Source separation by fuzzy-membership value aware beamforming and masking in ad hoc arrays

Sebastian Gergen[1], *Rainer Martin*[2], *Nilesh Madhu*[3]

[1]Bochum Institute of Technology gGmbH, Bochum, Germany
[2] Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[3]IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
{*sebastian.gergen@bo-i-t.de*}, {*rainer.martin@rub.de*}, {*nilesh.madhu@ugent.be*}

Abstract

This paper presents the concept of fuzzy-membership value (FMV) aware delay-and-sum beamforming for source separation in reverberant environments using ad hoc distributed microphones. Our approach employs a previously proposed fuzzy clustering algorithm to assign microphones of ad hoc arrays to individual source-dominated clusters and to compute fuzzy-membership values for each microphone and cluster. For each source-dominated cluster we first estimate relative time-differences-of-arrival (TDOA) information from the observed microphone signals and then apply both the TDOA and the FMV information in the beamforming stage. We show that such weighted beamforming improves upon the unweighted case. In a second enhancement stage we then apply cluster-related spectral masks to the output of the beamformers. We validate the proposed approach in three realistically-simulated rooms of different sizes. The method is evaluated by informal listening tests as well as by instrumental quality and intelligibility measures.

1 Introduction

Ad hoc acoustic sensor networks constitute an active field of research with many applications in smart home environments, surveillance and security, and hearing accessories. In these scenarios it is of interest to make optimal use of an arbitrary number of microphones as they are made available through smartphones, personal digital assistants, and Internet-of-Things (IoT) devices. As compared to traditional microphone arrays with a predefined geometry the relative locations of sensors is not known *a priori*, and typically their placement with respect to the audio sources of interest is rather arbitrary. Furthermore, the power budgets for processing data in each sensor node and for the exchange of data between nodes via wireless links is quite limited. Any application of *ad hoc* arrays, including those for audio enhancement and classification, has to deal with these constraints.

In this paper we focus on audio signal enhancement via delay-and-sum beamforming (DSB) [1, 2] and mask-based interference reduction (as discussed, e.g., in [3–5]). We thus extend our previous work [6–8] which introduced a source-related fuzzy clustering method for the aggregation of *ad hoc* microphones. In fuzzy clustering each microphone is allocated a fuzzy (soft) membership value (FMV) for each cluster. The clustering itself is based on predefined signal features and requires only a relatively coarse synchronisation accuracy ($\approx$ 100 ms) between the different microphone signals [9]. We introduce the fuzzy-membership value aware DSB (FMVA-DSB) to select those microphones for the DSB which are in the vicinity of a given source and which will be therefore most beneficial for the enhancement of the source signal. We furthermore use the mask-based source separation scheme proposed in [8] to further enhance the beamformed signals. The exchange of audio signals among nodes is confined to a local neighborhood around each source while all other data is transmitted in aggregated form as audio features or power spectra. This is because, for energy reasons, such arrays should also reduce the amount of data transmitted between nodes, without compromising on the audio quality. This latter aspect is, however, not a focus of the current paper.

The proposed method employs linear filtering-based separation as well as a non-linear mask-based method, the pros and cons of which have been discussed in detail, for instance, in [10, 11]. Similar to other approaches, e.g. based on independent component analysis (ICA, e.g. [12–14]), our method is fully blind, although at the current stage of development we still require knowledge about the number of sources that shall be extracted from the acoustic environment. While other works have opted for e.g. a distributed beamformer based on the transmission of compressed local information [15, 16] or a cascade of local and central beamformers [17], we here use the combination of local beamforming followed by a mask-based post-filter. The power spectra for the computation of the post-filter are derived from the the output of the beamformer such that only one spectrum per microphone cluster needs to be transmitted. Thus, we demonstrate how the fuzzy membership values can be used, along with the assumption of disjointness of the source spectra, to reduce the number of microphones necessary for the separation of audio sources captured by *ad hoc* arrays.

The remainder of the paper is structured as follows: the next section introduces the signal model and the *ad hoc* clustering approach. In Section 3, we explain the method for time-difference-of-arrival (TDOA) estimation and the FMVA-DSB approach. We further describe the method to extract time-frequency masks for each of the source-related clusters and discuss their use in the spatial separation scenario. We evaluate the proposed system on simulated data in Section 4.

2 Signal model and *ad hoc* clustering

The acoustic environment considered in this work (and depicted in Figure 1) consists in general of N acoustic sources and D microphones which are scattered within the environment. The acoustic signal transmission from the N sources to a microphone d may be described as:

$$x_d(t) = \sum_{n=1}^{N} \int_0^\infty h_{nd}(\tau) s_n(t-\tau) \mathrm{d}\tau, \qquad (1)$$

with $s_n(t)$ being the n-th source signal, $h_{nd}(t)$ the impulse response from source n to microphone d, and $x_d(t)$ repre-

senting the resulting microphone signal. The microphone signals are sampled, resulting in $x_d(l)$, where l is the time sample index, and then transformed to the short-time discrete Fourier domain:

$$X_d(k,b) = \text{STFT}[x_d(l)], \tag{2}$$

with k and b representing the frequency bin and time frame indices respectively.

In the fuzzy clustering procedure of our algorithm we utilize a feature set composed of MFCCs and their modulation spectra all of which are computed across signal segments of 4s duration. The effects of reverberation are reduced via cepstral mean normalization. For each microphone and each signal segment we obtain a feature vector v_d which is composed of A features, as described in more detail in [6].

Once we extract the set of A-dimensional feature vectors $\Omega = \{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_D\}$ from all D *ad hoc* distributed microphones, we estimate clusters of microphones which are dominated by one of the sources in the room [6, 18]. To this end, we evaluate a least-squared error functional which is given as

$$J_m = \sum_{d=1}^{D} \sum_{n=1}^{N} (\mu_{n,d})^{\alpha} ||\mathrm{v}_d - \mathrm{u}_n||_{\beta}^2 \tag{3}$$

where $\mu_{n,d} \in [0,1]$ denotes the FMV and the distance between an estimated cluster center u_n, $n \in \{1, ..., N\}$, and an observation v_d is computed as

$$||\mathrm{v}_d - \mathrm{u}_n||_{\beta}^2 = (\mathrm{v}_d - \mathrm{u}_i)^T \beta (\mathrm{v}_d - \mathrm{u}_n). \tag{4}$$

The weighting matrix β can be chosen to implement, e.g., the squared Euclidean norm (used in this work), diagonal norm or Mahalanobis norm [19]. As a result of the iterative optimization process we obtain a fuzzy membership value (FMV) of each microphone and for any source. This step is illustrated in Figure 1 for two localised sources. The bounding boxes around the sources indicate the sub-set of microphones selected for further processing. Note that this sub-set can be of a different cardinality for each cluster (as demonstrated in the figure).

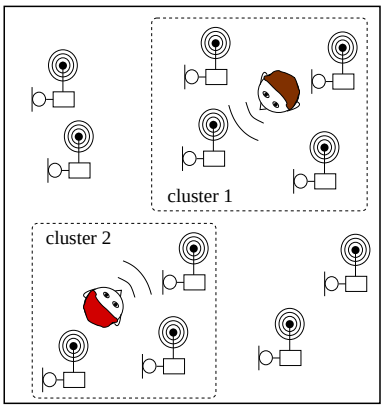

Figure 1: Clustering of microphones around two sources

3 FMV-aware signal enhancement

Beamforming can be carried out using the microphones of a source cluster, if at least the relative delays between the microphones were known for that source. This knowledge is required for both the simple (delay-and-sum) beamformers and the more powerful, statistics-driven ones (e.g. the generalised sidelobe canceller (GSC) [20]). Since the relative locations of the microphones with respect to each other and the dominant source are unknown, one way to estimate these delays is by correlating the microphone signals with that of a *reference* microphone. However, due to the presence of the interference signal and the ambient noise, this is not directly possible. Therefore, for each cluster n we proceed as follows: we first obtain an initial estimate of the source signal ($\widehat{s}_{i_n}(l)$) at all the microphones $d = i_n$ assigned to that cluster. Next, we select a reference microphone for each cluster and perform correlation analysis of all the other microphone signals with respect to this microphone to estimate the TDOAs. These TDOA estimates are subsequently used in the beamforming stage.

3.1 Initial source signal estimation

To compute the initial source estimate we assume that the localised sources are approximately disjoint in the short-time-frequency (T-F) plane. Therefore, only one source may be assumed to be dominant at any one T-F point (k,b). Thus, our goal is to estimate one spectral mask $\mathscr{M}_n(k,b)$ for each cluster and apply it onto the microphone signals (of that cluster). This will provide us with an estimate of the individual, underlying source signal with a reduced amount of interference from other sources. We consider here the case of the *binary* mask given by:

$$\mathscr{M}_n(k,b) = \begin{cases} 1, & \text{if source } n \text{ is dominant at } (k,b); \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This is the simplest separator in the T-F plane, in the absence of further information on the signal power spectra.

To estimate $\mathscr{M}_n(k,b)$ we begin by identifying, for each cluster n, the microphone $d = R_n$ with the highest FMV for that cluster. This microphone serves as the reference microphone for the source signal of that cluster, under the reasonable assumption that if a microphone has a high FMV for a particular cluster, the source in that cluster must dominate over the other sources for that microphone. We then compute the STFT representation $X_{R_n}(k,b)$ of this reference microphone signal of cluster n. The binary mask for cluster n is then obtained as:

$$\mathscr{M}_n(k,b) = \begin{cases} 1 & |X_{R_n}(k,b)| > \dfrac{1}{B} \displaystyle\sum_{b-B+1}^{b} |X_{R_j}(k,b)|, \\ & j = 1, \ldots, N \text{ and } j \neq n, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This is a generalisation of the binary mask traditionally used in the literature (where $B = 1$). The generalisation is required for the following reason: in *ad hoc* arrays the inter-microphone distances can be quite large. Thus, the inter-microphone delay between the different microphones for an impinging signal from a particular sound source is an appreciable fraction of the frame-size used for the STFT. This can lead to a possible *jitter* in the STFT spectral amplitudes across the different microphones. If the non-averaged spectra are used for the mask generation, the masks could flip randomly due to this jitter, leading to undesirable artefacts. By averaging the spectral amplitudes across time, we can reduce the effect of the jitter.

The $\mathscr{M}_n(k,b)$ are then applied to the respective spectra $X_{i_n}(k,b)$ of all microphones i_n assigned to cluster n[1]:

$$\widetilde{X}_{i_n}(k,b) = X_{i_n}(k,b)\mathscr{M}_n(k,b)\,. \tag{7}$$

By computing the inverse STFT of $\widetilde{X}_{i_n}(k,b)$ and reconstructing the time-domain signal by the overlap-add method we obtain $\widehat{s}_{i_n}$, which forms the *initial* estimate of the source signal of cluster n as received at microphone i_n.

3.2 Time-difference-of-arrival estimation

For each cluster n, we compute the TDOAs for all the microphones of that cluster with respect to the reference microphone R_n, using the $\widehat{s}_{i_n}(l)$ for the correlation analysis. This is realised as a *time-domain* cross-correlation, computed over segments of $\sim 4s$ in length, which is also the duration across which the audio features for the fuzzy clustering are computed. For additional accuracy, the cross-correlation function can be interpolated in the region around the correlation peak. A simple 3-point parabolic interpolation usually suffices.

3.3 Clustering-steered beamforming

Given the relative TDOAs for a cluster, a generalised DSB can be formulated, in the *time* domain, as a *weighted* combination of the microphone signals:

$$\widehat{s}_{n,\text{W-DSB}}(l) = \sum_{i_n} w_{n,i_n} x_{i_n}(l + D_{i_n})\,, \tag{8}$$

where the D_{i_n} are the relative TDOAs and w_{n,i_n} is the weight allocated to microphone i_n of cluster n. In [8], the weightings were uniformly set for all microphones of a cluster. Here, we set the weights proportional to the fuzzy-membership value (i.e. $w_{n,i_n} \propto \mu_{n,i_n}$, where μ_{n,i_n} is the FMV for microphone i_n for cluster n). Such weighting gives more importance to microphones with a higher FMV for a cluster (and, we hypothesise, a better SNR). It can be shown that in cases where the microphones have differing input SNRs, importance-weighted DSB yields a better output SNR than uniform weighting. In *ad hoc* arrays, since we have microphones at largely different distances from the source (and, thereby, differing SNRs), we expect the FMV-weighted DSB (output signal denoted as $\widehat{s}_{n,\text{FMVA-DSB}}(l)$) to be better than the simple DSB (denoted as $\widehat{s}_{n,\text{DSB}}(l)$). This hypothesis is tested in this paper.

Furthermore, in contrast to [8], where all the microphones allocated to a cluster were considered for the beamforming, we investigate what happens when we restrict ourselves here to the first I_n microphones with the highest FMV per cluster. Unlike compact arrays, adding an extra microphone to the DSB also introduces extra uncertainty (since the added microphone will have an FMV lower than the microphones currently in the array).

3.4 Mask re-estimation for post-filtering

We use the enhanced signal at the output of the DSB stage to compute a post-filtering mask ($\mathscr{M}_{n,\text{DSB}}(k,b)$) similar to (6). For the case of $\widehat{s}_{n,\text{FMVA-DSB}}(l)$, this mask is obtained as:

$$\mathscr{M}_{n,\text{DSB}}(k,b) = \begin{cases} 1 & |\widehat{S}_{n,\text{FMVA-DSB}}(k,b)| > \dfrac{1}{B}\displaystyle\sum_{b-B+1}^{b} |\widehat{S}_{j,\text{FMVA-DSB}}(k,b)|, \\ & j = 1,\ldots,N \text{ and } j \neq n, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The mask is then applied to $\widehat{S}_{n,\text{FMVA-DSB}}(k,b)$ and the time-domain signal is reconstructed, yielding the final, enhanced estimate of the source in each cluster. The algorithm in its entirety is shown schematically in Figure 2, for the case of two clusters.

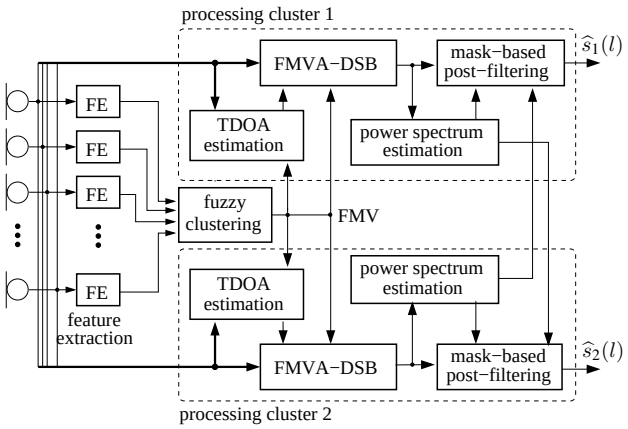

Figure 2: Schematic of the proposed algorithm for the case of 2 clusters.

4 Evaluation & results

For the evaluation we simulate 15 microphones and two active sound sources in three different rooms (see Tab. 2). For each room, we create 10 different scenarios of source-microphone setups. In each setup, $2 \leq D_n \leq 4$ microphones for cluster $n = 1,2$ are randomly located within the critical distance of the respective source. Additional $15 - D_1 - D_2$ microphones are placed randomly all over the room. The position of each of the sources is randomised in one or the other half of each room. We create RIRs using the method in [21]. To generate microphone signals which contain contributions from both sources we convolve male and female speech signals (clean and anechoic, English [22]) with the respective RIRs and add the signals from both sources. Based on the microphone data we extract the audio features from signals of 4 seconds duration, sampled at 16 kHz. The spectral and cepstral analysis is carried out with a frame length of 512 samples and a frame shift of 256 samples.

We use a freely available MATLAB® implementation of the fuzzy c-means algorithm [23] to estimate the clusters based on the extracted feature vectors. Main parameters for the FCM are the number of clusters, which we set to $N = 2$; a weighting exponent, which we select as $\alpha = 2$; and a weighting matrix $\boldsymbol{\beta}$ for the distance computations in the feature space, which we set as the identity matrix (resulting in the Euclidean norm).

The parameter B for the time-frequency masking in (6) and (9) was empirically set to 3.

[1] A microphone is said to be assigned to cluster n if its FMV for cluster n is larger than its FMV for all other clusters

Table 1: Instrumental performance evaluation. Results are averaged across all three rooms and all simulation scenarios. Performance measures for the reference signal are the absolute values and those for the enhanced signals are relative to that of the reference signal.

Reference values of input signals (absolute)						
	seg-SNR (dB)			PESQ		STOI
	-1.33			1.96		0.73
Method	**Beamformer only**			**Beamformer + post-processor**		
	seg-SNRi (dB)	ΔPESQ	ΔSTOI	seg-SNRi (dB)	ΔPESQ	ΔSTOI
$I_n = 3$ microphones						
$\widehat{s}_{n,\text{DSB}}(l)$	3.11	0.26	0.09	4.97	0.43	0.08
$\widehat{s}_{n,\text{FMVA-DSB}}(l)$	3.15	0.26	0.09	5.01	0.43	0.08
$\widehat{s}_{n,\text{pDSB}}(l)$	3.22	0.28	0.10	5.09	0.45	0.09
$\widehat{s}_{n,\text{FMVA-pDSB}}(l)$	3.26	0.28	0.10	5.14	0.45	0.09
$I_n = 4$ microphones						
$\widehat{s}_{n,\text{DSB}}(l)$	3.87	0.33	0.10	5.52	0.49	0.09
$\widehat{s}_{n,\text{FMVA-DSB}}(l)$	3.92	0.33	0.10	5.57	0.49	0.09
$\widehat{s}_{n,\text{pDSB}}(l)$	4.00	0.36	0.12	5.66	0.52	0.11
$\widehat{s}_{n,\text{FMVA-pDSB}}(l)$	4.05	0.36	0.12	5.71	0.52	0.11
$I_n = 5$ microphones						
$\widehat{s}_{n,\text{DSB}}(l)$	4.54	0.37	0.11	5.97	0.51	0.09
$\widehat{s}_{n,\text{FMVA-DSB}}(l)$	4.60	0.37	0.11	6.04	0.51	0.09
I_n : all microphones in cluster n (average $I_n = 7.5(\pm 1.8)$)						
$\widehat{s}_{n,\text{DSB}}(l)$	5.47	0.41	0.13	6.36	0.55	0.11
$\widehat{s}_{n,\text{FMVA-DSB}}(l)$	5.58	0.42	0.13	6.48	0.56	0.11
$\widehat{s}_{n,\text{pDSB}}(l)$	5.67	0.47	0.15	6.60	0.62	0.13
$\widehat{s}_{n,\text{FMVA-pDSB}}(l)$	5.76	0.47	0.15	6.71	0.63	0.14

Table 2: Sizes and information about reverberation time T60 and critical distance r_H of the simulated rooms.

	Size [m^3]	T60 [ms]	r_H [m]
Room 1	$4.7 \times 3.4 \times 2.4$	340	0.6
Room 2	$6.7 \times 4.9 \times 3.5$	490	0.9
Room 3	$9.3 \times 6.9 \times 4.9$	630	1.3

For the performance metrics, the enhanced signal is compared to the noisy mixture signal of the reference microphone of each cluster. The instrumental metrics used are the segmental SNR *improvement* seg-SNRi, the ΔPESQ i.e. the improvement in PESQ [24] and the ΔSTOI [25]. To give an idea of the upper performance bound, we also present the results when using a DSB that incorporates delays computed from the true positions of the source and microphones (pDSB, or position-informed DSB in Table 1). The results presented are the averaged results across all three simulated rooms, simulation scenarios and the two clusters. Note that for the reference signal, the values given are the *absolute values*. For all others, we provide the improvement relative to the reference signal.

We observe the following: (1) FMVA-DSB consistently yields a better performance compared to the simple DSB (mainly in the seg-SNRi). This holds even for the position-informed pDSB which does not use FMV weighting. However, the performance difference is not very large. While this proves the benefit of weighted beamforming, it also shows that a weighting that is simply proportional to the FMV is perhaps not the most optimal. (2) in our simulations, we observe that the average number of microphones per cluster is $7.5(\pm 1.8)$, and the best results are obtained when all the microphones in a cluster are utilised for source separation. However, already with 4 or 5 microphones per cluster, the improvements are quite close to that of using the full cluster. (3) the results using the pDSB is consistently better than that using the estimated TDOAs. However, for the case of limited I_n, we see than the results using estimated TDOAs in a cluster of $I_n + 1$ microphones is equal to, or better than, the results of the pDSB with I_n microphones (e.g. compare the performance of $\widehat{s}_{n,\text{FMVA-DSB}}(l)$ for 4 and 5 microphones to that of $\widehat{s}_{n,\text{FMVA-pDSB}}(l)$ for 3 or 4 microphones). This extra microphone is the price we pay for the inaccuracy in the estimates of the TDOAs. Applying the post-processor (based on the DSB outputs) further improves the PESQ and the segmental SNRi, as compared to the beamformer alone. However, the improvement in STOI is less than that for the beamformer. This is understandable, since the STOI measure is based on the fidelity of the signal envelopes and the binary mask tends to distort the envelope.

5 Conclusions

We have introduced the concept of fuzzy-membership value aware (FMVA) delay-and-sum beamforming (DSB) for source separation in reverberant environments using *ad hoc* distributed microphones. We have compared this approach with the uniformly-weighted beamformer and have demonstrated that while such weighted beamforming improves upon the uniformly-weighted case, the improvement is not significantly large. We believe this has to do with our simulation scenarios where, on average, the microphones are equally divided among the two sources and the constraint of FMV> 0.5 for cluster allocation would tend to produce weights that are roughly similar across microphones. We intend to investigate this in future work.

We have further shown that even choosing only 4-5 microphones within a cluster for the separation allows us to achieve most of the performance gain. The position informed beamformer has the best performance, due to the oracle knowledge incorporated. However, results similar to (or better than) the position informed beamformer can be obtained with the proposed approach, by incorporating one extra microphone into the *ad hoc* array. These are useful results for resource constrained networks.

References

[1] M. Brandstein and D. Ward, eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer-Verlag, 2001.

[2] P. Vary and R. Martin, *Digital Speech Transmission*. Wiley, 2006.

[3] N. Roman, D. Wang, and D. L. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, pp. 2236 – 2252, 2003.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, July 2004.

[5] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[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] S. Gergen and R. Martin, "Estimating source dominated microphone clusters in ad-hoc microphone arrays by fuzzy clustering in the feature space," in *Proceedings of the 12. ITG Fachtagung Sprachkommunikation*, pp. 1–4, 2016.

[8] S. Gergen, R. Martin, and N. Madhu, "Source separation by feature-based clustering of microphones in *ad hoc* arrays," in *Proceedings of 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–4, 2018.

[9] S. Gergen, *Classification of audio sources using ad-hoc microphone arrays*. Dissertation, Ruhr-Universität Bochum, 2016.

[10] N. Madhu and A. Gückel, "Multi-channel source separation: Overview and comparison of mask- based and linear separation algorithms," in *Machine Audition: Principles, Algorithms and Systems* (W. Wang, ed.), pp. 207–245, USA: IGI Global, 2010.

[11] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 692–730, April 2017.

[12] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation based on multi-stage ICA combining frequency-domain ICA and time-domain ICA," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–917–I–920, May 2002.

[13] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, Sept 2004.

[14] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 624–639, March 2011.

[15] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 38–51, Jan 2009.

[16] A. Bertrand and M. Moonen, "Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission," *IEEE Transactions on Signal Processing*, vol. 61, pp. 3447–3459, July 2013.

[17] D. Y. Levin, S. Markovich-Golan, and S. Gannot, "Distributed LCMV beamforming: Considerations of spatial topology and local preprocessing," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 244–248, Oct 2017.

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

[19] 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.

[20] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, pp. 27–34, Jan 1982.

[21] S. Gergen, C. Borß, N. Madhu, and R. Martin, "An optimized parametric model for the simulation of reverberant microphone signals," in *Proc. of the International Conference on Signal Processing, Communications and Computing (ICSPCC 2012)*, 2012.

[22] 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.

[23] J. Abonyi, *Fuzzy Clustering and Data Analysis Toolbox*, April 2005.

[24] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

[25] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125 – 2136, 2011.