Clustered Blind Beamforming from Ad-hoc Microphone Arrays

Ivan Himawan, *Student Member, IEEE,* Iain McCowan, *Member, IEEE,*
and Sridha Sridharan, *Senior Member, IEEE*

Abstract

Microphone arrays have been used in various applications to capture conversations, such as in meetings and teleconferences. In many cases, the microphone and likely source locations are known *a priori*, and calculating beamforming filters is therefore straightforward. In ad-hoc situations, however, when the microphones have not been systematically positioned, this information is not available and beamforming must be achieved blindly. In achieving this, a commonly neglected issue is whether it is optimal to use all of the available microphones, or only an advantageous subset of these. This paper commences by reviewing different approaches to blind beamforming, characterising them by the way they estimate the signal propagation vector and the spatial coherence of noise in the absence of prior knowledge of microphone and speaker locations. Following this, a novel clustered approach to blind beamforming is motivated and developed. Without using any prior geometrical information, microphones are first grouped into localised clusters, which are then ranked according to their relative distance from a speaker. Beamforming is then performed using either the closest microphone cluster, or a weighted combination of clusters. The clustered algorithms are compared to the full set of microphones in experiments on a database recorded on different ad-hoc array geometries. These experiments evaluate the methods in terms of signal enhancement as well as performance on a large vocabulary speech recognition task.

Index Terms

EDICS: SPE-GASR

Ivan Himawan, Iain McCowan, and Sridha Sridharan are with Speech and Audio Research Laboratory, Queensland University of Technology, S Block Lv 11, 2 George St. Brisbane, QLD, Australia.

Email: {i.himawan*, i.mccowan, s.sridharan}@qut.edu.au, Phone: +61 7 3138 9287, Fax: +61 7 3138 1516

Array signal processing, speech enhancement, speech recognition.

I. Introduction

A microphone array consists of multiple microphones that are combined to spatially filter a sound field. The geometrical configuration of the array allows filtering of desired signals, such as speech, from interfering signals, such as competing speech or room reverberation, based on the relative locations of the sources. This spatial filtering process is usually termed beamforming [1]. Microphone arrays offer an interesting alternative to close-talking microphones by facilitating more natural interaction by removing the constraint for the user to wear, or speak directly into, a microphone. In the context of an Automatic Speech Recognition (ASR) system, while close-talking microphones yield the highest accuracy, performance using microphone arrays is steadily approaching this through research [2], [3].

With advances in sensor and sensor network technology, there is potential for applications that employ ad-hoc networks of microphone-equipped devices collaboratively as a virtual microphone array [4]. In this new paradigm of pervasive computing, the user is free from intrusive devices while engaging in ordinary activities. While not an ad-hoc sensor network, conditions approaching this have in effect been imposed in recent NIST ASR evaluations on distant microphone recordings of meetings [5]. The NIST evaluation data comes from multiple sites, each with different and often loosely specified distant microphone configurations.

In scenarios where the microphone positions and likely source locations are not known, beamforming must be achieved blindly. The calculation of filters for most beamforming methods is dependent on two terms, the signal propagation vector and the spatial coherence of noise, and approaches can therefore be characterised by the way these terms are estimated. There are two general approaches to blindly estimate the steering vector for beamforming. The first is direct estimation without regard to the microphone and source locations. In the NIST meeting data evaluations, such an approach has been used for the Multiple Distant Microphone (MDM) condition in the AMI system [3] and the ICSI system [2], [6], among others. An alternative approach is to instead first determine the unknown microphone positions through array calibration methods [7], [8], and then use the traditional geometrical formulation for the steering vector [9]. Similarly the noise coherence matrix may either be estimated directly from signals during noise-only periods, or else by assuming some theoretical model of the noise field.

In purposefully-designed microphone arrays, the arrangement of microphones is carefully considered to achieve effective spatial enhancement for the likely speech source locations. In ad-hoc microphone arrangements, however, this will generally not be the case. An issue that has not received significant

attention in the research literature is therefore whether it is best to use all microphones in such a situation, or to select some optimal subsets of these. While the achievable array gain generally increases with the number of elements, it is commonly assumed that microphones are physically identical and located spatially close, and therefore have similar acoustic conditions.

Several simple practical ad-hoc scenarios have been discussed in the literature, generally resulting from meeting participants having a device with a single microphone element, such as a laptop, PDA or smartphone [4], [10], [11]. A similar, but distinct, scenario has been investigated in the NIST meeting transcription evaluations, in which data is recorded using a variety of randomly placed table-top microphones as well as small and large arrays placed in different locations in the room, with microphone position information of variable accuracy [5]. A further scenario that may be envisaged is audio surveillance using a number of audio sensors, distributed pseudo-randomly according to constraints of the environment. In this paper we consider the general problem of developing blind array processing methods capable of robustly handling variable numbers and relative placements of microphones in an environment, whether these microphones are from infrastructure, portable devices, arbitrarily placed sensors, or some combination of these. Although difficult to confirm empirically due to practical constraints, it is hypothesised that as the number of arbitrarily-placed microphones, people and noise sources increases, partitioning sensors into clusters will offer a more general solution than either selecting one microphone or using all of them.

When a microphone is arbitrarily placed, the signal acquired by the tranducer depends on its characteristics such as gain, directional response, the acoustic conditions of the room involving reverberation and the presence of localised noise sources. This means that some microphones would be better discarded due to their poor input SNR and signal quality. A particular consideration for blind beamforming is that large inter-microphone spacing may lead to erroneous Time Difference of Arrival (TDOA) computation, effectively causing delay inaccuracies in the steering vector of the beamformer. It is also undesirable to have spatial aliasing effects in the beamformer's directivity pattern. Therefore, it is hypothesised that using a cluster of microphones (ie, a sub-array), closely located both to each other and to the desired speech source may in fact provide more robust speech enhancement than the full array. In ad-hoc situations, the lack of prior knowledge of microphone and speaker locations means that the clustering of microphones and the selection of clusters must be done blindly.

Section II of this paper presents an overview of different approaches for beamforming when the microphone and speaker locations are unknown *a priori*. Following this, a novel approach for blindly clustering microphones is proposed in Section III. Microphones are first grouped into local clusters based on the Magnitude Squared Coherence (MSC) function during noise periods. The relative TDOA

between microphones is then used during speech periods as a basis for ranking the clusters according to their proximity to the source. Section III-E then presents two different methods of beamforming using these clusters, termed Closest Cluster (CC) Beamforming and Weighted Cluster Combination (WCC) Beamforming. The proposed methods are then evaluated and discussed in Section V on a speech database recorded on a variety of ad-hoc array geometries. This is followed by concluding remarks in Section VI.

II. Blind Beamforming Approaches

Many works on beamforming theory exist in the literature, such as [1]. This section reviews this theory as it relates to the particular case of blind microphone array beamforming, when microphone and speaker locations, as well as noise characteristics, are unknown *a priori*.

A. Beamforming Theory

Beamforming is an effective method of spatial filtering, differentiating desired signals from noise and interference based on their locations. Consider a desired signal received by an omni-directional microphone i sampled at discrete time t, in which the output is an attenuated and delayed version of the desired signal $a_i s(t-\tau_i)$ and noise v_i given by $x_i(t) = a_i s(t-\tau_i) + v_i(t)$. In the frequency domain by means of Fourier Transform, the signal model is written as:

$$x_i(f) = a_i s(f) e^{-j2\pi f \tau_i} + v_i(f) \tag{1}$$

The array signal model of N microphones is stacked into a vector and written as $\mathbf{x}(f) = s(f)\mathbf{d}(f) + \mathbf{v}(f)$, where $\mathbf{d}(f)$ represents the array steering vector which depends on the actual microphone and source location. In the near field, $\mathbf{d}(f)$ is given by [12]:

$$\mathbf{d}(f) = [a_1 e^{-j2\pi f \tau_1}, a_i e^{-j2\pi f \tau_i}, ..., a_N e^{-j2\pi f \tau_N}]^T, \tag{2}$$

$$a_i = \frac{d_{ref}}{d_i},\ \tau_n = \frac{d_i - d_{ref}}{c}, \tag{3}$$

where d_i and d_{ref} denote the Euclidian distance between the source and the microphone i, or the reference microphone, respectively. c is the speed of sound.

To recover the desired signal, each microphone output is weighted by frequency domain coefficients $w_i(f)$. The beamformer output is the sum of N weighted microphone outputs given by:

$$y(f) = \mathbf{w}^H(f)\mathbf{x}(f) \tag{4}$$

where $\mathbf{w}$ is a vector of size N x 1 of $w_i(f)$ and operator $(\cdot)^H$ denotes Hermitian transpose. The inverse Fourier transform results in time domain output signal $y(t)$.

In order to design beamformers with optimal noise suppression, the mean square of the noise at the array output may be minimized while giving an undistorted signal response in the desired look direction. The well known solution is usually termed the Minimum Variance Distortionless Response (MVDR) weights [13], given by:

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{\Gamma}_{\mathbf{vv}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}_{\mathbf{vv}}^{-1}\mathbf{d}} \tag{5}$$

in which the noise cross power spectral density has been normalised by assuming that the noise field is spatially homogeneous and expressed in terms of the coherence matrix, $\mathbf{\Gamma_{vv}}$ given by:

$$\Gamma_{v_iv_j}(f) \triangleq \frac{\Phi_{v_iv_j}(f)}{\sqrt{\Phi_{v_iv_i}(f)\Phi_{v_jv_j}(f)}} \tag{6}$$

where $\Phi_{v_iv_i}(f)$ and $\Phi_{v_iv_j}(f)$ are auto- and cross-power spectral densities, respectively.

From the MVDR solution, the beamformer filter weights are functions of two parameters which are: the array steering vector $\mathbf{d}$ and the noise coherence matrix $\mathbf{\Gamma_{vv}}$. The choice of how to estimate each of these parameters leads to different beamformer designs, as explained in the following sections.

The sensitivity of the MVDR beamformer to array peturbation increases as $||\mathbf{w}||^2$ increases [13]. Therefore to increase the robustness of the beamformer, the quadratic constraint $\mathbf{w}^H\mathbf{w} = T_o$ is usually applied in the optimisation problem for the beamforming weights formulation. This constraint is used to limit incoherent noise amplification and often referred to as a white noise gain constraint. Solving for $\mathbf{w}$, the solution is to add a scalar value μ to the main diagonal of coherence matrix in Equation 5.

B. Blind Estimation of the Array Steering Vector

If the true locations of the microphones and speaker are known, the array steering vector can be easily constructed using Equations 2 and 3. When these locations are unknown, however, the delay and gain scaling factors which characterise the incoming signal must be estimated from the received signals directly. This direct estimation of steering vector is referred to as the *blind estimation* approach. Note that this task is constrained to the case of a single active speaker in this article.

A two step approach is proposed to estimate the gain scaling factor a_i. First, the gain level on each channel due to signal acquisition is normalised. In order to do this, the normalisation factor β is first calculated and used to normalise the input level on each channel i, where:

$$\beta_i = \sqrt{\frac{E_{ref}}{E_i}} \tag{7}$$

The E_i is calculated as the average of the M lowest energy frames on each channel i. The highest energy channel is selected as the reference channel E_{ref}.

The gain scaling factor a_i is then estimated as the ratio of frame energies between the reference channel and each channel, for each time step corresponding to each new delay vector estimate. Assuming a single speaker, calculating the gain factor in this manner should reflect the speech level arriving on each microphone.

Following gain calibration, the appropriate delay for each microphone n, denoted as τ_i, can be determined by finding the Time Difference of Arrival (TDOA) with respect to the reference channel, which corresponds to the peak in Generalised Cross Correlation (GCC) function [14]. The PHAT weighting has been shown to emphasise the peak in the GCC seen at the time lag corresponding to the true source location, and de-emphasise those caused by reverberation and noise [14]. The PHAT-weighted cross correlation between microphone i and j is defined as:

$$\hat{G}^{ij}_{PHAT}(f) = \frac{x_i(f)x^*_j(f)}{|x_i(f)x^*_j(f)|} \tag{8}$$

The cross correlation will exhibit a global maximum at the lag value which correspond to the relative delay which is the TDOA between microphone i and j:

$$\tau(i,j) = arg\max_{\tau}\left(\hat{R}^{ij}_{PHAT}(\tau)\right) \tag{9}$$

where $\hat{R}^{ij}_{PHAT}(\tau)$ is the inverse Fourier Transform of Equation 8.

To have a robust TDOA estimate, the cross correlation measure can be calculated over a large window incorporating several frames, which constitutes a tradeoff between robustness and capturing rapid variations in the TDOA. To aid with the estimation accuracy of the delay, the input signals are pre-emphasised using a two-tap high pass filter $([1 - .95z])$ to attenuate the low frequency band of the signal where the TDOA estimation is less reliable due to the presence of low frequency noise. To increase the precision of the delays, time domain interpolation is performed to improve the temporal resolution of the peak estimation.

An alternative to this blind delay estimation approach is to first perform array shape calibration and then construct the steering vector directly using Equations 2 and 3 [15]. While information on microphone positions obtained from array shape calibration is useful in multi-speaker scenarios, previous work has shown no significant benefit over blind delay estimation in single speaker scenarios [9], [16]. As experiments in the current article focus on evaluating the clustered methods proposed in Section III with a single target speaker, the simpler blind estimation method presented above is used to obtain the beamformer steering vector throughout.

C. Estimation of the Noise Coherence Matrix

Different beamforming techniques are mostly characterised by the formulation they use for the noise coherence (or correlation) matrix in Equation 5. In general, the coherence matrix $\mathbf{\Gamma_{vv}}$ may be directly estimated using noise samples, either offline or adaptively, or calculated based on an assumed theoretical noise field model. In the following, three different methods for blindly estimating the noise coherence matrix are presented.

In environments such as offices or meeting rooms, the principal noises come from stationary sources such as computer fans and air conditioners. Assuming stationarity, the noise field may be measured once over a period of time. The noise auto- and cross-spectral densities are typically estimated using a recursive peridogram [17] as:

$$\Phi_{v_i v_j}(f) = \alpha \Phi^{'}_{v_i v_j}(f) + (1-\alpha) v_i(f) v_j(f) \tag{10}$$

where Φ is the density estimate for the current frame, and $\Phi^{'}$ is the estimate from the previous frame. The term α is a number close to unity which is given by $\alpha = exp(-T/\tau_\gamma)$, where T is the step size in seconds, and τ_γ is the decay time constant. Using the measured spectral densities and Equation 6, the measured $\mathbf{\Gamma_{vv}}$ can simply be substituted in the MVDR solution. In time-varying noise environments, adaptive algorithms such as variants of the Generalized Sidelobe Canceler [18], may instead be employed to achieve the same purpose.

An alternative to such direct noise field estimation is to instead assume a particular theoretical noise field model. In a spatially uncorrelated (or incoherent) noise field, the correlation of noise signals received at two microphones at any given spatial location is zero. Substituting an identity matrix as the coherence matrix, $\mathbf{\Gamma_{vv}} = \mathbf{I}$ in Equation 5, therefore yields the optimal beamforming weights for this case. This solution corresponds to the delay-sum beamformer. An incoherent noise model is often appropriate when the inter-microphone spacings are large and there are no major coherent noise sources in the environment.

In spherically isotropic (or diffuse) noise, two spatially separated microphones receive equal energy and random phase noise signals from all directions simultaneously. Such a noise field is a close approximation of moderately reverberant environments, such as inside offices or cars. To optimise the directivity factor, which is the ability of the array to suppress a diffuse noise, the diffuse noise coherence defined as:

$$\Gamma^{\text{diffuse}}_{v_i v_j} = \text{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) \tag{11}$$

is used in place of the noise coherence matrix in the MVDR solution, $\mathbf{\Gamma_{vv}} = \mathbf{\Gamma}^{\text{diffuse}}_{\mathbf{vv}}$. The beamformer obtained by maximising the directivity factor is commonly termed the superdirective beamformer.

The diffuse noise model requires knowledge of the distances between each pair of microphones. In a scenario when these are unknown *a priori*, these could first be estimated by fitting a diffuse noise model to the measured noise coherence. Based on previous work [15], a *blind superdirective beamforming* method is proposed in which the distance between each microphone pair is estimated by fitting Equation 11 to the measured noise coherence from Equation 6 in the least-squares sense:

$$\varepsilon_{ij}(d) = \sum_{f=0}^{fs/2} \left| \mathcal{R}\{\Gamma_{ij}(f)\} - \text{sinc}\left(\frac{2\pi f d}{c}\right) \right|^2 \tag{12}$$

$$d_{ij} = \underset{d}{\text{argmin}}\, \varepsilon_{ij}(d) \tag{13}$$

where d_{ij} is the distance between microphones i and j.

Table I summarises three different blind beamforming methods based on the above methods for estimating the noise coherence matrix. All beamformers use the same direct estimation of gain and TDOA for the steering vector, as explained in Section II-B.

TABLE I

Different blind beamforming methods

Beamformer	**Noise Coherence Matrix Estimation**
Blind Delay-Sum (DS)	Spatially uncorrelated noise model
Blind Superdirective (SD)	Diffuse model using distance estimate
Blind MVDR	Estimation from noise samples

III. Clustered Blind Beamforming

This section proposes a novel clustered approach for beamforming from ad-hoc microphone arrays. As mentioned in the introduction to this article, for ad-hoc microphone arrangement, a particular issue is that large inter-microphone spacings may lead to erroneous TDOA computation, effectively causing delay inaccuracies in the steering vector of the beamformer. In such situations, using a subset of microphones instead of the full array may prove to a more robust approach. Theoretical justifications for this hypothesis are analysed in Section III-A below, motivating methods for selecting a cluster of microphones that are both close to each other, as well as to the speaker.

With this motivation, first, an inter-microphone proximity measure is proposed based on the Magnitude Squared Coherence (MSC) function during noise periods. Two different algorithms for forming clusters

based on this measure are proposed, followed by a method for ranking the formed clusters according to their proximity to the desired speaker. Using the ranked clusters, beamforming may then be achieved by applying the blind methods from Table I to either the closest cluster, or else a weighted combination of all clusters.

A. Theoretical Justification

The array gain G measures the SNR improvement between one sensor and the output of the whole array. This is given by:

$$G = \frac{|\mathbf{w}^H \mathbf{d}|^2}{\mathbf{w}^H \Gamma_{\mathbf{vv}} \mathbf{w}} \tag{14}$$

where the signal model and $\mathbf{d}$ have been given in Equation 1 and 2 respectively. For delay-sum beamformer assuming noise is uncorrelated from sensor to sensor (i.e $\Gamma_{\mathbf{vv}} = \mathbf{I}$ thus $\mathbf{w} = \frac{1}{N}\mathbf{d}$ in Equation 5) and the steering delays are matched to the wave's direction of propagation, the array gain can be simplified to:

$$G = ||\mathbf{w}||^{-2} = \frac{1}{\sum_{i=1}^{N} |w_i|^2} = N \tag{15}$$

where N is the number of sensors in the array.

Adding one further microphone to the array, the array gain which consists of previously N microphones can be written as:

$$G = \frac{\left| \sum_{i=1}^{N} \frac{a_i}{N+1} + \frac{\beta}{N+1} e^{j2\pi\Delta\tau} \right|^2}{\sum_{i=1}^{N+1} |w_i|^2} \tag{16}$$

where $\Delta\tau$ is the small variation in delay estimates due to mismatches of the steering delays to the wave's direction of propagation. Assuming each microphone gain has $\beta = a_i = 1 \ \forall i$ and $\Delta\tau = 0$ for including a microphone with matched delays, leading to in-phase addition of the complex frequency signals, the maximum gain achieved is $G_{hi} = N + 1$. If the addition of a microphone causes phase difference of π radians (i.e. antiphase addition), the lower bound of array gain for including this one microphone is equal to:

$$G_{lo} = \frac{\left| \frac{N}{N+1} - \frac{1}{N+1} \right|^2}{1/(N+1)} = \frac{(N-1)^2}{N+1} \tag{17}$$

where the complex exponential in the numerator of Equation 16 is equal to $e^{-j\pi} = -1$ due to the $\theta - \pi$ phase shift from the sum of other microphones signals with matched delays (in which θ is assumed equal to 0).

In this case of including an additional one microphone, the array gain will vary between $G_{lo} \leq N \leq G_{hi}$. As a concrete example, for $N = 9$, the theoretical gain from a single additional microphone with

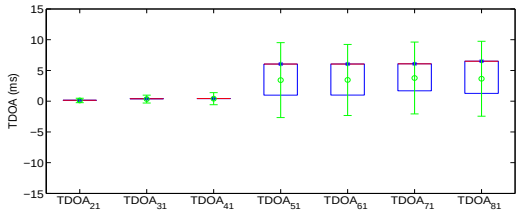


Fig. 1. Box plot of the TDOA values applied to 8-element microphone arrays. The lines are drawn for the lower quartile, median, and upper quartile values. Whiskers extend to one standard deviation above and below the mean (circle) of the data. The expected ground truth of TDOA values are shown in asterisk. In the x axis, TDOA_{21} corresponds to the TDOA between the second closest and the closest microphone in the array, similary, TDOA_{31} corresponds to the TDOA between the third closest and the closest microphone in the array, and so on.

incorrect phase alignment may vary from $G = 6.4$ to $G = 10$. It is then a robust strategy to exclude a microphone with large delay errors since the overall gain may be less than N.

Figure 1 shows TDOA values calculated from the speech frames between a 8-element microphone array. The TDOAs are calculated between a reference microphone to 3 closely located microphones (within 20cm of distance) and other 4 microphones of about 2m from the reference microphone. The empirical observation shows that TDOAs are accurate between closely spaced sensors, with accuracy decreasing with greater distance between the pair used to calculate TDOA (e.g. TDOA_{51}, TDOA_{61}, TDOA_{71}, and TDOA_{81}) due to the large distance between the reference microphone and the 5^{th}, 6^{th}, 7^{th}, 8^{th} closest microphone to the speaker respectively.

Consider the effect of variations in microphone gain a, $0 < a < 1$ in Equation 1. From Equation 3, a relates to the distance from source to the microphone in an ideal propagation model. Assuming all microphones in a cluster have matched delays, the overall gain is the sum over gain factors for a cluster,

$$G = \frac{\left|\sum_{i=1}^{N} \frac{a_i}{N}\right|^2}{\sum_{i=1}^{N} |w_i|^2} = \frac{\frac{1}{N^2}\left|\sum_{i=1}^{N} a_i\right|^2}{1/N} = \Lambda N \tag{18}$$

where $\Lambda = \frac{1}{N^2}\left|\sum_{i=1}^{N} a_i\right|^2$, in which Λ has value between 0 and 1. From Equation 18, the total gain of a cluster depend on the two variables, the gain factor a and the size of cluster N.

Assume an array of N microphones consists of two clusters having N_1 and N_2 number of microphones respectively (i.e. $N = N_1 + N_2$), the ratio of cluster gains,

$$\frac{G_1}{G_2} = \frac{\Lambda_1 N_1}{\Lambda_2 N_2} = \frac{\frac{1}{N_1}\left(\sum_{i=1}^{N_1} \frac{d_{ref}}{||s - m_{1i}||}\right)^2}{\frac{1}{N_2}\left(\sum_{i=1}^{N_2} \frac{d_{ref}}{||s - m_{2i}||}\right)^2} \tag{19}$$

where the sensor gain a has been replaced with the ratio between distance from source to the reference microphone d_{ref} and the distance from source to each microphone in the cluster. Here, the d_{ref} is the distance from source to the closest microphone in the array which is the same for both clusters. From Equation 19, the ratio of cluster gains depends in general on the size of cluster and inversely on the average distance of that cluster from the sound source.

From this analysis, robust gain will be achieved when:

1) Only elements with accurate inter-microphone delays estimates are included, and
2) The gain factor a is close to 1 which means that the microphones are closer to the source.

B. Inter-Microphone Proximity Measure

The MSC between two microphone signals i and j at discrete frequency f, $C_{ij}(f)$ is calculated in the following manner:

$$C_{ij}(f) \triangleq \frac{|\Phi_{v_i v_j}(f)|^2}{\Phi_{v_i v_i}(f)\Phi_{v_j v_j}(f)} \tag{20}$$

where the auto- and cross-power spectral densities are estimated as in Equation 10.

Environments such as offices or meeting rooms are usually considered to represent diffuse noise fields. The MSC function between two microphones in a diffuse noise field can be modelled as [19]:

$$C_{ij}^{\text{diff}}(f) = \text{sinc}^2\left(\frac{2\pi f d_{ij}}{c}\right) \tag{21}$$

According to this model, the noise coherence between two microphones depends principally on the distance d_{ij} between them. The first minimum of this MSC function occurs at:

$$f_m(i,j) = \frac{c}{2d_{ij}} \tag{22}$$

and beyond this frequency the coherence approaches zero.

This dependence of the diffuse noise coherence on the distance can be used to indicate how close two microphones are, since closely-spaced microphones will have wider main lobes in the coherence function compared to distantly-spaced pairs as shown in Figure 2. To give a measure of overall coherence between microphones, and hence a measure of their proximity, the MSC may be integrated across frequencies:

$$T_{MSC}^{ij} = \sum_{0}^{f_{max}} C_{ij}(f) \tag{23}$$

where the summation range is limited by f_{max} to improve robustness. Typically, this may be set to be $f_m(i,j)$ from Equation 22, as the measured coherence function often varies significantly from the theoretical model for frequencies much beyond the main lobe.

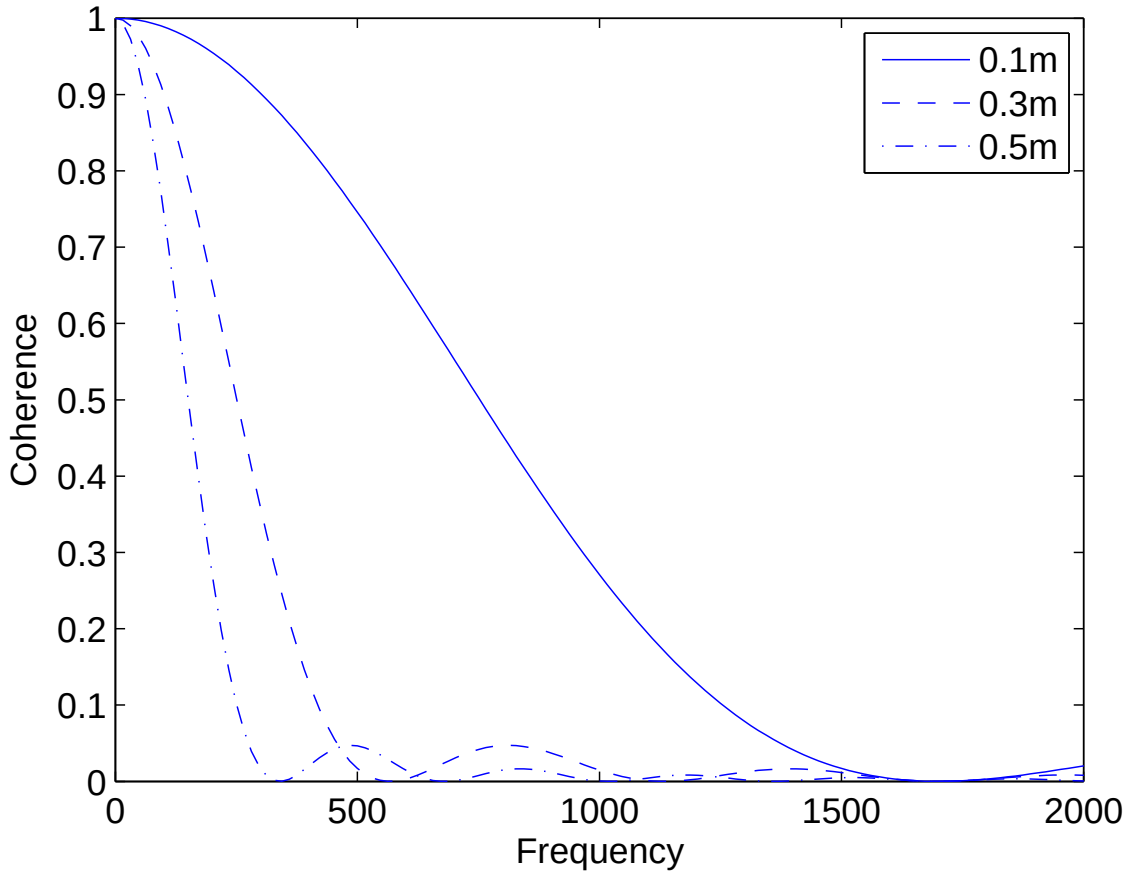


Fig. 2. *Theoretical magnitude squared coherence as a function of distance between two microphones. The sampling frequency is 16kHz.*

C. Microphone Clustering

1) Rule-based Clustering: In order to cluster microphones, the measure in Equation 23 may be compared to some threshold value to determine if two microphones are sufficiently close to each other. A threshold value can be computed to correspond to a desired distance, d_ε, by using the theoretical coherence model from Equation 21 and summing up to a threshold frequency $f_\varepsilon = c/2d_\varepsilon$ (corresponding to its first minimum):

$$T_\varepsilon = \sum_{l=0}^{f_\varepsilon} \text{sinc}^2 \left(\frac{2\pi f_l d_\varepsilon}{c} \right) \tag{24}$$

The measured value for T_{MSC}^{ij} may then be compared to this *intra-cluster* threshold T_ε. If $T_{MSC}^{ij} \geq T_\varepsilon$, then microphones i and j are grouped in the same cluster, otherwise they belong to separate clusters.

This conservative binary classification is evaluated over all microphone pairs to form an initial set of clusters in which all microphones are within the specified distance of all others in the cluster. A subsequent merging pass then combines clusters for which at least one inter-cluster microphone pair is within a more restrictive distance. Without this second pass, the method would only capture clusters with maximum extent defined by the distance threshold - the merging step allows larger clusters to be formed from these if continuity exists at the boundaries. The proposed clustering algorithm is as follows:

1) Assign $b_{ij} = 1$ if $T_{MSC}^{ij} \geq T_{\varepsilon}$, for $i, j = 1, \cdots, N$.
2) Compute $B_i = \sum_{j=1}^{N} b_{ij}$, for $i = 1, \cdots, N$.
3) Select the microphone belonging to the most pairs as the centre microphone of the first cluster, ie $\hat{m}_1 = \arg\max_i B_i$.
4) Form cluster 1, Q_1 with $Q_1 = \{j | b_{\hat{m}_1 j} = 1\}$.
5) Remove microphones belonging to cluster 1 from consideration, then repeat the above steps to form clusters $k = 2 : K$ until all microphones have been assigned a cluster.
6) Once the set of initial clusters $Q_{1:K}$ is formed, a merging pass is conducted. Two clusters are merged if $T_{MSC}^{ij} \geq T_{\kappa}$, where microphone i belongs to one of the clusters and j belongs to another, and T_{κ} is an *inter-cluster* threshold calculated using a more restrictive distance criteria d_{κ} in Equation 24
7) In the case the above steps result in the formation of any single-element clusters, these may be merged with the closest cluster if a relaxed inter-cluster threshold is satisfied.

2) Spectral Clustering: As an alternative to the rule-based clustering algorithm explained above, the use of spectral clustering is also investigated. Spectral clustering finds group structure in data by using the spectrum of a similarity matrix. The algorithm is based on spectral graph theory and has been widely used for pattern recognition [20]–[22]. Here spectral clustering is applied on the coherence measure to automatically group microphones in the spatial domain without need for hard decision rules. The key to the partitioning is the construction of the similarity matrix as a weighted adjacency matrix S modeling the neighborhood relationship between data points. To cluster microphones using the MSC measure, this matrix may be defined as:

$$s_{ij} = exp\left(-\alpha(\frac{1}{L_{f_{\varepsilon}}} T_{MSC}^{ij} - 1)^2\right) \tag{25}$$

where $L_{f_{\varepsilon}}$ is the DFT length up to frequency f_{ε}, used to normalise the sum of MSC values T_{MSC}^{ij}, and α is the scaling parameter in the Gaussian filtering function.

The algorithm of spectral clustering investigated in this article follows [22], [23]:

1) Construct matrix $S = (s_{ij})_{i,j=1,\cdots,N}$ defined in Equation 25.

2) Compute $D = diag(d_1, \cdots, d_i)$, where d_i is defined as:

$$d_i = \sum_{j=1}^{N} s_{ij} \tag{26}$$

3) Solve the generalised eigenproblem $(D - S)v = \lambda Dv$ for the first k eigenvectors, where k is the number of clusters.
4) Form matrix $V \in \mathbb{R}^{N \times k}$ containing the vectors $v_1, \cdots, v_k$ as columns.
5) Partition matrix V into vector $y_i \in \mathbb{R}^k$ for $i = 1, \cdots, N$ which correspond to the i^{th} row of V.
6) Cluster $y_1, \cdots, y_N$ into k clusters with k-means algorithm in $\mathbb{R}^k$ into clusters $A_1, \cdots, A_k$.
7) Microphones will be clustered into $Q_1, \cdots, Q_k$ with $Q_i = \{j | y_j \in A_i\}$.

The eigengap which indicates the stability of the eigenstructure can reveal the number of clusters [20]. In the ideal case of k completely disconected clusters, the eigenvalue 0 has multiplicity of k and there will be a gap to the $(k+1)^{th}$ eigenvalue which is greater than zero. Thus the sudden increase of the k^{th} eigengap ξ_k,

$$\xi_k = |\lambda_{k+1} - \lambda_k| \tag{27}$$

where λ_k is the k^{th} smallest eigenvalue, may indicate the true number of clusters. Similar to step 7 in the rule-based clustering algorithm, in the case of formation of any single-element clusters, these may be merged with the closest cluster if a relaxed inter-cluster threshold is satisfied.

Since the spectral clustering algorithm involves eigendecomposition, the computation time could be an issue for a large matrix due to clustering the large number of microphones compared to rule-based approach which only involves conservative binary classification evaluated over all microphone pairs.

D. Ranking Clusters

Once microphone clusters have been formed according to one of the above methods, it is necessary to somehow select which cluster, or clusters, will be used for beamforming. While a range of signal quality criteria may be used to achieve this more generally, this section proposes one method to achieve this based on the assumption that the best clusters will be those located closest to the speaker of interest.

Assuming a known period of speech from a single person, the delay in receiving a sound wave between clusters indicates their relative distance to that speaker. The detailed steps to rank clusters based on their proximity to the speaker are outlined below. The algorithm considers both the proximity of the closest microphone within each cluster (using the TDOAs between a reference microphone from each cluster), as well as the spatial extent of the cluster (using a measure of the spread of TDOAs within each cluster).

1) *Find the closest microphone to the speaker within each cluster and set it as reference* m_k. To do this, for cluster k, choose an initial arbitrary reference microphone m'_k and calculate $\tau(i, m'_k)$ for each microphone i in the cluster during speech-only segments. Update the reference microphone for the cluster to be the closest microphone by selecting the one having the minimum TDOA, ie $m_k = \arg\min_i \tau(i, m'_k)$.

2) *As a measure of cluster spread, calculate the mid-range TDOA offset for each cluster* δ_k *relative to its reference microphone*. To do this, for cluster k, calculate the TDOA of each other microphone with respect to the reference microphone selected in the previous step, ie $\tau(i, m_k)$ for each microphone i in the cluster. Set the mid-range TDOA offset for the cluster to be half of the maximum TDOA, ie $\delta_k = \frac{1}{2}\max_i \tau(i, m_k)$.

3) *Find the reference cluster* c_{ref} *as the one with its reference microphone closest to the speaker.* To do this, first choose an arbitrary reference cluster k_r and calculate the set of TDOAs between its reference microphone and the reference in all clusters k, $\tau_(m_k, m_{k_r})$. Update the reference cluster to be the one which has the minimum TDOA , ie $c_{ref} = \arg\min_k \tau(m_k, m_{k_r})$.

4) *Form the final proximity score* D_k *for each cluster by compensating the inter-cluster TDOAs with the cluster mid-range offsets.* Given the set of mid-range offsets for all clusters, δ_k, and the set of TDOAs with respect to the reference cluster c_{ref}, $D_k = \tau(m_k, m_{c_{ref}}) + \delta_k - \delta_{c_{ref}}$.

The clusters may then be ranked according to their proximity to the speaker according to the score D_k. Note that it is possible that this score may be negative, indicating that the reference cluster from step 3 did not turn out to be the closest cluster when considering the mid-range offsets. Considering these mid-range TDOA offsets is a means to compensate for the differing spatial extents of clusters. For instance, while some clusters may have a reference microphone that is close to the speaker, they may also be large clusters with other microphones that are quite far from the speaker.

E. Clustered Beamforming

Following clustering of microphones and the subsequent ranking of clusters according to their proximity to the speaker, blind beamforming may be performed for speech enhancement and recognition. This section presents two methods for beamforming using the clustering information: beamforming using the

closest cluster only, and forming a weighted combination over multiple clusters.

1) Closest Cluster Beamforming (CC): In the spatially distributed microphones scenario, a speaker is usually relatively close to one or subset of microphones in the same time. The simple strategy for speech enhancement in this situation is to choose the signal from microphones which are closest to the speaker or ones which have the best SNR. Assuming identical microphone gains and no obstructions in the line of sight from the speaker to microphones, the microphones which have the highest SNR will be the ones which are closest to the speaker.

2) Weighted Cluster Combination Beamforming (WCC): While the closest cluster may have the best SNR, it is hypothesised that contributions from every cluster may in fact improve the overall performance. In this article, delay-sum beamforming is used to combine the beamformed signals of each clusters. To calculate the overall combination, each cluster output is phase-aligned with an estimated fixed delay before a weighted summation.

Given the $y_k(f)$ is the beamforming output of cluster k as defined in Equation 4, the weighted combination of cluster beamforming output is obtained from

$$z(f) = \sum_{k=1}^{K} {w'_k}^H(f) y_k(f) \tag{28}$$

The weight $w'_k(f)$ is defined as

$$w'_k(f) = \alpha_k e^{-j2\pi f \tau_k} \tag{29}$$

where α_k represent the contribution of each cluster's k and τ_k is the relative delay of each cluster's output signal with the respect to the reference cluster.

IV. Experimental Setup

Experiments were conducted in a meeting room of size 5.3m x 4.4m x 2.7m, as shown in Figure 3. The main sources of noise were a PC, laptop, a projector, and air conditioning. To experiment with different ad-hoc array geometries, microphones were mounted in varying positions on two cork boards placed on top of the meeting table. A total of 8 microphones (AKG C417 omnidirectional condenser microphones) were used for each ad-hoc geometry. The microphones were recorded using a MOTU 8pre audio interface and SONAR 8 software, allowing simultaneous, fully synchronised playback and recording of multiple audio channels.

Evaluating methods to deal with ad-hoc microphone placements clearly requires a trade-off between the desire to test as many configurations as possible, and practicalities of conducting and analysing a large body of experiments. To make some coherent conclusions from this initial investigation into microphone

clustering, four data sets were recorded to investigate algorithm behaviour in different circumstances, as described in the following sub-sections. The recordings were constrained to a typical meeting room deployment, with 8 microphones placed in different configurations on a meeting table. While endeavouring to explore a variety of configurations in the spirit of ad-hoc situations, scenarios focus on those which highlight the potential benefits and limitations of clustering.

A. Data Set A

The first data set was collected to evaluate the proposed microphone clustering algorithm and consisted of noise recordings from 20 different ad-hoc array geometries. The 8 microphones were placed within an area of approximately 1m x 1m. While constrained, this area serves to test the algorithm behaviour around the range of the inter-cluster distance parameters designed in the rule-based algorithm. Within those recordings, microphones were positioned on the meeting table to reflect variants of ad-hoc positioning into clearly separated cluster of microphones, closely spaced clusters, or just a single large cluster.

B. Data Set B

The second data set was collected to evaluate the clustered approach for blind beamforming from ad-hoc arrays. The microphones and speakers were configured such that there were clearly separated microphone subsets with a speaker positioned relatively closer to one of these. To achieve this, two clusters with an equal number of microphones were placed near opposite edges of the table, as shown in Figure 3. Three different speaker locations were used: Position S1 where the speaker is facing all microphones but is closer to one cluster than the other, Position S2 where the speaker faces only one of the clusters, and Position S3 where the speaker faces both clusters at varying angles. A total of 30 utterances from the WSJCAM0 evaluation set [24] were recorded by playing the clean speech files through a loudspeaker at the various speaker positions. To simulate random microphone placement within these constraints, the microphone positions were rearranged for every 10 recorded sentences.

To simulate the effect of localised noise sources such as a competing speaker for a given ad-hoc configuration, babble noise was played during utterance recordings. The babble noise was taken from NOISEX database [25] and the volume was set to be approximately 10dB lower than the main speaker. Three different noise source positions were used in this experiment: at Position N1 and N2 where the noise source is in close proximity to one of the clusters, and at Position N3 where the noise source is located with large distance to any clusters. For the completion of experiment, the recording session is

also performed without the localised noise. The illustration of the ad-hoc array setup for Data Set B is shown in Figure 3.

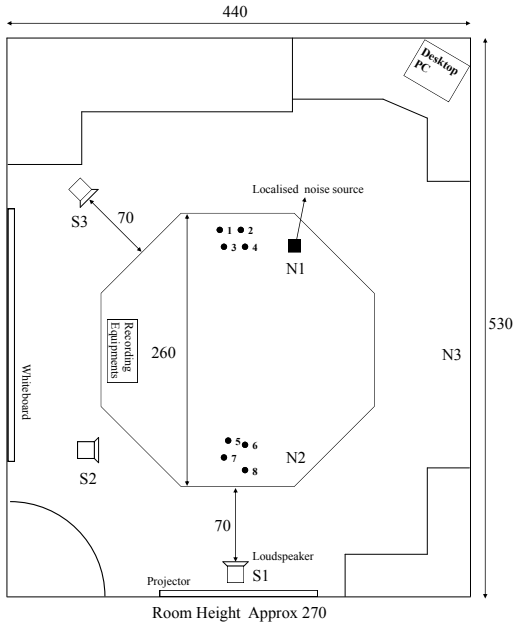


Fig. 3. *Meeting room used in Data Sets B and D (measurements in cm).*

C. Data Set C

The third data set serves the same purpose as Data Set B. However in this set of experiments, the number of clusters and speaking orientations were configured differently. The illustration of the ad-hoc array setup for Data Set C is shown in Figure 4. There are 3 clusters of microphones with 2 of these containing 3 microphones and 1 cluster containing 2 microphones. The two clusters with 3 microphones were placed near opposite edges of the edge of the table, and the cluster with 2 microphones was placed on the right edge of the table. The speaker at Position S1 is facing all microphones but is relatively closer to one cluster than the others, the speaker at Position S2 is facing all microphones from different angles and close to two of the clusters, and the speaker at Position S3 is also facing all microphones with somewhat equally large distances to every cluster. Four localised noise conditions were positioned in the same manner as in Data Set B.

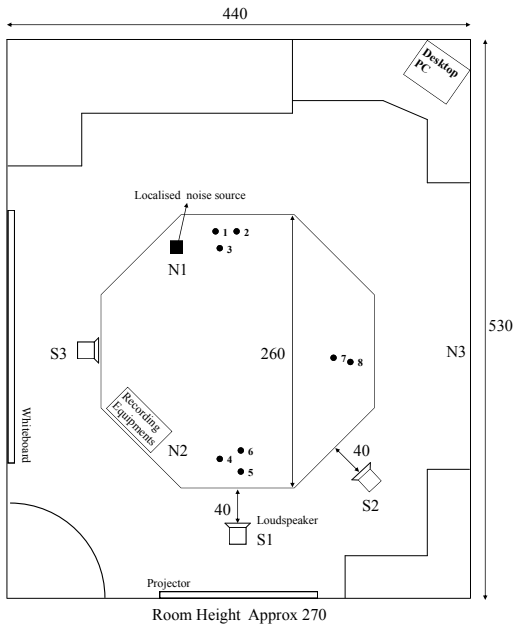


Fig. 4. *Meeting room used in Data Set C (measurements in cm).*

D. Data Set D

This data set was recorded for the purpose of evaluating speech recognition performance. The ad-hoc array arrangement is similar to that described in Data Set B (Figure 3) with three speaker positions and no localised noise sources. The data consists of 182 utterances from the evaluation set of the WSJCAM0 corpus for each speaker orientation. For every 10 recorded utterances, microphone positions were rearranged to simulate random placements.

V. Experimental Results

A. Microphone Clustering Evaluation

Compared to classification, clustering can be difficult to objectively evaluate, as often there is no correct grouping that can be considered as ground-truth. To evaluate use of the noise coherence feature, the results of the proposed automatic *cluster* algorithm based on noise recordings were therefore compared to *sub-arrays* formed by applying the same algorithm to ground-truth distances between known microphone positions. The comparison is illustrated for three ad-hoc geometries in Figure 5. For the rule-based clustering algorithm, the intra-cluster distance threshold d_ε and inter-cluster distance threshold d_κ are set to be 30cm and 20cm respectively. Single element clusters are merged to the nearest cluster if they satisfy

a relaxed threshold of 50cm. For the spectral clustering algorithm, the α value used in Equation 25 is 15.

To measure overall performance, the purity measure used in speaker clustering literature is adapted to the current context [26], [27]. Dual purity measures are used to evaluate how well closely the automatic *clusters* match the 'ground-truth' *sub-arrays*, and vice versa. First, define

N_s: total number of true *sub-arrays*.

N_c: total number of found *clusters*.

N: total number of microphones.

n_{ij}: total microphones in *cluster* i that are from *sub-array* j.

$n_{i.}$: total microphones in *cluster* i.

$n_{.j}$: total microphones in *sub-array* cluster j.

The purity of a *cluster* $p_{i.}$ is defined as:

$$p_{i.} = \sum_{j=1}^{N_s} n_{ij}^2 / n_{i.}^2 \tag{30}$$

and the average *cluster* purity acp is:

$$acp = \frac{1}{N} \sum_{i=1}^{N_c} p_{i.} . n_{i.} \tag{31}$$

Similarly, the *sub-array* purity $p_{.j}$ and asp are defined as

$$p_{.j} = \sum_{i=1}^{N_c} n_{ij}^2 / n_{.j}^2 \tag{32}$$

and

$$asp = \frac{1}{N} \sum_{j=1}^{N_s} p_{.j} . n_{.j} \tag{33}$$

The asp gives a measure of how well a sub-array matches only one cluster, and the acp gives a complementary measure of how well a cluster matches only one sub-array. These scores can be combined to obtain an overall score, $K = \sqrt{acp \times asp}$. Table II presents the average score results for acp, asp, and K for 20 different ad-hoc geometries recorded in Data Set A for both the rule-based and spectral clustering algorithms described in Section III-C. Four illustrative examples are plotted in Figure 5.

The high purity measures in Table II indicate that both clustering algorithms using the magnitude squared coherence feature well-approximate the sub-arrays formed from known microphone positions. When the separation between clusters is clear, as in Figure 5(i), the algorithm succeeds.

The lower value of $a\bar{c}p$ compared to $a\bar{s}p$ in Table II for recorded geometries shows that the automatic measure tends to create larger clusters for microphone separations near the threshold value, indicating that

TABLE II

Clustering results in terms of average acp, asp and K over 20 ad-hoc geometries.

Clustering technique	$a\bar{c}p$	$a\bar{s}p$	$\bar{K}$
Rule-based clustering	0.855	0.945	0.887
Spectral clustering	0.867	0.968	0.906

the measured coherence tends to exceed that predicted by the diffuse model in this particular environment. Examples of this occurring are shown in Figure 5(ii)-(iii). For Figure 5(iii), however, the *spectral* clustering separates them as two distinct clusters since elements in each are tightly spaced making the cut between clusters more obvious.

The array geometry in Figure 5(iv) presents a case when all microphones are somewhat continuously close to each other. Assuming known microphone positions, the algorithm merges all microphones together to form a single cluster. When the noise signal is used, the *rule-based* algorithm forms 2 distinct clusters as the measured coherence between microphone 5 and 6 is below the threshold in this particular situation. For *spectral* clustering however, the clusters are overlapping and it is very difficult to determine the number of clusters since the eigenvalues are likely to have continuous values with no well-defined gap. Therefore, the number of clusters is set to be one. Such a situation shows that perhaps more sophisticated methods may benefit from incorporating multiple features into the clustering decision, for instance to encode structural regularity as well as just inter-element proximity.

B. Cluster Proximity Ranking

For the subsequent speech enhancement and recognition evaluation, the *rule-based* clustering was used to obtain microphone clusters for Data Sets B, C and D. While the *spectral* showed slightly better performance in Table II, the clusters for the enhancement and recognition experiments are well separated and both methods gave the same clusters.

Following rule-based clustering, each cluster was ranked for its proximity to the speaker by means of the TDOA-based algorithm described in Section III-D. Table III details the automatically-determined cluster configurations for the data sets used in the subsequent enhancement and recognition experiments.

C. Speech Enhancement Evaluation

The conventional method to measure the noise reduction is to compute the amount of speech energy over the noise energy as signal-to-noise ratio (SNR) after the enhancement. In practice, calculating the

TABLE III

Microphone membership and the average proximity score $\bar{D}_k$ (in ms) to clusters for the ad-hoc array setup in Data Sets B, C and D.

Data Set	Spk. Loc.	Closest Clust. Mic.		2^{nd} Closest Clust. Mic.		3^{rd} Closest Clust. Mic.	
		Mics.	$\bar{D}_k$	Mics.	$\bar{D}_k$	Mics.	$\bar{D}_k$
B, D	S1	5,6,7,8	0	1,2,3,4	5.6	-	-
	S2	5,6,7,8	0	1,2,3,4	4.4	-	-
	S3	1,2,3,4	0	5,6,7,8	3.8	-	-
C	S1	4,5,6	0	7,8	3.8	1,2,3	5.7
	S2	4,5,6	0	7,8	0.5	1,2,3	3.9
	S3	1,2,3	0	4,5,6	0.2	7,8	2.3

true SNR is rarely possible as the speech and noise cannot be separated at the output. Therefore in this article, the average segmental signal-plus-noise to noise ratio is computed [28].

In addition to calculating segmental SNR to evaluate the noise suppression capability, the overall quality of enhanced speech signal is also evaluated using the Perceptual Evaluation of Speech Quality (PESQ) measure. It has been shown that the PESQ measure is well correlated with the subjective listening test compared to the segmental SNR since it considers distortions and artifacts introduced in the processing of speech signals by the speech enhancement algorithms [29], [30]. For experiments in this paper, the PESQ software[1] was used to predict the Mean Opinion Score (MOS) of the enhanced speech.

The output segmental SNRs and PESQ measures of the blind beamforming for the closest cluster, the second closest cluster, and using all microphones are presented in Table IV and Table V for Data Set B respectively. Because the different blind beamforming methods exhibited very similar performance in terms of these enhancement quality measures, only the Blind MVDR results are shown to simplify presentation and analysis. For comparison, the closest single input channel measure is presented in the same tables.

1) Experiments on Data Set B: Experiments on Data Set B provide an example situation where the speaker is relatively close to one of the clusters and far from the other. In this situation, considering the results in in Table IV and V, using the closest cluster for beamforming gives the best performance in

[1][Online]. Available: http://www.utdallas.edu/∼loizou/speech/software.htm

TABLE IV

Segmental SNR (dB) of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. For the localised noise condition, the result shows the average over positions N1, N2 and N3 in the database. The best result for each scenario is highlighted.

Spk. Position	Noise Condition	Single Channel	Blind MVDR		
			Closest cluster	Second closest cluster	All microphones
S1	Localised	9.07	**10.65**	6.38	9.45
	No noise	12.08	**13.42**	9.09	12.76
S2	Localised	9.64	**11.68**	5.11	6.90
	No noise	12.61	**13.25**	7.46	9.65
S3	Localised	8.72	**9.17**	6.98	9.00
	No noise	10.50	**11.09**	8.62	10.73

TABLE V

PESQ measures of ad-hoc array cluster beamforming in Data Set B. Results are averaged over 30 utterances. For the localised noise condition, the result shows the average over positions N1, N2 and N3 in the database. The best result for each scenario is highlighted.

Spk. Position	Noise Condition	Single Channel	Blind MVDR		
			Closest cluster	Second closest cluster	All microphones
S1	Localised	1.56	**1.78**	1.27	1.58
	No noise	1.72	**1.99**	1.35	1.64
S2	Localised	1.52	**1.75**	1.08	1.60
	No noise	1.58	**1.91**	1.22	1.74
S3	Localised	1.35	1.60	1.31	**1.65**
	No noise	1.42	1.71	1.44	**1.77**

terms of SNR and PESQ compared to the closest channel or the full array. For the speaker at position S3 however, the speaker is roughly the same distance from both clusters. This is reflected in the lower proximity score D_k in Table III for the second closest cluster in the S3 case. In this case, the SNR and PESQ measures show similar performance between the closest cluster and the full array, with the full

TABLE VI

Segmental SNR (dB) of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. For the localised noise condition, the result shows the average over positions N1, N2 and N3 in the database. The best result for each scenario is highlighted.

Spk. Position	Noise Condition	Single Channel	Blind MVDR		
			Closest cluster	Weighted cluster combination	All microphones
S1	Localised	9.96	10.87	**11.27**	7.77
	No noise	12.90	13.67	**14.00**	10.90
S2	Localised	8.89	9.58	**9.99**	8.84
	No noise	11.48	11.64	**11.97**	11.32
S3	Localised	7.26	6.82	8.11	**9.04**
	No noise	10.26	10.44	10.43	**12.73**

TABLE VII

PESQ measures of ad-hoc array cluster beamforming in data set C for closest cluster and second closest cluster. Results are averaged over 30 utterances. For the localised noise condition, the result shows the average over positions N1, N2 and N3 in the database. The best result for each scenario is highlighted.

Spk. Position	Noise Condition	Single Channel	Blind MVDR		
			Closest cluster	Weighted cluster combination	All microphones
S1	Localised	1.69	**1.92**	1.91	1.63
	No noise	1.86	**2.16**	2.14	1.85
S2	Localised	1.32	1.43	**1.62**	1.51
	No noise	1.35	1.52	**1.72**	1.72
S3	Localised	1.21	1.32	**1.48**	1.45
	No noise	1.39	1.48	**1.68**	1.53

array offering higher PESQ.

For the second closest cluster, the beamforming techniques do not show improvement over the single channel. This is likely attributed to the minor steering errors which effectively result in signal degradation, as well as the lower input SNR. Depending on the location of the speaker, the beamformer output from the

full array may not give improvement over the output of closest cluster with less number of microphones. While using more microphones in theory increases the overall array gain, in this practical instance the degradation is attributable to the large inter-microphone spacings leading to erroneous blind TDOA estimation. To phase-align the microphone signals, the TDOA is computed by selecting a reference microphone and calculating delays relative to this reference. Unfortunately the calculation of delays in this way causes inaccuracies between more distant pairs.

The cross-correlation function between two microphones which are spatially close will be dominated by a peak corresponding to the TDOA difference, as they receive signals which have otherwise undergone very similar acoustic transfer functions from the source. For microphones with large distances however, the two impulse responses are likely to be different, increasing the probability of reflections obscuring the cross-correlation peak. To illustrate this phenomenon, the TDOA values computed from the speech frames between microphones in the closest cluster, second closest cluster, and using all microphones are presented in the form of box plot in Figure 6.

Figure 6 illustrates that delays calculated from microphones in the closest cluster are reasonably accurate, in which the mean, median, and expected ground truth delays have consistently similar values. Similar delay accuracy is observed in the second closest cluster but with a larger standard deviation. Delays calculated between distant microphone pairs, however, shows large inconsitencies (e.g. TDOA_{51}, TDOA_{61}, TDOA_{71}, and TDOA_{81}) due to the large distance between the closest microphone as the reference microphone and the 5^{th}, 6^{th}, 7^{th}, 8^{th} closest microphone to the speaker respectively.

2) Experiments on Data Set C: In the second set of experiments conducted on Data Set C, results in Table VI and VII show that in general the weighted cluster combination of the two closest clusters improves speech quality compared to either the closest cluster or the full array. For a speaker at position S2 and S3, the extremely low proximity score of the second closest cluster of 0.5 and 0.2 in Table III suggests that the distance from the speaker to the two closest clusters is approximately the same. In this situation, it is beneficial to combine multiple clusters rather than simply select the closest one. This shows that the proximity score D_k may be used to indicate whether it is worth combining clusters based on their relative distance to the speaker.

D. Speech Recognition Evaluation

While improvement in SNR gives a good indication of performance of the enhanced signal in terms of noise reduction, it is appropriate to confirm the proposed clustered beamforming approach when it is used as a front-end for automatic speech recognition. All speech recognition results quoted in this paper

are the percentage of Word Error Rate (WER) [31]. Comparing WER performance between methods in these ASR tests provides a more meaningful indication of relative speech quality than the simple SNR.

Experiments were conducted on Data Set D using the WebASR web service[2], which was configured to run a system based on the AMI MDM system [3], [32]. The WebASR system is a state-of-the-art large vocabulary system, optimised for conversational speech such as in meetings, from international English speakers, and trained using table-top microphones. As a benchmark comparison, speech recognition results generated from the original clean database files gave a WER of 26.1%. As it is not optimised specifically for read speech from the WSJ corpus, this baseline performance is lower than might be achieved with a task-specific ASR system, however the WebASR system was chosen as it provides robust results when using speech re-recorded on distant microphones. WebASR also makes it easy for other researchers to benchmark array processing methods using the same ASR system used in this paper. The speech recognition results are presented in Table VIII for single channel input and the outputs from the blind Delay-Sum, MVDR, and Superdirective techniques.

The speech recognition performance from the closest cluster beamforming output shows significant improvement over a single channel for all investigated beamforming techniques. Delay-sum, MVDR, and Superdirective show similar improvement albeit slightly higher for the last two techniques. Position S3 shows the smallest improvement of 7.1% compared to the single channel input, with WER of 53.4% compared to position S1 which improves 7.5% from 43.2%, and position S2 which improves 8.9% from 45.7% for delay-sum. For this position, the lower single channel WER and its beamforming output in the closest cluster is attributed to the lower input SNR and PESQ measures (Table IV and V) as the speaker is not facing the microphones directly.

In the second closest cluster, while delay-sum shows improvement compared to the single channel, this is not the case for MVDR and superdirective beamforming. This degradation of performance is likely attributed to the steering error due to inaccuracy in direct delay estimations, as discussed above. This is more evident in the MVDR and superdirective as they are more sensitive to steering errors than delay-sum [13]. For this same reason, together with the fact that some of the more distant microphones have lower speech quality, beamforming using the full array also degrades the performance.

As explained in Section II-A, to reduce the sensitivity of MVDR and superdirective to this type of steering error, a small scalar is usually added to the diagonal of the coherence matrix. The scalar μ depends on the choice of constraint T_o. Decreasing the sensitivity will result in a large value of μ,

[2][Online]. Available at http://www.webasr.org/

TABLE VIII

% word error rate from various blind beamforming methods on Data Set D. For comparison, WER of 26.1% is achieved using the original clean corpus recordings on the same ASR system.

Speaker Position	Technique	Closest clust.	2^{nd} closest clust.	All microphones
S1	Single Ch.	43.2	64.0	-
	Delay-sum	35.7	56.5	44.4
	MVDR	36.0	59.8	48.1
	Superdirective	34.8	61.4	55.5
S2	Single Ch.	45.7	69.8	-
	Delay-sum	36.8	67.8	49.6
	MVDR	36.2	76.4	54.8
	Superdirective	35.5	83.3	64.3
S3	Single Ch.	53.4	62.6	-
	Delay-sum	46.3	56.5	49.9
	MVDR	45.8	64.2	54.4
	Superdirective	46.0	67.4	61.2

biasing the effective noise matrix towards the identity matrix as in delay-sum: as μ approaches infinity, the performance of MVDR and superdirective is therefore expected to be equal to the delay-sum. While MVDR and superdirective prove to be beneficial when using a cluster of spatially close microphones due to their optimised noise coherence suppression, in the case of blind estimation of delays, the delay-sum shows to be a more robust beamforming technique.

E. Experiments on Weighted Cluster Combination

To investigate whether performance may be improved by combining cluster beamformer outputs, a first set of experiments was conducted on Data Set B to study the effect of cluster weight. The Weighted Cluster Combination (WCC) beamforming method described in Section III-E was used to combine the two clusters with varying cluster weights. Note that only the blind Delay-Sum beamforming method was used for these experiments (hence results vary from those reported using MVDR on the same data in Table IV). As there are only two clusters, the weighting parameters α_k from Equation 29 were set to α_1 and $\alpha_2 = 1 - \alpha_1$, with α_1 varied from 0 to 1.

The WCC beamforming is evaluated in terms of the SNR improvement, determined by subtracting the input SNR of the closest microphone from the SNR of the WCC output. Figure 7 shows the SNR improvement of the weighted cluster combination for Data Set B in different speaker positions for the varying weights. The SNR improvement obtained from the closest cluster and using all microphones are also plotted for comparison.

These preliminary results suggest two effects. First, there seems little benefit to combining clusters when one of the clusters is considerably closer to the speaker and further from other noises than the others (i.e. for S1 and S2. See Figure 4 and the second cluster proximity measure in Table III). Second, when it is worth combining multiple clusters due to their similar distance to the speaker (i.e. for S3), the weighting of each depends on their relative size (number of microphones), as suggested by the theoretical motivations in Section III-A. The plot of SNR improvement versus cluster weight in Figure 7(c) shows that even though the two clusters independently offer different SNRs due to their positioning with respect to speech and noise sources, it may still optimal to equally weight the two clusters, as they have the same size and are at a similar distance to the speaker.

Following these observations, a second set of experiments was conducted on Data Set C. For this dataset, the clusters are less separated and so there is generally more motivation for their combination. Tables VI and VII give the SNR and PESQ results obtained from a weighted combination of the first two clusters, where the weight was set to reflect their relative sizes: that is, weights of (0.6,0.4), (0.6,0.4) and (0.5, 0.5) were used for speaker positions S1, S2 and S3, respectively. This choice of cluster weights based on their relative sizes is equivalent to performing delay and sum on all microphones in the combined clusters, but with the delays computed in a more robust clustered fashion, rather than with respect to a single global reference microphone.

Experiments on the weighted cluster combination beamforming show that, depending on the relative positions of speakers, microphones and noises, optimal performance may be obtained when the contributions of cluster are taken into account. To give more insight into cluster combination, Figure 7(a) presents a plot of SNR with the peak result occuring when α_1 and α_2 are set to be 0.8 and 0.2 respectively. Similarly in Figure 7(c), the optimal performance for speaker position S3 is obtained when α_1 and α_2 are set to be equal 0.5. For speaker at position S2, however, using the closest cluster by itself yields the optimal performance. The best performance of weighted cluster combination for speakers at position S1 and S3 are only slightly higher than their closest cluster beamforming performance. This is likely attributed to the large distance that separates both clusters, causing small contributions of the second closest cluster to the overall performance. For position S1 in particular, only a small contribution is expected due to the fact

that the second closest cluster is located close to the noise source. Using all microphones directly without any clustering results in negative SNR improvement - this is somewhat counter-intuitive, as it may be expected that this would correspond to the performance of the equally weighted cluster combination (e.g. $\alpha_1 = \alpha_2 = 0.5$). The difference is again due to the inaccuracy in TDOA estimation for more distant microphone pairs, as explained above - in the case of weighted cluster combination, only the single inter-cluster delay is calculated over a longer distance, while for the unclustered approach, many pairs will involve relatively large distances.

To confirm the effect of cluster combination on speech recognition performance, the delay-sum WCC beamformer for S3 was tested on Data Set D, gave 42.3% WER, compared with 46.3% using the closest cluster and 49.9% using all microphones.

These results, as well as theoretical motivations, indicate that optimal cluster weights will depend on cluster size and relative proximity. The effect of combining clusters with weights according to cluster size on Data Set C was discussed in Section V-C above. Note that, in more general circumstances, the cluster size and relative proximity may not be the only the two factors which influence the cluster weights for optimal performance. Factors which affect speech quality such as noise and reverberation may also play a part and more experiments are needed to verify how this will affect the choice of cluster weights.

VI. Conclusion

In this article, a novel clustered approach to blind beamforming from ad-hoc microphone arrays has been proposed. For the first step, two microphone clustering algorithms were proposed to group microphones using only the knowledge of noise coherence. In a second step, the clusters are ranked based on their proximity to the speaker using TDOA information. Finally, two methods for achieving microphone array beamforming using these clusters were investigated in the context of speech enhancement and recognition: closest cluster (CC) beamforming, and weighted cluster combination (WCC) beamforming. Experiments were conducted on a new database recorded for this purpose in a typical meeting room environment. Eight microphones were used in a variety of placements to simulate ad-hoc arrangements on a meeting table. Experiments validated the partitioning provided by the clustering algorithms, as well as the effect of subsequent beamforming on the SNR, PESQ, and the word error rate.

Depending on the relative distance from the cluster of microphones to the speaker as indicated by their proximity score, using a cluster or multiple clusters in the same time can provide better performance than a larger array. An underlying cause of this improvement is the fact that larger inter-microphone distances can lead to erroneous delay estimation for blind steering vector formulations. The speech recognition

experiments further confirm the benefit of clustering, as both clustered beamforming methods (CC and WCC) show significant improvement over both the single channel input and the full set of microphones.

Amongst the blind beamforming techniques investigated, it was found that delay-sum beamforming is the most robust when using the full set of microphones, as it is less sensitive to steering errors than MVDR and superdirective beamforming. When only the closest cluster is used, however, MVDR and superdirective improve more relative to delay-sum, as these methods are optimal for reducing the ambient noise when more accurate steering vectors are used. Overall, however, all beamforming methods offer similar speech recognition performance from the clustered beamforming, and therefore delay-sum is a good candidate in practical deployments due to its simplicity and decreased sensitivity to delay estimation errors. While not investigated in this paper, it is however noted that MVDR and superdirective beamforming provide a potential means of trading off robustness to steering errors with noise reduction by varying the white noise constraint.

Experiments on the weighted cluster combination beamforming indicate that combining microphones at the cluster level, rather than individual level, offers improved robustness for blind beamforming when the clusters are similarly proximate to the speaker. The proposed cluster proximity measure offers a promising means of detecting when this cluster combination should be used.

Based on the above findings, ongoing research into beamforming from ad-hoc microphone arrays will investigate other measures for ranking clusters, and automatically determining cluster weights for combination. In true ad-hoc situations, characteristics other than just distance to the speaker should be considered, as microphones may be of widely varying quality and response. Finally, it is noted that while this research has been constrained to propose solutions for unknown geometries, in true ad-hoc scenarios robust methods must also consider potential differences in microphone response and synchronisation.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 2–24, 1988.

[2] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, “The SRI-ICSI spring 2007 meeting and lecture recognition system,” *Lecture Notes in Computer Science*, vol. 4625, pp. 450–463, 2008.

[3] T. Hain et al., “The AMI system for the transcription of speech in meetings,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, 2007, pp. 357–360.

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

[5] J. Fiscus, J. Ajot, M. Michet, and J. Garofolo, “The rich transcription 2006 spring meeting recognition evaluation,” *Lecture Notes in Computer Science*, vol. 4299, pp. 309–322, 2006.

[6] A. Janin et al., "The ICSI-SRI Spring 2006 Meeting Recognition System," in *Proc. of the Rich Transcription 2006 Spring Meeting Recognition Evaluation, Washington, USA*, 2006.

[7] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown location–part i: Far-field sources," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 286–299, Mar 1987.

[8] ——, "Array shape calibration using sources in unknown locations–part ii: Near-field sources and estimator implementation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 35, no. 6, pp. 724–735, Jun 1987.

[9] I. Himawan, S. Sridharan, and I. McCowan, "Dealing with uncertainty in microphone placement in a microphone array speech recognition system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008, pp. 1565–1568.

[10] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, April 2007, pp. 761–764.

[11] J. Dmochowski, Z. Liu, and P. Chou, "Blind source separation in a distributed microphone meeting environment for improved teleconferencing," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 31 2008-April 4 2008, pp. 89–92.

[12] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[13] H. L. V. Trees, *Optimum Array Processing - Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley, 2002.

[14] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[15] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array calibration in diffuse noise fields," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, pp. 666–670, 2008.

[16] I. Himawan, M. I., and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin, 2008, pp. 295–305.

[17] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[18] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan 1982.

[19] R. K. Cook et al., "Measurement of correlation coefficients in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 27, pp. 1072–1077, 1955.

[20] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On Spectral Clustering: Analysis and an algorithm*. MIT Press, 2001, pp. 849–856.

[21] L. Zelnik-manor and P. Perona, *Self-tuning spectral clustering*. MIT Press, 2004, pp. 1601–1608.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[23] U. von Luxburg, "A tutorial on spectral clustering," Max Planck Institute for Biological Cybernetics, Tech. Rep., 2006.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1995, pp. 81–84.

[25] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," DRA Speech Research Unit, Tech. Rep., 1992.

[26] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, 12–15 May 1998, pp. 757–760.

[27] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," in *In Proceedings of ICSLP*, 2002, pp. 573–576.

[28] S. R. Quackenbush, T. P. B. III, and M. A. Clements, *Objective measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality PESQ-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2001, pp. 749–752.

[30] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 –238, January 2008.

[31] S. Young et al., *The HTK Book, 3rd ed.* Cambridge University Engineering Department, December 2006.

[32] T. Hain, A. El Hannani, S. Wrigley, and V. Wan, "Automatic speech recognition for scientific purposes - WebASR," in *Proc. of Interspeech*, 2008.

Ivan Himawan (S'08) received the B.E. degree in electrical and computer engineering from the Queensland University of Technology (QUT), Brisbane, in 2004. He completed his PhD with the Research Concentration in Speech, Audio, Image, and Video Technology at QUT in 2010, including a period of research at CSTR in University of Edinburgh.

He is currently with the QUT mobile multimedia research group as a research fellow involved in multimedia analysis, indexing, and delivery research. His current research interests include microphone array processing, speech enhancement and recognition, and content-based retrieval of multimedia.

Iain McCowan (M'97) received the B.E. and B.InfoTech. from the Queensland University of Technology (QUT), Brisbane, in 1996. In 2001, he completed his PhD with the Research Concentration in Speech, Audio and Video Technology at QUT, including a period of research at France Telecom R&D.

In 2001 he joined the IDIAP Research Institute, Switzerland, progressing to the post of Senior Researcher in 2003. While at IDIAP, he worked on a number of applied research projects in the areas of automatic speech recognition, content-based multimedia retrieval, multimodal event recognition and modeling of human interactions. From 2005-2008 he was with the CSIRO eHealth Research Centre, Brisbane as Project Leader in the area of multimedia content analysis. In 2008 he founded Dev-Audio Pty Ltd to commercialise microphone array technology. He holds an adjunct appointment as Associate Professor at QUT in Brisbane.

Sridha Sridharan has a BSc (Electrical Engineering) degree and obtained a MSc (Communication Engineering) degree from the University of Manchester Institute of Science and Technology (UMIST), UK and a PhD degree in the area of Signal Processing from University of New South Wales, Australia. He is a Senior Member of the Institute of Electrical and Electronic Engineers - IEEE (USA).

He is currently with the Queensland University of Technology (QUT) where he is a full Professor in the School of Engineering Systems. Professor Sridharan is the Deputy Director of the Information Security Institute and the Leader of the Research Program in Speech, Audio, Image and Video Technologies at QUT. He has published over 300 papers consisting of publications in journals and in refereed international conferences in the areas of Image and Speech technologies during the period 1990-2010. During this period he has also graduated 20 PhD students as their Principal Supervisor in the areas of Image and Speech technologies. Prof Sridharan has also received a number of research grants from various funding bodies including Commonwealth competitive funding schemes such the Australian Research Council (ARC) and the National Security Science and Technology (NSST) unit. Several of his research outcomes have been commercialised.

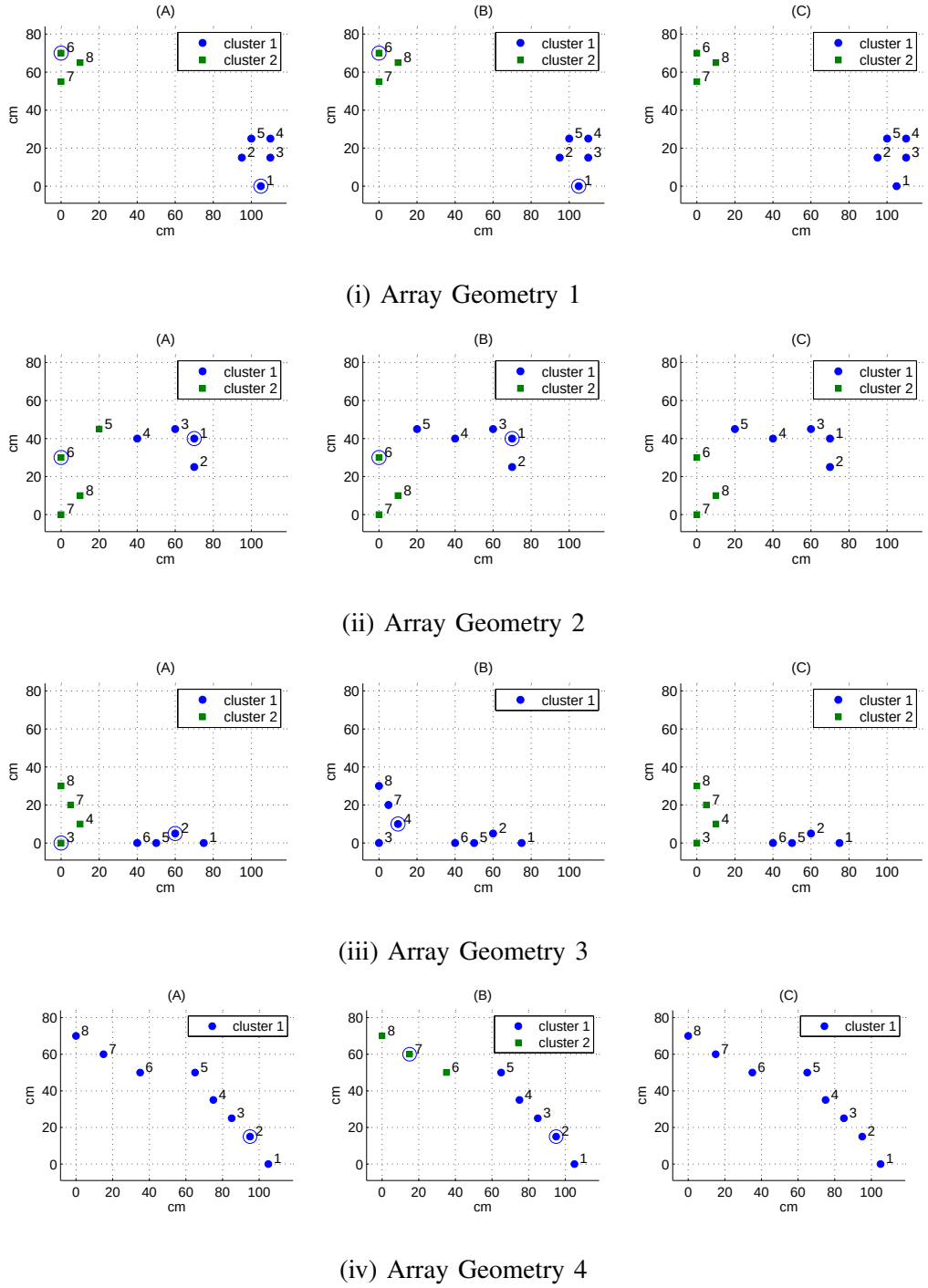
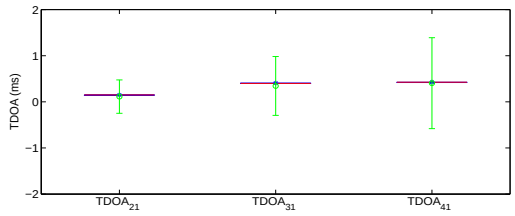
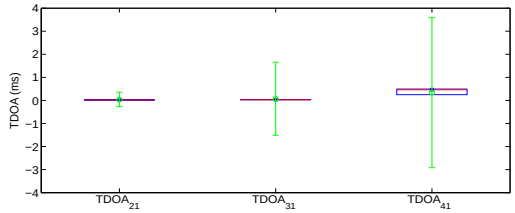


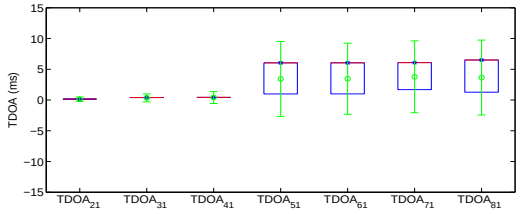
Fig. 5. *Cluster assignments on the ground truth positions for four geometries, with one geometry per row. In each case, (A) shows the sub-arrays obtained from known microphone positions while (B) and (C) shows the result of* rule-based clustering *and* spectral clustering *from the measured noise coherence. The centre microphone in a cluster is shown by encircling line for* rule-based clustering.



(a) closest cluster delays

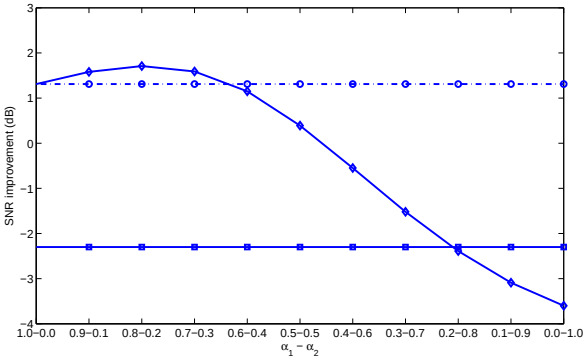


(b) second closest cluster delays

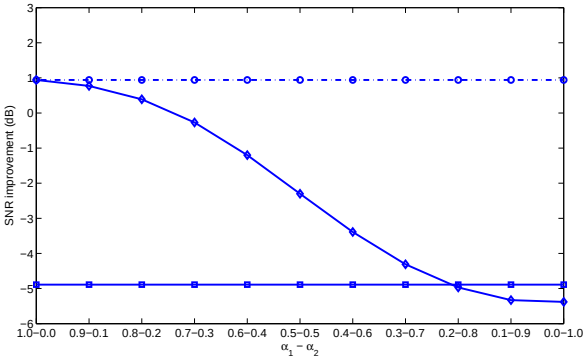


(c) all microphone delays

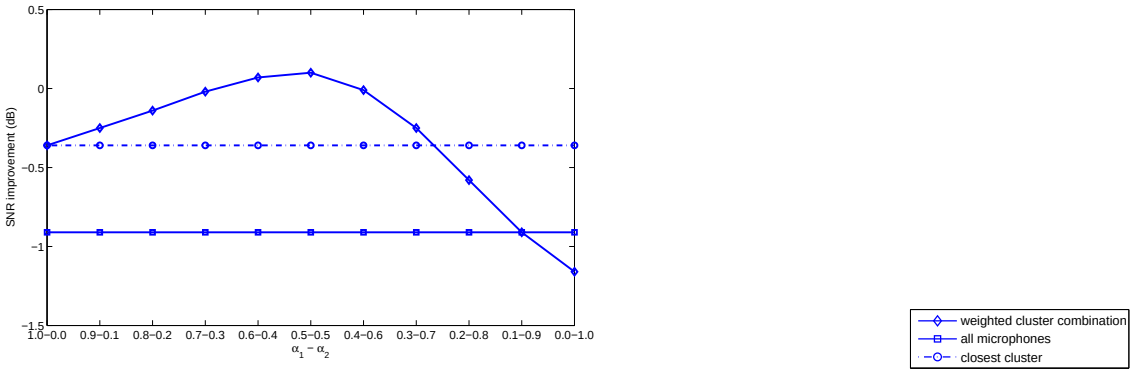
Fig. 6. Box plot of the TDOA values applied to closest cluster, second closest cluster, and all microphones for speaker at position S1 with noise condition N1 from Data Set B. The attributes of plots are similar to box plot in Figure 1.



(a) Speaker Position S1



(b) Speaker Position S2



(c) Speaker Position S3

Fig. 7. The delay-sum SNR improvement of weighted cluster combination for noise condition N1 in different speaker positions by varying weight α_1 for the closest cluster and weight α_2 for the second closest cluster.