10m \times 10m \times 3m}$ room, $\boldsymbol{RT_{60} = 600ms}$ based on the clarity features

V. CONCLUSION

This paper proposed a novel multi-talk detection method through localisation of simultaneously active sources for multi-party meeting scenarios. The method is based on deriving distance cues from microphones spatially distributed across a room of known geometry and joint analysis of the derived features. The experiments of this research show the correlation between the extracted features and microphone-source distances. It is shown that C_{50} cues and speech Magnitude Square Coherence (MSC) can detect frames with more than one active speaker and localise active sources (up to four simultaneously active sources). It is concluded that C_{50} yields more accurate multi-talker detection and source counting rates but it cannot be applied to real time scenarios, on the other hand it is possible to apply MSC features to short frames and localise and count the simultaneously active sources during each frame. The analysis of a simulated meeting room by the proposed method achieved an average of 75% successful multi-talker detections however the success rate is a function of the chosen threshold. Exploiting the number of active sources and their location information along

with the state of the art source separation and speech diarization algorithms will be covered in future work.

REFERENCES

[1] T. J. Han, K. J. Kim and H. Park, "Location Estimation of Predominant Sound Source with Embedded Source Separation in Amplitude-Panned Stereo Signal," in *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1685-1688, Oct. 2015.

[2] Y. Murota, D. Kitamura, S. Koyama, H. Saruwatari and S. Nakamura, "Statistical modeling of binaural signal and its application to binaural source separation," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, pp. 494-498.

[3] R. Takashima, T. Takiguchi and Y. Ariki, "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 4295-4299.

[4] Z. Liu, Z. Zhang, L. W. He and P. Chou, "Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, 2007, pp. II-761-II-764.

[5] M. Souden, K. Kinoshita and T. Nakatani, "An integration of source location cues for speech clustering in distributed microphone arrays," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 111-115.

[6] I. Himawan, I. McCowan and S. Sridharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010, pp. 2814-2817.

[7] S. Pasha, Y. X. Zou and C. Ritz, "Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses," *International Conference on Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit*, Chengdu, 2015, pp. 84-88.

[8] 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.

[9] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, pp. 2814-2818.

[10] I. Dokmanić, Y. M. Lu and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 321-324.

[11] I. Dokmanić, L. Daudet and M. Vetterli, "How to localize ten microphones in one finger snap," *2014 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014, pp. 2275-2279.

[12] R. Parhizkar, I. Dokmanić and M. Vetterli, "Single-channel indoor microphone localization," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 1434-1438.

[13] Y. Ji, Y. Baek and Y. c. Park, "A priori SAP estimator based on the magnitude square coherence for dual-channel microphone system," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, pp. 4415-4419.

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

[15] Reuven Berkun, Israel Cohen, " Microphone array power ratio for quality assessment of reverberated speech" EURASIP journal on advances in signal processing. December 2015

[16] P. P. Parada, D. Sharma and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 4718-4722.

[17] D. Pavlidi, A. Griffin, M. Puigt and A. Mouchtaris, "Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193-2206, Oct. 2013.

[18] X. Zheng, C. Ritz and J. Xi, "Collaborative Blind Source Separation Using Location Informed Spatial Microphones," in *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 83-86, Jan. 2013.

[19] R. Takashima, T. Takiguchi and Y. Ariki, "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing- ICASSP 13*, Vancouver, BC, 2013, pp. 4295-4299.