Estimating Source Dominated Microphone Clusters in Ad-Hoc Microphone Arrays by Fuzzy Clustering in the Feature Space

Sebastian Gergen, Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
Email: {sebastian.gergen,rainer.martin}@rub.de
Web: www.ruhr-uni-bochum.de/ika

Abstract

Microphones in an ad-hoc microphone array provide distributed signal acquisition in an acoustic environment in which multiple sound sources may be active. For some applications, e.g., in the context of signal classification or noise reduction it is helpful if clusters of microphones can be determined which are dominated by one of the sources, and one additional cluster of microphones which receives contributions from all sources with similar power and a high amount of reverberation. We propose to perform a fuzzy clustering of audio features which are extracted from microphone signals and evaluate the fuzzy membership values to assign microphones either to source dominated clusters or to a background cluster. Simulations and recordings of scenarios with two sound sources and multiple microphones are used to evaluate the proposed assignment procedure.

1 Introduction

Ad-hoc microphone arrays (MA) [1] may consist of devices like mobile phones, tablet computers and laptops, which all provide the ability of audio capturing, signal processing and integrated connectivity. However, traditional audio signal processing methods, e.g., noise reduction [2, 3], blind source separation [4, 5] or classification [6] cannot be easily implemented. For these methods, often, the knowledge about the position of the array elements is important, and is thus also an aspect in research in the field. For example, an accurate estimation of the position of such devices is aimed for by using voice activity detection and coherence models [7], energy decay information [8], calibration signals [9] or available compass information [10].

Instead of an estimation of the precise position of microphones in the room, we aim at building sub-arrays, by clustering the microphones into groups based on the similarity of their signals in the feature domain. Further, we evaluate the available information from a fuzzy clustering algorithm to determine which sub-arrays are dominated by the direct-path related signals of a sound source in a room, and which cluster receives mixtures of all sources and a high amount of reverberation. In [6], classification is performed for source-dominated clusters and also for the background cluster, because no differentiation between cluster was made. Now, based on an accurate clustering of microphones into source related clusters and a background cluster, accurate analyses of spatially diverse acoustic scenes can be performed [11].

The paper is structured as follows. In Section 2, we introduce the feature extraction and the clustering method. Further aspects on the clustering of microphones and the analyzed scenario are covered in Section 2.4. The evaluation and the results are presented in Section 3 and 4, before we conclude the paper in Section 5.

2 Methods

2.1 Acoustic Scenario

We assume that Q audio sources are simultaneously active in a room and D receivers in that room pick up mixtures of source signals and reverberation. Due to the ad-hoc MA scenario that we consider, we may assume that some microphones are relatively close to an active audio source, such that they are clearly dominated by the signal from that source and such that the direct path sound energy is higher than the reverberated sound energy at those microphones. Other devices are located elsewhere in the room, possibly close to other sources, or in a region where the energy of reverberation is larger than the direct path sound energy of any source. The dominance of a source signal at a microphone is described by the critical distance [12]. For our ad-hoc MA scenario, we thus assume that some microphones are located within the critical distance of a particular source, and others are located outside of the critical distances of all sources, such that these receive a mixture of different source signals, by reverberation, and by noise.

2.2 Mod-MFCC Feature vector computation

For our investigation we consider a cepstro-temporal representation of the signal which has provided good accuracy in classification problems before [13, 14]. The feature vector is based on the modulations of the Mel-frequency cepstral coefficients (MFCC) [15] of a signal. Thus, we call the features Mod-MFCC features. The feature extraction has been introduced in detail, e.g., in [11, 6]. First, a short-time Fourier transform (STFT) representation of a microphone signal is computed, based on which we derive the MFCCs. Then, we compute the modulation spectrum of the MFCCs in order to consider the temporal evolution of a signal. As we aim to generate a feature representation which behaves rather stationary in comparison to short-time audio features, the absolute values of the modulation spectra are averaged over time which results in averaged absolute modulation amplitudes $\tilde{X}_{\text{mfcc}}(\nu,\eta)$ for each modulation frequency bin ν and MFCC η. To obtain an even more compact representation we compute two cepstral modulation ratios (CMR)

$$\rho_{\nu_1|\nu_2}(\eta) = \frac{\sum_{\nu=\nu_1}^{\nu_2} \tilde{X}_{\text{mfcc}}(\nu,\eta)}{(\nu_2-\nu_1+1)\tilde{X}_{\text{mfcc}}(0,\eta)}, \tag{1}$$

as well as an averaged modulation amplitude $\bar{X}_{\text{mfcc}}(\eta)$ over all modulation frequencies.

For the work, we sample microphone signals at $f_s = 16$ kHz and process them in blocks of $T = 4$ seconds and extract Mod-MFCC features as described. For the spectral and cepstral analysis, the frame length is 512 and the frame shift is 256 samples. For the cepstral modulation analysis the frame length and shift is $L = 16$ and $R = 8$,

ISBN 978-3-8007-4275-2

respectively. Feature vectors are computed for the first 30 MFCCs. The CMRs $\rho_{\nu_1|\nu_2}(\eta)$ for $\nu_1 = 1, \nu_2 = 1$ as well as for $\nu_1 = 2, \nu_2 = 8$ are computed. The CMRs and the averaged modulation spectrum $\bar{X}_{\text{mfcc}}(\eta)$ are rewritten in vector notation (for $\eta = 1, \ldots, 30$) as $\boldsymbol{\rho}_{1|1}$, $\boldsymbol{\rho}_{2|8}$ and $\bar{\mathbf{X}}_{\text{mfcc}}$, and finally stacked into one vector $\mathbf{v} = (\bar{\mathbf{X}}_{\text{mfcc}}^T, \boldsymbol{\rho}_{1|1}^T, \boldsymbol{\rho}_{2|8}^T)^T$. Thus, $A = 90$ features summarize 4 seconds of audio data.

2.3 Fuzzy clustering

For the grouping of microphones we use a soft-clustering method as it provides insights into the homogeneity of the captured spatial data. Data from devices closer together is more similar than data from devices placed at a larger distance as those would probably be close to different sources. Moreover, data from devices that receive balanced mixtures of a variety of signals should have an intermediate similarity value. Therefore, it is rather useful to determine soft weights instead of hard-clustering the data. In the following we use the term cluster for microphones that are in one neighborhood and are dominated by one type of audio signal, e.g., by a specific source signal or by reverberation. The assignment of a microphone to a cluster is performed by evaluation of the fuzzy membership values (FMV) with respect to each of the clusters, which we obtain in the process of the fuzzy clustering. Thus, a microphone is assigned to a cluster for which its FMV is maximal.

Given is a set of D observations $\Omega = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$ in the A-dimensional Euclidean Space $\mathbb{R}^A$. $\mathbf{v}_d$ is the d-th observation (feature) vector and v_{dj} the j-th feature of $\mathbf{v}_d$, with $j \in \{1, 2, \ldots, A\}$. A hard partition of Ω into N clusters $\Omega_1, \Omega_2, \ldots, \Omega_N$, with $2 \leq N \leq D$ satisfies the following three conditions: each of the clusters is a non-empty set and therefore contains some cluster members. An intersecting set of two clusters is the empty set and the union of all N clusters yields the observation set Ω. Corresponding to the definition of Fuzzy Sets [16], one may generalize this to a fuzzy partition of Ω by replacing the assignment of an observation d to a cluster n $\mu_{nd} \in \{0, 1\}$ by $\mu_{nd} \in [0, 1]$, which now can be interpreted as a continuous membership grade of $\mathbf{v}_d$ in the fuzzy clusters of Ω.

Several algorithms have been proposed to estimate an optimal fuzzy partition of Ω. The most studied and most popular method is the Fuzzy c-Means algorithm (FCM). It evaluates a least-squared error functional, given as [17]

$$J_m(\nabla, \mathbf{u}) = \sum_{d=1}^{D} \sum_{n=1}^{N} (\mu_{nd})^{\alpha} ||\mathbf{v}_d - \mathbf{u}_n||_{\boldsymbol{\beta}}^2 \qquad (2)$$

and iteratively estimates the matrix ∇ (which contains all μ_{nd}) and the cluster centers $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$, where $\mathbf{u}_n = (u_{n,1}, u_{n,2}, \ldots, u_{n,A})^T$ is the center of cluster n. In (2), a weighting exponent α with $1 \leq \alpha \leq \infty$ and a distance norm $||\,||_{\boldsymbol{\beta}}$ on $\mathbb{R}^A$ are used, where different types of weighting matrices $\boldsymbol{\beta}$ result in, e.g., the squared Euclidean norm, diagonal norm or Mahalanobis norm [18].

2.4 Using fuzzy membership values to detect the background cluster

An assignment of microphones to clusters based on the highest soft-clustering value does not exploit the full information of soft-cluster estimation. However, probably not all microphones in our scenario are located in the critical distance of one of the sources and thus, not all have a high signal-to-interference ratio (SIR). For the detection of microphones with low SIRs, the estimation of a *background* cluster is useful. Therefore, we aim to estimate $N = Q + 1$ clusters: one cluster for each of the Q sources and one additional cluster. The similarity of signals received by microphones in a first source-related cluster and in this background cluster should be higher than the similarity of microphone signals which are in clusters that are dominated by different sources. Thus, when fuzzy clustering is performed, the second highest FMV for microphones in a source-related cluster should be assigned to the background cluster which receives a mixture of signals at rather even levels. As all other remaining clusters are dominated by different sources, they should have a lower amount of similarity to our first cluster which should be reflected in the respective FMVs. Note, that in this work we assume the number of sources as known.

Table 1: Example of fuzzy membership values for 7 microphones that are clustered into 3 clusters. Based on its highest FMV (bold), a microphone is assigned to a cluster.

	Cluster n		
Mic. d	1	2	3
1	0.1	*0.3*	**0.6**
2	**0.5**	*0.3*	0.2
3	**0.7**	*0.2*	0.1
4	*0.25*	0.2	**0.55**
5	*0.25*	**0.6**	0.15
6	0.15	**0.65**	*0.2*
7	0.05	*0.15*	**0.8**

Table 2: Averaged fuzzy membership values for microphones in clusters 1-3. For those clusters which are dominated by a source, the cluster with the second highest average FMV (italic) indicates the background cluster.

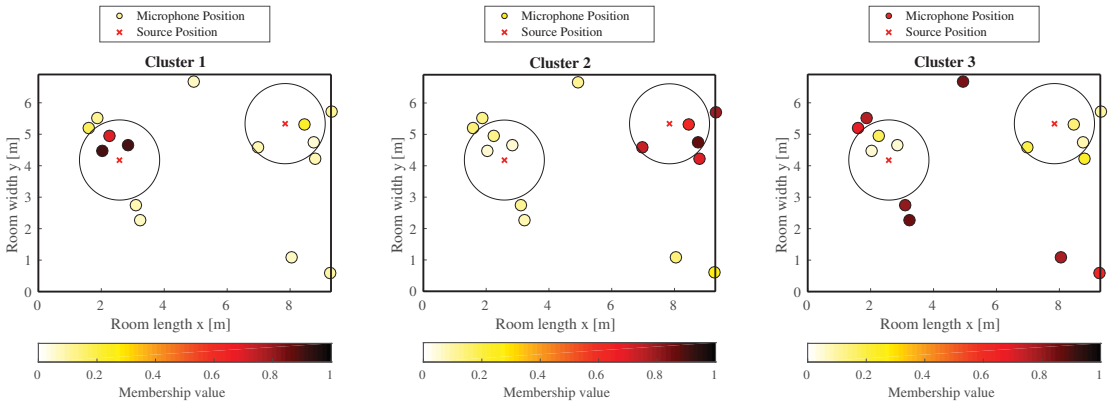
	Cluster n		
	1	2	3
Mics. in Cluster 1	**0.6**	*0.25*	0.15
Mics. in Cluster 2	*0.2*	**0.63**	0.17
Mics. in Cluster 3	0.13	*0.22*	**0.65**

We illustrate the operation of the algorithm with an example. For the detection of the background cluster, Tab. 1 presents FMVs of a clustering result. With respect to the highest FMV for a microphone, we can assign the microphones to cluster 1 (mic. 2+3), cluster 2 (mic. 5+6) and cluster 3 (mic. 1+4+7). Now, we average the FMVs over all microphones in a cluster (Tab. 2). After this, we search for the second highest values of FMVs (in each row of Tab. 2). Both, cluster 1 and cluster 3, have a higher similarity in terms of the FMV for cluster 2, than for the respective other cluster (which is cluster 3 for cluster 1 and vice versa) which leads to the conclusion that cluster 2 is the background cluster. Note that the detection of the background cluster is based on a majority decision and can only performed for the scenario of at least 2 sources which results in a 3 cluster estimation task.

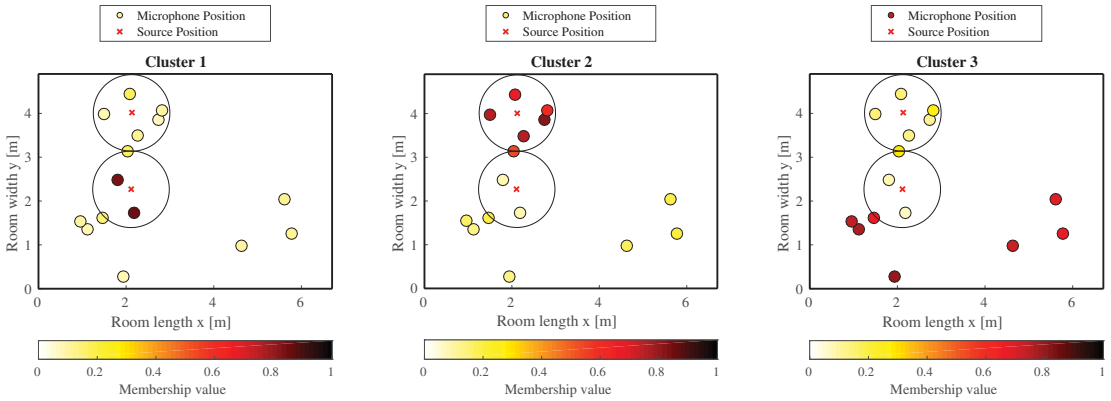
3 Evaluation

For the evaluation of the above assignment scheme we simulate a scenario with two sources and 15 randomly placed microphones. For a meaningful evaluation of clustering results, we ensure in our simulations that the number of microphones within the critical distance of each source is between two and four. We simulate three different room

ISBN 978-3-8007-4275-2



(a) Distant source positions.



(b) Close source positions.

Figure 1: Fuzzy clustering based on extracted audio features from simulated male/female speech signals. The coloring indicates the fuzzy membership value for a microphone in each cluster. The critical distance is indicated.

sizes (with different reverberation characteristics), which are described in detail in [6]. For each simulated room size, we generate 20 scenarios of randomly placed sources and microphones where 10 follow the first and the other 10 the second definition regarding the source placement in the room:

1. *Distant source positioning*: source 1 is randomly positioned in the left and source 2 is randomly positioned in the right half of each room. This scenario represents a good separation of sources in a room.
2. *Close source positioning*: Both sources are positioned randomly in the room, and their distance is exactly twice the critical distance. This scenario still provides well separated sources. However, the distance between the sources decreases and thus the contribution of the competing source signal in a microphone increases.

Furthermore, we record a scenario of two well separated sources with 16 microphones in an acoustic laboratory for reverberation times $T_{60} = \{300, 500, 700\}$ ms. The signals that are played by the sources are for the first experiment speech signals (source 1) and music signals (source 2) and for the second experiment male speech signals (source 1) and female speech signals (source 2). Speech signals are taken from [19]. Music signals are from a private database and include different genres such as rock and pop, instrumental classic and jazz, as labeled using the *allmusic* guidelines [20]. In each of the simulated rooms and in the acoustic laboratory, 100 different combinations of speech and music (or male and female speech) are used, and the results are averaged over all of these signal combinations as well as over all of the respective configurations of rooms. In our work we use a freely available MATLAB® implementation of the fuzzy c-means algorithm [21]. Main parameters for the FCM are the number of clusters which we set to $N = 3$, a weighting exponent which we select as $\alpha = 2$ and a weighting matrix $\boldsymbol{\beta}$ for the distance computations in the feature space. We select $\boldsymbol{\beta}$ as the identity matrix which results in the Euclidean metric.

The evaluation criteria for the clustering are on the one hand the normalized cluster-to-source distance (described in the following paragraph) and the percentage of correct detection of cluster 3 as the background cluster. Note, that without loss of generality we label the cluster which is close to source 1 as cluster 1, etc. The last cluster $N = Q + 1$ denotes the background cluster.

To evaluate the performance of cluster estimation we introduce the normalized cluster-to-source distance $\tilde{d}_{q,n}$, from cluster n to source q as

$$\tilde{d}_{q,n} = \frac{||\mathbf{r}_{\mathrm{s},q} - \bar{\mathbf{r}}_n||}{||\mathbf{r}_{\mathrm{s},1} - \mathbf{r}_{\mathrm{s},2}||}. \tag{3}$$

In (3), $\mathbf{r}_{\mathrm{s},q}$ is the geometrical position of source q in the room. Respectively, $\bar{\mathbf{r}}_\mathbf{n}$ is the average of the positions of all microphones that are assigned to cluster n, and $||.||$ represents the Euclidean norm. Note, that the information about the position of the microphones and sources is only used for evaluation purposes.

The normalization to the actual distance between the sources in a room is necessary, as this distance and therefore the desired distance between the microphone clusters is different for each scenario of source positions.

When the sources are separated by a large distance, well estimated clusters ($n = 1$ and $n = 2$) result in a small normalized distance of a cluster to its respective source,

ISBN 978-3-8007-4275-2

Table 3: Normalized distance $\tilde{d}_{1,n}$ of cluster n to Source 1 and $\tilde{d}_{2,n}$ to Source 2 based on the fuzzy clustering algorithm. The results are averaged over all room sizes and reverberation scenarios.

(a) Simulation, Distant source positions, Speech+Music.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.11**	1.00	0.50
$\tilde{d}_{2,n}$	1.02	**0.17**	0.89

(b) Simulation, Close source positions, Speech+Music.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.16**	1.07	0.89
$\tilde{d}_{2,n}$	1.04	**0.32**	1.2

(c) Recording, Distant source positions, Speech+Music.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.05**	0.99	0.57
$\tilde{d}_{2,n}$	0.97	**0.07**	0.42

(d) Simulation, Distant source positions, Male+Female Speech.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.16**	0.95	0.67
$\tilde{d}_{2,n}$	0.97	**0.19**	0.75

(e) Simulation, Close source positions, Male+Female Speech.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.24**	0.88	1.25
$\tilde{d}_{2,n}$	0.93	**0.34**	1.15

(f) Recording, Distant source position, Male+Female Speech.

	$n=1$	$n=2$	$n=3$
$\tilde{d}_{1,n}$	**0.05**	0.99	0.99
$\tilde{d}_{2,n}$	0.97	**0.07**	0.43

$\tilde{d}_{1,1}$ or $\tilde{d}_{2,2}$. Further, the normalized distance between a source and the cluster of microphones which is located near the other source in the room, $\tilde{d}_{1,2}$ or $\tilde{d}_{2,1}$, should be close to 1. The source-to-source distance is the normalization factor in (3) and decreases over the two previously mentioned definitions of source placements (*distant source positions*, *close source positions*). This influences the results of our normalized distance measure as we see in the following experiments. However, for a successful cluster estimation, $\tilde{d}_{1,1}$ and $\tilde{d}_{2,2}$ should always be smaller than $\tilde{d}_{1,2}$ and $\tilde{d}_{2,1}$. The effect of the decreasing trend of the normalization factor will influence the normalized distance of cluster 3 to the sources, as well. The microphones of cluster 3 are distributed all over a room. For decreasing source-to-source distances and a rather large cluster distance of the third cluster to each of the sources, the normalized distance is expected to increase.

4 Results

Figure 1 presents two of the simulated scenarios of distributed sources (male and female speech) and microphones. One scenario shows a distant source placement and the other shows a close source placement. For the respective scenarios, the result of the feature-based fuzzy clustering of the microphones into cluster 1-3 is shown in the three plots from left to right. For each cluster, the FMV of the microphones is indicated by the colormap. We observe that the signal similarity for microphones within the critical distance of a source is very well reflected in the feature domain, such that the fuzzy clustering algorithm detects very reliably source related clusters of microphones and the background cluster.

Table 4: Average rate of detection of a cluster as the background cluster that is not dominated by one of the source signals but receives mainly reverberant signal mixtures.

(a) Speech and music sources

	Speech cluster (Cluster 1)	Music cluster (Cluster 2)	Background cluster (Cluster 3)
Simulations	2.0%	8.3%	89.7%
Recordings	2.3%	2.6%	95.1 %

(b) Male and female speech sources

	Male Sp. cluster (Cluster 1)	Female Sp. cluster (Cluster 2)	Background cluster (Cluster 3)
Simulations	3.4%	3.1%	93.5 %
Recordings	3.4%	2.7%	93.9 %

Tables 3a-3c show the resulting normalized distances of the detected clusters to the sources for an active speech and an active music source. For both simulated source distributions as well as for the recordings, the distance of a cluster to its respective source ($d_{1,1}$ and $d_{2,2}$) is relatively small, while the distance to the cluster of the other source ($d_{1,2}$ and $d_{2,1}$) is close to one. The normalized distance of the background cluster to the sources has intermediate values, especially for the case of well separated sources. The results are very similar for the case of an active male and and active female speech source (Tab. 3d-3f).

Using the information provided by the FMV, we obtain the results for the detection of cluster 3 as the background cluster as shown in Tab. 4, averaged over all simulated or all recorded scenarios and signals. For all scenarios, the background cluster is detected correctly in at least 89.7% of the cases for the scenario with speech and music sources and in over 93% of the cases with a male and a female speech source.

5 Conclusion and Discussion

In this work we performed the clustering of microphones in an ad-hoc microphone array based on extracted feature information. In this way, insights into the spatial distribution of sources and microphones in a room can be obtained without the actual knowledge about geometrical positions of the sources and the receivers. The evaluation of information that we obtain in the fuzzy clustering allows to estimate which of the microphones are in the vicinity of a sound source in a room, and which microphones contribute to a cluster that receives mainly signal mixtures and a high amount of reverberation. Experiments with two sound sources show a good clustering performance as well as a high detection rate for the background cluster. The described approach should work in similar fashion for more than two sources as long as the number of sources is given or estimated correctly. However, experiments with a higher number of sources as well as experiments with the number of clusters being larger than N for Q sources should be conducted. Furthermore, the constraint on having some microphones within the critical distance of each source should be relaxed.

ISBN 978-3-8007-4275-2

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proceedings of the IEEE Symposium on Communications and Vehicular Technology (SCVT)*, pp. 1–6, 2011.

[2] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[3] T. C. Lawin-Ore and S. Doclo, "Analysis of the average performance of the multi-channel wiener filter for distributed microphone arrays using statistical room acoustics," *Signal Processing*, vol. 107, pp. 96–108, Feb. 2015.

[4] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[5] Y. Hioka and B. Kleijn, "Distributed blind source separation with an application to audio signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 233–236, 2011.

[6] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Processing*, vol. 107, pp. 21–32, 2015.

[7] I. Himawan, *Speech Recognition Using Ad-Hoc Microphone Arrays*. PhD thesis, Queensland University of Technology, Brisbane, 2010.

[8] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 22–25, 2007.

[9] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Closed-form self-localization of asynchronous microphone arrays," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 139–144, 2011.

[10] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 127–132, 2011.

[11] S. Gergen and R. Martin, "Linear combining of audio features for signal classification in ad-hoc microphone arrays," in *Proceedings of the 11. ITG Fachtagung Sprachkommunikation*, pp. 1–4, 2014.

[12] H. Kuttruff, *Room Acoustics*. London: Applied Science Publishers Ltd, 1979.

[13] R. Martin and A. Nagathil, "Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 321–324, 2009.

[14] M. McKinney and J. Breebaart, "Features for audio and music classification," in *Proceedings of the International Conferece on Music Information Retrieval (ISMIR)*, 2003.

[15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, p. 357, 1980.

[16] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[17] M. S. Yang, "A survey of fuzzy clustering," *Mathematical and Computer Modelling*, vol. 18, pp. 1–16, 1993.

[18] J. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993. Linguistic Data Consortium, Philadelphia.

[20] "Allmusic." http://www.allmusic.com/.

[21] J. Abonyi, "Fuzzy clustering and data analysis toolbox," April 2005. http://www.abonyilab.com/software-and-data/fclusttoolbox.

ISBN 978-3-8007-4275-2