`

FORMING AD-HOC MICROPHONE ARRAYS THROUGH CLUSTERING OF ACOUSTIC ROOM IMPULSE RESPONSES

S.Pasha[1], Y. X. Zou[2], C. Ritz[1],

[1]School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia
[2]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, China

ABSTRACT

This paper investigates the formation of ad-hoc microphone arrays for the purpose of recording multiple sound sources by clustering microphones spatially distributed within a room. A novel codebook-based unsupervised method for cluster formation using features derived from the Room Impulse Responses (RIRs) corresponding to each microphone is proposed and compared with baseline clustering and classification methods. The features correspond to the sequence of arrival time and time delays of echoes as estimated by peaks of the RIRs along with peak amplitudes. Results suggest that the proposed codebook based clustering algorithm can outperform KNN supervised classification method and kmeans unsupervised clustering method applied to microphone segmentation and clustering, in terms of clustering success rate and noise robustness.

Index terms- Ad-hoc microphone arrays, Codebook based clustering, Informed clustered beamforming, Room impulse response, Time delays of arrivals.

1. INTRODUCTION

An ad-hoc microphone array is formed from sets of microphones randomly positioned in a room and can be used to record multiple spatially distributed sound sources with a better and more flexible spatial coverage compared with a single microphone array located at one position. Such arrays could be formed from microphones attached to mobile phones or other portable computing devices such as tablets. In such scenarios where no information about the sources are available, tasks such as Direction of Arrival (DOA) estimation of individual sources must be done blindly and rely on automatic position calibration of the ad hoc microphone arrays.

Some recent recording methods [1,2] using ad-hoc microphone arrays utilize partial information to help guide applications such as sound source separation and classification. These informed signal processing approaches [3] are more effective compared to blind approaches for the analysis of complex acoustic scenes and sound source separation. As an example, in [1] a novel method for exploiting relative microphone and source spatial locations was introduced and evaluated for microphone clustering and signal classification. This method relies on accurate knowledge of the total number of sources as well as the total number of clusters to form.

In [2] the authors showed that rather than using all microphones in a room, forming ad-hoc microphone arrays using small clusters of microphones each located close to one source can yield better separation quality. The approach removes microphones from the ad-hoc array that are located far from target sources, which may be corrupted by other sources and hence have a low target to interference signal ratio. Such an approach also reduces the beamforming steering error [2], [4] and is based on measuring the coherence between microphones in noise-only periods as well as the relative Time Difference of Arrival (TDOA) between neighboring microphones during speech periods. Their approach assumed small subsets of microphones were located close to desired speakers.

The TDOA is based on the difference in arrival time of the direct sound at each adjacent microphone and is generally calculated using cross correlation-based methods. These methods suffer from room reverberation and hence techniques to suppress the effects of reverberation on TDOA estimation accuracy are often required. In contrast, recent research [5] has instead made use of the reverberation to localize virtual image sources and reflectors based on counting the echo peaks detected in the acoustic Room Impulse Response (RIR) recorded by sets of microphones.

More recently [6], this approach has been used to localize sets of microphones using a single sound source as opposed to other methods that generally require more sources (typically 5) and microphones (typically 10) [7]. Motivated by this work, the approach proposed in this paper utilizes information inferred from echoes as discriminative features to cluster microphones. In particular, arrival and delay times and peak amplitudes are extracted from the RIRs and used as features within existing clustering algorithms to form ad-hoc microphone arrays.

Section 2 of this paper describes the general scenario of recording with ad-hoc microphone arrays and describes the model adopted for this problem. In section 3 the proposed methodology based on clustering RIR features is described. In section 4 the implementation and evaluation process of the baseline algorithms and the proposed model are depicted. In section 5 the paper is concluded.

1.1. Relation to Prior Work

Compared to existing ad-hoc microphone array approaches [1,2] the proposed approach is based on clustering of features derived from RIR recordings rather than recorded speech signals and does not require complex calculations of noise coherence and inter-microphone cross correlations. Similar to [2], information about the sources, room and microphone array is unavailable. In contrast to [1], in this research reverberation is not suppressed (e.g. using cepstral mean normalization) but instead is exploited to cluster the microphones. This is motivated by the approaches in [5,6], where similarly we estimate echoes as the peaks in the RIR recordings. These are then used within alternative clustering algorithms for forming the ad-hoc microphone arrays. The limitation of the proposed method in [1] is that it assumes microphones are located close to the sources and there is no microphone at an equal distance from two sources. This limitation is covered in this paper by exploiting features derived from RIRs which discriminate microphones effectively. Discrimination of symmetric clusters by using two asynchronous sources located at two different locations is the novelty of this proposed method.

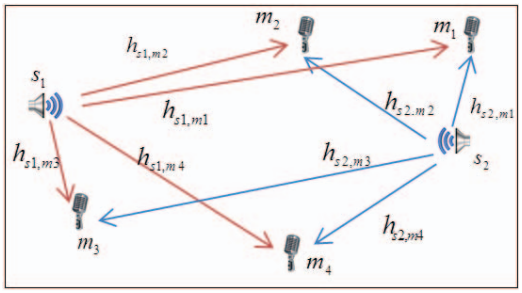

Fig. 1. RIRs and time differences of arrivals in an ad-hoc array

2. AD-HOC MICROPHONE ARRAY MODEL

In the general form of the problem let M microphones be distributed in a room of unknown geometry and labelled $m_1, m_2, \dots m_j \dots, m_M$, which record N sources $s_1, s_2, \dots s_k \dots, s_N$ (Figure1 illustrates this for an example of 4 microphones recording 2 sources). The sound recorded by each of these microphones is the convolution of the acoustic RIR corresponding to it's location in the room and the source signal [8]. It is assumed that all microphones are synchronized and the lengths of the RIRs are equal. These RIR sequences contain impulses received from direct paths between sources and microphones and reflections from the walls, ceiling and floor and can be modelled mathematically as a train of impulses as [9]:

$$\hat{R}_{sk,mj}(n) = \sum_l a_{sk,mj}(l)\delta\left(n - d_{sk,mj}(l)\right) + N(n) \quad (1)$$

where $d_{sk,mj}(l)$ represents the propagation delay from source and reflectors to the microphone m_j when source k is active, $a_{sk,mj}(l)$ represents the amplitudes of each impulse corresponding to an echo and l=0 to L represents the number of impulses. $N(n)$ represents the noise in the general form. In practice, RIRs can be estimated by techniques such as recording a sine-sweep covering a range of frequencies (e.g. 20Hz to 20 KHz) and digitally sampling this signal as a pre-recording phase or they can be extracted from speech signals by the proposed method in [10, 11]. However, in this paper we assume perfect knowledge of the RIRs.

Let $\hat{h}_{sk,mj}$ represent the first $L+1$ impulses of (1), which correspond to the direct path component and first L echoes of source k and can be represented as:

$$\hat{h}_{sk,mj} = \left[h_{sk,mj}(0), \dots, h_{sk,mj}(L)\right] + N(n) \quad (2)$$

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

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

The peak sample numbers representing the propagation delays, $d_{sk,mj}(l)$, corresponding to the peaks of $\hat{h}_{sk,mj}$ of (2) are represented here by the vector of delays, $\hat{d}_{(sk,mj)}(l) = \left[d_{sk,mj}(0), d_{sk,mj}(1), \dots, d_{sk,mj}(L)\right]$, where $d_{sk,mj}(0)$ is the arrival time from the source k to the microphone m_j for the direct path signal and $d_{sk,mj}(1), \dots, d_{sk,mj}(L)$ represent

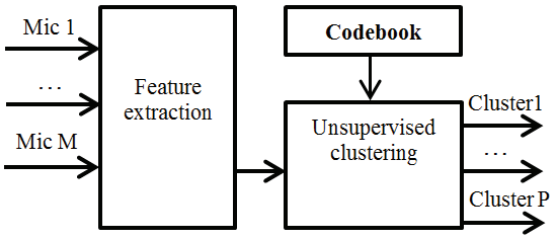

Fig. 2. Proposed system overview

the delays for the first L echoes. The delay matrix for microphone m_j can be constructed as D_j, where j=1 to M:

$$D_j = \begin{bmatrix} d_{s1,mj}(0) & \dots & d_{sN,mj}(0) \\ \vdots & \ddots & \vdots \\ d_{s1,mj}(L) & \dots & d_{sN,mj}(L) \end{bmatrix} \quad (4)$$

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

$$A_j = \begin{bmatrix} \left|\hat{h}_{s1,mj}(0)\right| & \dots & \left|\hat{h}_{sN,mj}(0)\right| \\ \vdots & \ddots & \vdots \\ \left|h_{s1,mj}(L)\right| & \dots & \left|\hat{h}_{sN,mj}(L)\right| \end{bmatrix} \quad (5)$$

$\left|\hat{h}_{sk,mj}(0)\right| = a_{sk,mj}(0)$ from (1) and the extension to more sources and microphones than in Figure1 is straightforward. Feature vectors derived from these matrices may be more efficient to use within machine learning algorithms as opposed to state of the art methods which use cepstral features such as MFCC and LP-CMRARE [1]. In section 3.1 and 4.2 the effectiveness and accuracy of these proposed discriminative features for microphone clustering applications are investigated.

3. METHODOLOGY

In a randomly distributed microphone array, the objective is to compare microphones RIRs and cluster microphones into a flexible number of clusters (Figure 2). Translating received signals to small and light discriminative feature sets (Feature extraction) is the first step (Section 3.1) of this process. These discriminative features are then used for clustering of microphones into ad-hoc arrays (Section 3.2).

3.1. Feature Extraction and RIR Simulation

For both supervised and unsupervised methods, using original raw RIRs received by each microphone leads to time consuming and inefficient training and evaluation procedures. Hence, similar to the applied features in [12] the time delay of peaks and their corresponding amplitudes are extracted from RIRs, representing the delays described in (4) above. For simulating RIRs, reflections are modelled as virtual image sources, for example the magnitude of the direct path signal (first peak of the RIR) and arrival time and also the magnitudes of the first three echoes and corresponding echo times are depicted in Figure 3.

In this paper, the implementation of [13] is used for simulating RIRs. The first four maximum peak values of RIRs (5) and corresponding sample indices (4) are detected by a peak-picking process to extract discriminative features from RIRs. It is assumed

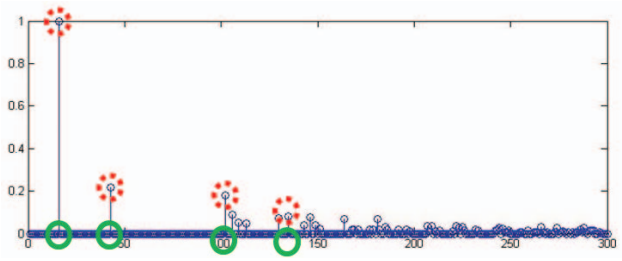

Fig.3. Arrival and echo times and corresponding magnitudes

only one source is active at each RIR recording process.

Input: RIR of each microphone, Codebook
Output: Clustered microphones based on spatial locations

1. ***Choose P center points in the room, obtain arrival time and echo delays and assign a zone label to each center point***
   (Codebook generation)
2. ***For each randomly distributed microphones in the microphone array:***

A. Obtain the recorded RIR
B. Derive discriminative features
C. Compare each microphone's feature vector with the generated codebook
D. Assign the closest center point's zone to the microphone

3. ***The number of assigned zones labels show the number of clusters***

Fig.4. Codebook-based clustering

3.2. Microphone Clustering

This section describes three approaches to microphone clustering: K Nearest Neighbour (KNN) classification; k-means clustering; and a proposed codebook-based approach. Requirements and limitations of each method are described with respect to this specific application.

3.2.1. Supervised classification

The original RIRs, extracted features of RIRs and parameterized RIRs can be applied as a training set to train a *k* nearest neighbors (KNN) classifier with *K*=1 and a predefined number of classes. This first requires training data and here this is generated using simulated RIRs on a 3D grid with varying distances between points including 0.1m, 0.2m and 0.4m (smaller step sizes yield more accurate classification results). Once trained, the classifier then processes the RIRs for microphones at unseen positions (not in the training set). Based on their detected classes the microphones are localized and grouped together. A standard KNN classifier is used here [14].

3.2.2. Unsupervised microphone clustering by k-means

To avoid the need for training data, microphone clustering can be applied directly to all RIR recordings. The goal is to assign a microphone to a group with small mean intra-group distances compared to all data points. This goal may be achieved by minimizing an objective intra-cluster distance function in an unsupervised manner (without training). The distance function can be represented as:

$$\varepsilon_p(xi, xj) = \left(\sum_{k=1}^{K} |x_{ik} - x_{jk}|^p\right)^{\frac{1}{p}} \quad (6)$$

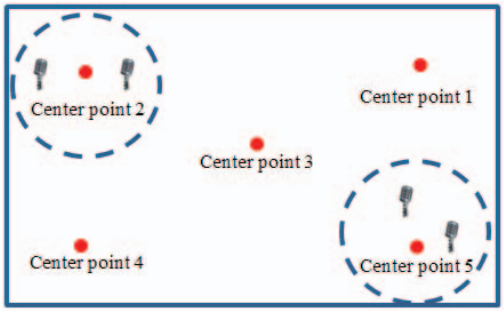

Fig.5. Center points and clusters

where k is the data point dimension, x_i and x_j are parameterized RIRs (i.e. time delays and peak values) of two microphones and p depends on the applied distance function. In this paper, the Euclidian distance (p=2) is used as the dissimilarity function. This is implemented here using the standard k-means algorithm [14]. Choosing the number of clusters is a critical and important task in unsupervised methods. In this research the number of clusters is chosen blindly and based on the total number of microphones. This uncertainty in choosing the number of microphones is the main drawback of unsupervised approaches.

3.2.3. Codebook Based Clustering

In this approach it is assumed that *P* reference RIRs are known (or have been previously recorded) within the room (referred here also as center point of a cluster). These center points can be chosen blindly with a uniform distribution within the room however if there is prior information about possible locations of sources and microphones they can be chosen in an informed manner. Discriminative features as described in section 3.1 are derived from recorded RIRs. For *M* microphones the goal is to assign each data point for each microphone at an unknown position to the closest center point based on features similarities. Similar to Vector Quantization (VQ), microphones are clustered based on the closest matching center points. This is illustrated in Fig. 5 for an example with 4 microphones and 5 center points. Similar to Section 3.2.2, the Euclidean distance is used here for comparing RIRs. Also *P* defines the maximum number of clusters in a flexible manner and depending on the array distribution there might be 1 to *P* clusters.

4. SIMULATION AND RESULTS

This section describes the simulation setups and evaluation results for the proposed clustering methods (section 3.2). Results in section 4.2 are average results for 35 different setups with one or two asynchronously active sources and 4 to 25 microphones. Effects of the room size, noise, discriminative features and the applied clustering methods have been investigated. Limitations and advantages of each method are also discussed. Symmetry and clustering symmetrically positioned microphone clusters together is also addressed by using two asynchronous sources at different positions and concatenating the feature vectors.

L (number of echoes) is an important factor in codebook and discriminative feature vector generation. For all the experiments *L*=3, which means direct path signal along with the first three echoes are exploited as discriminative features. Effects of *L* on clustering performance and feature extraction can be investigated more in future experiments.

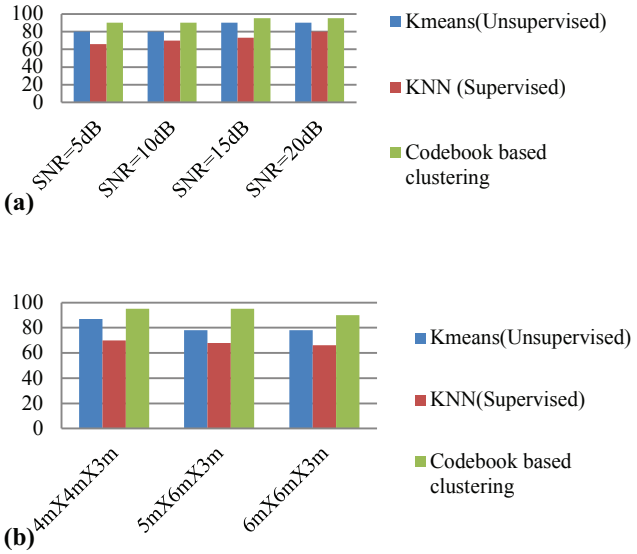

Fig.6. (a) Microphone clustering Success Rate (SR) for a $5m \times 6m \times 3m$ room and 5 center points at different noise levels **(b)** Microphone clustering SR for different simulated rooms at an SNR=10dB.

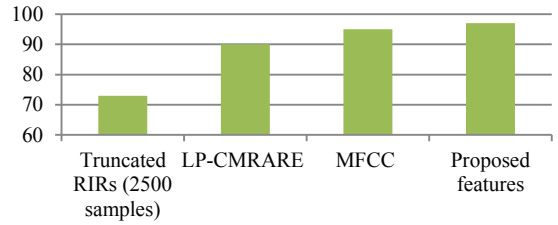

Fig.7. Effect of extracted features (Clean data sets)

4.1. Experimental setup

Approaches discussed in section 3.2, have been examined in different scenarios of Signal to Noise Ratio (SNR) (5dB to 20dB, 5dB steps), room sizes and different microphone distribution models with one active source applied for each RIR recordings. Airport noise from Loziou data base [15] in different SNR's is added to the simulated RIRs with a 44.1 kHz sampling rate and $RT_{60} = 500ms$ for all experiments.

4.2. Results

For M randomly positioned microphones, if microphone m_j is clustered with other spatially close microphones (inter-cluster distances compared with mean intra-cluster distance (6)), the microphone m_j clustering result is labeled *"V"* (Valid) otherwise is labelled *"I"* (Invalid). The success rate, SR, is applied to evaluate all methods and is calculated as:

$$SR = \frac{n(V)}{n(I+V)} \times 100 \quad (7)$$

Figure 6 (a) indicates the average success rate for the three clustering approaches examined for 4 to 25 microphones and one

Applied method	Number of clusters	Training step size (x, y and z axes)	SNR	SR
KNN	4 segments	0.2 m	10 dB	73.5%
K-means	Predefined as 4 (fixed)	No training	10 dB	81%
Codebook based clustering	Flexible (1-5)	No training	10 dB	92%

Table 1 Comparison of each method ($4m \times 4m \times 3m$ room)

Or two asynchronously active sources in different noise levels. Results suggest that mismatch between the clean training set and noisy test set affects the success rate of supervised classification method (i.e. KNN) significantly. Effect of noise can also be observed in the other applied methods where increase in the noise level can distort the peaks and decrease the accuracy of extracted features and clustering results. Figure 6 (b) shows the decrease of success rate in rooms with larger dimensions which can be a result of the direct path signal attenuation. For a symmetric room and only one source located at the center, it is not possible to discriminate two microphones located at two opposite corners with the same distance from the source and reflectors (walls, ceiling and floor) but by using two asynchronous sources and concatenating the feature vectors this issue can be addressed.

Based on these results it can be concluded that the proposed codebook-based method provides the highest success rates for all SNR conditions and room sizes. Results also show that the proposed method is robust against noise. Table 1 provides another comparison of the three methods, showing the SR for SNR of 10 dB and the amount of training data required. Compared to the KNN approach, the proposed codebook-based approach does not require training data while compared to the k-means approach it does not require predetermining the number of clusters and can flexibly detect the number of clusters based on the recorded RIRs. For a $4m \times 4m \times 3m$ room supervised classification process with step size of 0.2m needs 6000 RIRs as training set.

In Figure 7 Proposed extracted features are compared with state of the art approaches to microphone clustering [1], the proposed method in this paper yields more accurate results for a similar scenario and setup of two clusters. The limitation of the proposed method in [1] is that it needs to assume one source is dominant in each cluster, where in the proposed method by using two asynchronous sources at different locations microphones can be discriminated more effectively without any prior information about sources.

5. CONCLUSION AND FUTURE WORKS

This paper described a new approach to clustering microphones to form ad-hoc arrays based on discriminative features derived from the RIRs. These features represent the time delays of echoes and peak amplitudes received by the microphones and provide a compact set of parameters for use within supervised and unsupervised learning algorithms including a proposed codebook-based approach. Investigations and simulations of this research showed that by using a relatively small codebook (5 center points), it is possible to cluster microphones in noisy environments accurately and independently from sources, in terms of number, position and signal variation. Effects of the number of applied echoes (L), center points and RT_{60} time on clustering performance are needed to be investigated more in future research.

`

6. REFERENCES

[1] Gergen, S., Nagathil, A, Martin, R., "Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.3692,3696, 26-31 May 2013

[2] Himawan, I, McCowan, I, Sridharan, S., "Clustering of ad-hoc microphone arrays for robust blind beamforming," *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* , vol., no., pp.2814,2817, 14-19 March 2010

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

[4] Asaei, A, Davies, M.E., Bourlard, H., Cevher, V., "Computational methods for structured sparse component analysis of convolutive speech mixtures," *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* , vol., no., pp.2425,2428, 25-30 March 2012

[5] Dokmanic, I, Lu, Y.M., Vetterli, M., "Can one hear the shape of a room: The 2-D polygonal case," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on , vol., no., pp.321,324, 22-27 May 2011

[6] Dokmanic, I., Daudet, L., Vetterli, Martin" How to Localize Ten Microphones in One Fingersnap" 22nd European Signal Processing Conference, Lisbon, Portugal, September 1-5, 2014

[7] Pollefeys, M., Nister, D., "Direct computation of sound and microphone locations from time-difference-of-arrival data," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* , vol., no., pp.2445,2448, March 31 2008-April 4 2008

[8] Samarasinghe, P.N., Abhayapala, T.D., Polettfi, M.A, Betlehem, T., "On room impulse response between arbitrary points: An efficient parameterization," *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on* , vol., no., pp.153,156, 21-23 May 2014

[9] Dokmanić, I.,Parhizkar R., Walther A., Yue M., Vetterli M. "Acoustic echoes reveal room shape" *Proceedings of the National Academy of Sciences* 110, no. 30: 12186-12191. 2013

[10] Takashima, R., Takiguchi, T., Ariki, Y., "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.4295,4299, 26-31 May 2013

[11] Ajdler, T., Vetterli, M., "The plenacoustic function, sampling and reconstruction," *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on* , vol.5, no., pp.V,616-19 vol.5, 6-10 April 2003

[12] Ribeiro, F., Florencio, D., Ba, D., Cha Zhang, "Geometrically Constrained Room Modeling With Compact Microphone Arrays," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.20, no.5, pp.1449,1460, July 2012

[13] Habets, E., "Room Impulse Response Generator for MATLAB", [online], The Netherlands 2003-2010. http://home.tiscali.nl/ehabets/rir_generator.html

[14] Rogers S., Girolami M., " A First Course in Machine Learning ", Chapman & Hall/Crc, October 2011

[15] Hu, Y. and Loizou, P. (2007). "Subjective evaluation and comparison of speech enhancement algorithms,"*Speech Communication*, 49, 588-601

[16] Jacob, F., Schmalenstroeer, J., Haeb-Umbach, R., "DOA-based microphone array position self-calibration using circular statistics," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.116,120, 26-31 May 2013

[17] Souden, M., Kinoshita, K., Nakatani, T., "An integration of source location cues for speech clustering in distributed microphone arrays," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.111,115, 26-31 May 2013

[18] Joder, C., Weninger, F., Virette, D., Schuller, B., "Integrating noise estimation and factorization-based speech separation: A novel hybrid approach," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.131,135, 26-31 May 2013

[19] Raykar, V.C., Duraiswami, R., "Automatic position calibration of multiple microphones," *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on* , vol.4, no., pp.iv-69,iv-72 vol.4, 17-21 May 2004