Joint Identification and Localization of a Speaker in Adverse Conditions Using a Microphone Array

Daniele Salvati, Carlo Drioli, Gian Luca Foresti
Department of Mathematics, Computer Science and Physics
University of Udine, Italy
Email: {daniele.salvati, carlo.drioli, gianluca.foresti}@uniud.it

***Abstract*—We discuss a joint identification and localization microphone array system based on diagonal unloading (DU) beamforming, which has been recently introduced for acoustic source localization. First, we propose a DU beamformer version for the signal enhancement problem. Then, we propose a enhanced DU steered response power (SRP), in which the first estimate of the source position is further refined with the information gathered from the speaker recognition module. The enhanced SRP-DU is obtained by weighting the frequency components with respect to the spectral characteristics of the speaker. The approach does not add significant computational load to the array processing. Experiments conducted in noisy and reverberant conditions show that the use of the DU beamformer provides better speaker recognition performance if compared to the conventional one since it reduces deleterious effects due to the spatially white noise and point-source interferences. Simulations also show that the speaker identification can improve the localization accuracy, and it is thus interesting for applications and systems which rely on integrated localization and speaker identification.**

***Index Terms*—Acoustic source localization, speaker identification, beamforming, diagonal unloading, microphone array.**

I. Introduction

Microphone array processing techniques are of great interest in various applications such as teleconferencing systems, audio surveillance, autonomous robots, human-computer interaction, and have a central role in a number of applications in the speech technology area. These include speaker localization [1]–[6], speech enhancement [7]–[10], speaker/speech recognition [11]–[15]. Multichannel processing for enhancing the acoustic front end involved in speaker/speech recognition can be advantageous if compared to single-channel case due to the ability in reducing background noise, reverberation, source-point interference, especially in distant-talking conditions [16]. Sensor array techniques such as beamforming [17] and multichannel noise reduction [18] can greatly improve the recognition accuracy in adverse environments (due to noise, reverberation, and multisource conditions). Some possibilities of exploiting the information gathered from a multichannel system have been discussed for example in [11]–[15]. It was demonstrated in [11] that the matched-filter microphone arrays are capable of producing high speaker identification scores in a hostile acoustic environment. In [12], multichannel acquisition and sub-band processing are integrated to design a speech recognition framework in which the channels are exploited to compute a set of separate sub-band mel-frequency cepstral coefficients (MFCCs), and to drive a separate hidden Markov model (HMM) recognizer for each sub-band. By including angle of arrival information in a speaker identification system based on a HMM, a performance improvement was reported in [13] compared to an algorithm based on the speech spectrum only. In [14], the conventional beamforming approach is revisited by the use of a deep neural network (DNN), which improves the computation of the beamforming weights and provides in turn an enhanced MFCC set in the speech recognition. Speaker/speech recognition was enhanced in [15] by using a microphone array and by combining the dynamic time warping associated with all the microphones through a fuzzy control scheme.

A microphone array system for distant speaker/speech recognition typically consists of a localization step, a beamformer and a recognition module [16]. Given the speaker position estimate, the beamformer emphasizes sound waves coming from the direction of interest. The signal enhancement output is then fed into a recognizer. While each module has been deeply investigated (an overview of such system can be found in [16]), the study of joint localization and recognition is rather limited in literature. An example is the work in [19], that however has been developed in the context of binaural applications.

In this paper, we discuss joint speaker identification and localization using a microphone array. We recently introduced the low-complexity robust diagonal unloading (DU) beamforming [20] for the acoustic source localization problem. We propose the use of a novel DU beamforming version in the signal enhancement module and a new approach that refines the localization step based on the speaker recognition. This scenario can be of interest for example in videoconferencing applications, in which the estimation of sound coordinates can be used to automatically steer a videocamera towards an active speaker. We show that the DU beamformer can be successfully adopted for the signal enhancement problem by an appropriate DU procedure improving performance against spatially white noise and point-source interference if compared to the conventional beamformer. The DU beamformer is implemented without any *a priori* knowledge of noise-plus-interferences information, which is in general request in the implementation of high resolution beamformers such the well-known minimum variance distortionless response (MVDR) beamformer [8]. The identification step is computed on the enhanced signal using a Gaussian mixture model (GMM) of the MFCC statistics.

ISBN 978-90-827970-1-5

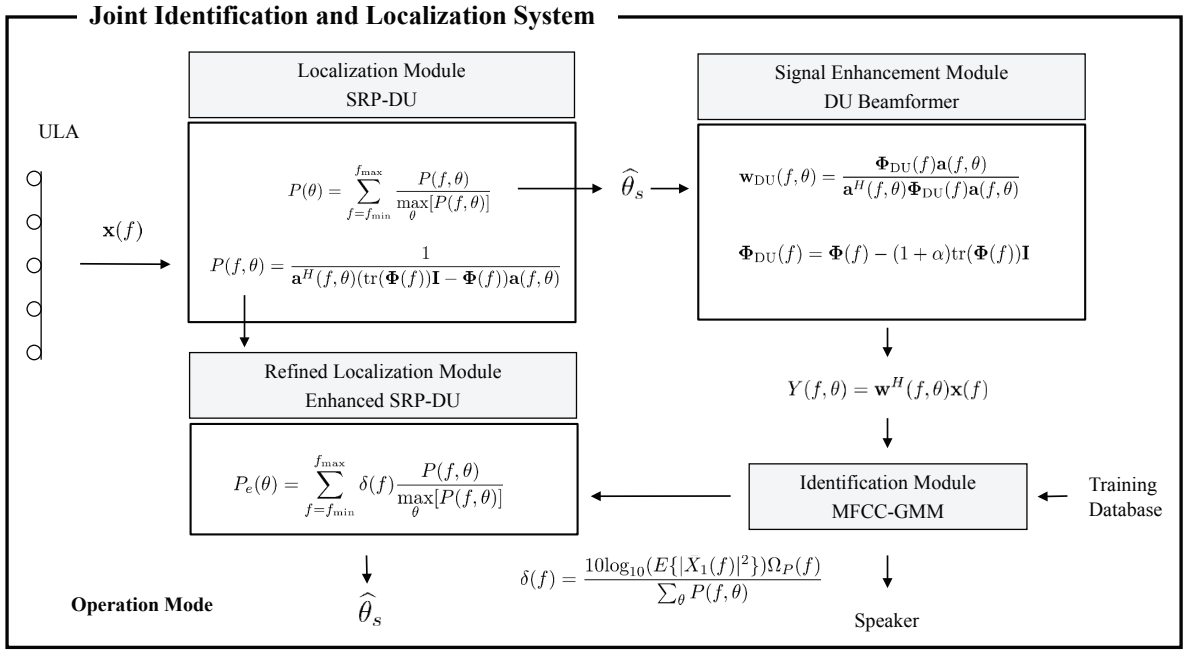

Fig. 1. Schematic diagram of the proposed system for joint identification and localization of a speaker.

II. Joint Identification and Localization

This section describes the proposed system. An overview is given in the first subsection. Then, we summarize the SRP-DU described in [20], and we present the new DU beamformer for the signal enhancement problem. Next, we summarize the identification module used in this paper, and we finally present the new enhanced DU beamformer for the localization problem.

A. System Overview

The speaker identification and localization system of choice in this study is made of a uniform linear array (ULA) of M sensors located in a noisy and reverberant environment. The proposed system consists of four main building blocks: 1) the localization module, 2) the signal enhancement module, 3) the speaker identification module, 4) the refined localization module. A steered response power (SRP) DU beamformer [20] operated in far-field mode to estimate the direction of arrival (DOA) θ is used for the localization, a DU beamformer for the signal enhancement module, a GMM-MFCC scheme for the identification process. When the speaker identification is available, the spectral information is used to refine the speaker localization step in the enhanced SRP-DU computation. The organization of the speaker identification and localization components is illustrated in Figure 1.

B. The Localization Module: SRP-DU

In general, a beamformer is a spatial filter technique which is aimed at leaving untouched the acoustic energy related to sources located along a given direction of arrival, while minimizing the energy of noise and sources coming from directions other than the desired one. An acoustic SRP beamformer is typically computed in the frequency-domain by calculating the response power of each frequency bin and by fusing the narrowband components. The power spectral density (PSD) of the DU beamformer [20] for a frequency bin f is given by

$$P(f,\theta) = \frac{1}{\mathbf{a}^H(f,\theta)(\mathrm{tr}(\mathbf{\Phi}(f))\mathbf{I} - \mathbf{\Phi}(f))\mathbf{a}(f,\theta)}, \quad (1)$$

where $\mathrm{tr}(\cdot)$ is the operator that computes the trace of a matrix, $\mathbf{a}(f,\theta)$ is the steering vector, $\mathbf{I}$ is the identity matrix, H denotes the Hermitian (complex conjugate) transpose, and $\mathbf{\Phi}(f) = E\{\mathbf{x}(f)\mathbf{x}^H(f)\}$ is the PSD matrix of the microphone signals ($E\{\cdot\}$ denotes mathematical expectation and $\mathbf{x}(f) = [X_1(f), X_2(f), \ldots, X_M(f)]^T$ is the vector of the frequency-domain signals sensed by the M sensors, T denotes the transpose operator). The DU beamformer exploits the orthogonality property between signal and noise subspace by the removal or the attenuation of the signal subspace from the PSD matrix using a diagonal unloading procedure (for a detailed discussion on the DU procedure, see [20]). The $P(f,\theta)$ is related to the energy contribution of a single frequency bin, and a function providing the energy information of the whole frequency spectrum (the broadband PSD) can be obtained by merging the contribution by some fusion criterion. In [21], we proposed the following normalized incoherent frequency fusion

$$P(\theta) = \sum_{f=f_{\min}}^{f_{\max}} \frac{P(f,\theta)}{\max_{\theta}[P(f,\theta)]}, \quad (2)$$

where $\max[\cdot]$ denotes the maximum value, $f_{\min}$ and $f_{\max}$ denote the considered frequency range. The normalization lends a high resolution to the spatial spectrum, but emphasizes

ISBN 978-90-827970-1-5 © EURASIP 2018

the noise in those narrowband beamformers in which the SNR ratio is low, thus providing a misleading contribution to the fusion. The broadband PSD resulting from the fusion is characterized by high energy peaks corresponding to those directions from which acoustic energy is sensed, thus the DOA estimation of the active source is provided by maximum energy peak search:

$$\widehat{\theta}_s = \underset{\theta}{\operatorname{argmax}}[P(\theta)]. \tag{3}$$

C. The Signal Enhancement Module: DU Beamformer

The output of a narrowband beamformer $Y(f,\theta)$ for frequency f and look direction θ, is obtained as

$$Y(f,\theta) = \mathbf{w}^H(f,\theta)\mathbf{x}(f), \tag{4}$$

where $\mathbf{w}(f,\theta)$ is a column vector containing the beamformer coefficients for time-shifting, weighting, and summing the data, so to steer the array in the direction θ. The DU transformation is obtained by subtracting an opportune diagonal matrix from the covariance matrix $\mathbf{\Phi}(f)$ of the array output vector. As a result, the DU beamforming removes as much as possible the signal subspace from the covariance matrix and provides a high resolution beampattern. In practice, the design and implementation of the DU transformation is simple and effective, and is obtained by computing the matrix (un)loading factor.

Given the matrix $\mathbf{\Phi}(f)$ which represents the array output vector covariance, the DU transformed matrix can be written as

$$\mathbf{\Phi}'_{\text{DU}}(f) = \mathbf{\Phi}(f) - \mu\mathbf{I}, \tag{5}$$

where μ is a real-valued, positive scalar, selected in such a way that the resulting PSD matrix is negative semidefinite for exploiting the orthogonality property between subspaces. In the single source case with spatially white noise the optimal DU implementation can be obtained by imposing that the eigenvalue corresponding to the signal subspace is null, and that the eigenvalues corresponding to the noise subspace are non-zero. The value of μ that verifies such constraints can be shown to be

$$\mu = \text{tr}(\mathbf{\Phi}(f)) - (M-1)\sigma^2, \tag{6}$$

where σ^2 is the noise variance for all sensors. Theoretically, this solution totally removes the signal subspace from the PSD matrix and the beamformer output is therefore null in the source direction. Note that in practice the orthogonality property is partially exploited due to the reverberation and the multisource scenario. The DU beamformer is formulated by using an optimization problem with an orthogonality constraint that aims to achieve the signal subspace removal [20], the optimization reads as:

$$\begin{aligned} &\text{minimize} \quad ||\mathbf{w}(f,\theta) - \mathbf{a}(f,\theta)||^2, \\ &\text{subject to} \quad \mathbf{u}_s^H\mathbf{w}(f,\theta) = 0, \end{aligned} \tag{7}$$

where $\mathbf{u}_s$ is the signal subspace of $\mathbf{\Phi}(f)$. The solution to the optimization problem is $\mathbf{w}_{\text{DU}}(f,\theta) = \left(\frac{1}{\lambda'_v}\mathbf{I}\right)\mathbf{\Phi}'_{\text{DU}}(f)\mathbf{a}(f,\theta)$, where λ'_v is the noise eigenvalue of the matrix $\mathbf{\Phi}'_{\text{DU}}(\mathbf{f})$ (more details on the analytical solution can be found in [20]). In order to take into account the signal enhancement problem, we generalized here the DU beamformer by defining a parametrized solution given by

$$\mathbf{\Phi}_{\text{DU}}(f) = \mathbf{\Phi}(f) - (1+\alpha)\text{tr}(\mathbf{\Phi}(f))\mathbf{I}, \tag{8}$$

where $\alpha \geq 0$ is a tradeoff parameter between noise reduction and speech distortion. When $\alpha = 0$ the maximum noise reduction and interference suppression is obtained, leading however on a distortion of the speech signal since the PSD matrix contains both signal, noise and interference. By increasing the value α, we can reduce the noise suppression and distortion. We adopt successfully a value $\alpha = 0.5$ in the simulation section. The DU procedure provides a solution for the following beamforming coefficients $\mathbf{w}$:

$$\mathbf{w}_{\text{DU}}(f,\theta) = \frac{\mathbf{\Phi}_{\text{DU}}(f)\mathbf{a}(f,\theta)}{\mathbf{a}^H(f,\theta)\mathbf{\Phi}_{\text{DU}}(f)\mathbf{a}(f,\theta)}. \tag{9}$$

Given the beamforming coefficients, is thus possible to compute the spatially filtered signal with respect to the estimated DOA, which will provide an enhanced version of the target source signal.

D. The Identification Module: MFCC-GMM

Given that Q is the number of triangular filters in the mel filter bank, the output of the equalized filter bank in the signal enhancement output is computed as

$$e_i = \log\Big(\sum_{f=f_{\text{min}}}^{f_{\text{max}}} E\{|Y(f,\theta)|^2\}\cdot\psi_i(f)\Big),\ i = 1,...,Q, \tag{10}$$

where $\psi_i(f)$ is the i-th triangular bandpass filter. The MFCC vector is then computed as:

$$\mathbf{m} = \mathbf{C}\mathbf{e}, \tag{11}$$

where $\mathbf{e} = [e_1, e_2, \ldots, e_Q]^T$, and $\mathbf{C} = [c_{ij}]$ is the discrete cosine transform (DCT) matrix

$$\begin{aligned} c_{ij} = \sqrt{\frac{2}{Q}}\cos\Big(\frac{\pi(i-1)(j-0.5)}{Q}\Big), \\ i = 1,2,\ldots,N, \quad j = 1,2,\ldots,Q, \end{aligned} \tag{12}$$

where N is the dimension of the MFCC vector that includes the zero-th order coefficient. The speaker identification is finally based on a conventional GMM statistical model of the MFCC cues computed as above.

E. The Refined Localization Module: Enhanced SRP-DU

When the speaker identification is available, the enhanced SRP-DU is adopted in the localization computation. Similarity to [22], the weighting factors are used to attenuate the errors in the broadband SRP fusion process. The weighting factors are related to the spectral information of a known speaker, and on some features of the PSD function: the skewness and the overall energy of the PSD. The enhanced DU is given by

$$P_e(\theta) = \sum_{f=f_{\text{min}}}^{f_{\text{max}}} \delta(f)\frac{P(f,\theta)}{\max\limits_{\theta}[P(f,\theta)]}, \tag{13}$$

ISBN 978-90-827970-1-5

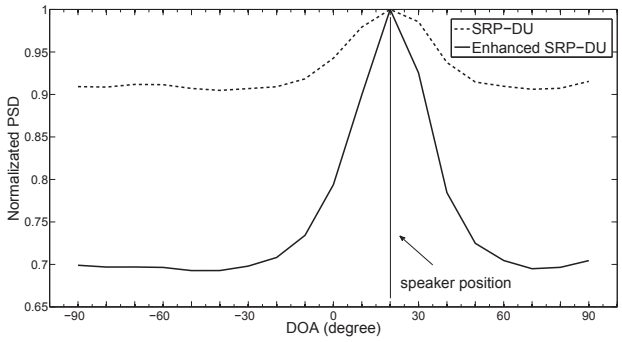

Fig. 2. The PSD of the SRP-DU and the enhanced SRP-DU.

TABLE I
THE IDENTIFICATION AND LOCALIZATION PERFORMANCE IN SPATIALLY WHITE NOISE CASE WITH A RT_{60} OF 0.2 s.

	Speaker identification			Speaker localization	
SNR (dB)	AI (%) SC	DS	DU	RMSE (degree) SRP-DU	Enhanced SRP-DU
20	96.67	100	100	5.77	3.03
15	81.67	98.33	98.33	6.09	3.08
10	53.33	95.00	95.00	5.48	4.11
5	23.33	78.33	80.00	6.65	3.59
0	15.00	38.33	73.33	7.27	4.09
-5	11.66	18.33	55.00	7.93	4.49
-10	5.00	16.66	41.66	8.65	5.42

where

$$\delta(f) = \frac{10\log_{10}(E\{|\bar{X}_1(f)|^2\})\Omega_P(f)}{\sum_\theta P(f,\theta)}, \qquad (14)$$

where $E\{|\bar{X}_1(f)|^2\}$ is the averaging signal power at reference sensor obtained during the training phase, $P(f,\theta)$ is calculated using equation (1), and $\Omega_P(f)$ is the skewness of the PSD $P(f,\theta)$ for all candidate DOAs in the analysis frame defined as

$$\Omega_P(f) = \frac{\frac{1}{N_\theta}\sum_\theta [P(f,\theta) - \bar{P}(f,\theta)]^3}{(\frac{1}{N_\theta - 1}\sum_\theta [P(f,\theta) - \bar{P}(f,\theta)]^2)^{\frac{3}{2}}}, \qquad (15)$$

where $\bar{P}(f,\theta) = \frac{1}{N_\theta}\sum_\theta (P(f,\theta))$ and N_θ is the number of candidate DOAs. The skewness measure gives information of the peakedness distribution of a PSD function [22]. An ideal normalized PSD function is characterized by a delta Dirac in the source position with a null value in the other directions. This means that the overall energy of the PSD spectrum is equal to 1, and the skewness measure has the maximum possible value. Hence, the enhanced DU assigns higher importance to the narrowband PSD components that are more similar to an ideal SRP and that have affinities with the mean spectrum of the speaker. Figure 2 shows the enhancing effect of weighting the narrowband power responses, visible in terms of a higher resolution for the DOA related to the source position and of a larger attenuation of the power for the other directions.

III. SIMULATIONS

The speaker identification and localization performance observed is illustrated through a set of simulated experiments. A reverberant room of 7 m $\times$ 5 m $\times$ 3 m simulated with an improved image-source model [23] was used. A ULA of 8 microphones, with distance between microphones of 0.04 m, was used. A localization task in two-dimensions was considered. Therefore, both microphones and the source were positioned at a distance from the floor of 1.3 m. The spatial resolution was 1 degree. The sampling frequency was 48 kHz, the block size L was 2048 samples with an overlap of 512 samples. A Hann window was used. A frequency range between $f_{\text{min}} = 80$ Hz and $f_{\text{max}} = 8000$ Hz was used. The GMM model parameters were set to 50 for the components and 0.01 for the regularization parameter. MFCC vectors of length $N = 22$ were used as input to the GMM models. Speech signals from the TSP speech database was used [24]. In the training phase, we consider the speaker source positions at a distance from the array of 0.8 m and a DOA of 0 degree. Recordings of 1 min each from 20 different speakers (10 males, 10 females) were used to train their respective GMM models through a conventional expectation-maximization algorithm. The training of the GMMs was conducted by setting an averaging SNR of 30 dB, which was obtained by adding mutually independent white Gaussian noise to each channel. The identification tests were conducted for spatially white noise and point-source interference conditions by using recordings of 3 s each from the 20 speakers. We compared the performance of the speaker identification based on conventional single channel (SC) MFCC-GMM, on the microphone array system with conventional beamforming, i.e., the delay and sum (DS) beamformer [17], and with the proposed DU. To reduce the temporal fluctuations due to the nonstationary nature of the speech signal, the broadband PSD was smoothed with a single-pole recursive filter for both SRP-DU and enhanced SRP-DU, which are implemented with a single snapshot. Three positions were used during the Monte Carlo simulation with distance from the array and DOA: (2 m, -8.8 degrees), (1.3 m, -28.6 degrees), (1.6 m, 15.3 degrees). In each position, the identification and localization performance is measured for all 20 speakers. Performance is reported in terms of the percentage of identification accuracy (IA) estimates for the speaker identification and by the root mean square error (RMSE) for all the estimates for the speaker localization. Tables I and II show the speaker identification and DOA estimation results for the spatially white noise case at variation of SNR and in two reverberant time (RT_{60}) conditions: 0.2 s and 0.5 s. As we can observe, the identification based on DU beamformer outperforms the one based on SC and on DS beamformer when the noise level increases. In low SNR conditions, range [5,-15] dB, we obtain a significant IA increment by using the proposed DU beamformer. We can also note the improved performance of the enhanced SRP-DU in all conditions. Next, the evaluation with a point-source interference was performed by considering a fan noise signal at position: 3 m far from the center of the array with a DOA of 84.4 degrees. The RT_{60} was set to 0.2. The results depicted in

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE II
THE IDENTIFICATION AND LOCALIZATION PERFORMANCE IN SPATIALLY WHITE NOISE CASE WITH A RT_{60} OF 0.5 S.

	Speaker identification			Speaker localization	
SNR (dB)	AI (%) SC	DS	DU	RMSE (degree) SRP-DU	Enhanced SRP-DU
20	100	100	100	5.94	3.76
15	90.00	96.66	96.66	6.31	3.96
10	56.66	93.33	93.33	6.79	4.30
5	28.33	70.00	78.33	6.72	4.63
0	16.66	33.33	65.00	8.17	5.12
-5	13.33	16.66	51.66	8.71	5.82
-10	5.00	15.00	25.00	10.47	7.65

TABLE III
THE IDENTIFICATION AND LOCALIZATION PERFORMANCE IN POINT-SOURCE INTERFERENCE CASE WITH A RT_{60} OF 0.2 S AN SNR OF 30 DB.

	Speaker identification			Speaker localization	
SIR (dB)	AI (%) SC	DS	DU	RMSE (degree) SRP-DU	Enhanced SRP-DU
20	96.66	98.33	98.33	5.85	5.55
15	81.67	90.00	90.00	6.10	5.89
10	10.00	75.00	75.00	6.79	6.34
5	21.66	51.66	55.00	8.47	7.47
0	8.33	31.66	40.00	17.72	13.04
-5	5.00	11.66	23.33	21.45	18.64
-10	5.03	6.67	6.67	29.78	28.08

Table III at variation of the signal-to-interference ratio (SIR) show the improvement performance for both identification and localization up to an SNR of -5 dB. However, we can note smaller differences between DS and DU if compared to the adverse spatially white noise condition.

IV. CONCLUSIONS

In this paper, we presented a joint identification and localization microphone array system based on DU beamforming. We presented a DU version specifically designed for the signal enhancement problem. We showed that the use of DU beamformer can improve the speaker identification in adverse spatially white noise and point-source interference conditions. Beside that, we introduced an enhanced SRP-DU algorithm that is based on the spectral information of the speaker and on skewness and overall energy measures of narrowband PSD functions. We showed that the localization accuracy can be improved if the speaker is known.

REFERENCES

[1] F. Ribeiro, C. Zhang, D. A. Florencio, and D. E Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1781–1792, 2010.

[2] D. Salvati, A. Rodà, S. Canazza, and G. L. Foresti, "A real-time system for multiple acoustic sources localization based on ISP comparison," in *Proceedings of the International Conference on Digital Audio Effects*, 2010, pp. 201–208.

[3] Y. Tian, Z. Chen, and F. Yin, "Distributed imm-unscented kalman filter for speaker tracking in microphone array networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1637–1647, 2015.

[4] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognition Letters*, vol. 84, pp. 15–21, 2016.

[5] D. Salvati, C. Drioli, and G. L. Foresti, "Sound source and microphone localization from acoustic impulse responses," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1459–1463, 2016.

[6] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 586–601, 2017.

[7] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1513–1523, 2013.

[8] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 67–79, 2014.

[9] J. Traa, D. Wingate, N. D. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 493–503, 2016.

[10] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3351–3361, 2016.

[11] Qiguang Lin, Ea-Ee Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 622–629, 1994.

[12] I. A. McCowan and S. Sridharan, "Multi-channel sub-band speech recognition," *EURASIP Journal on Applied Signal Processing*, pp. 45–52, 2001.

[13] J. W. Stokes, J. C. Platt, and S. Basu, "Speaker identification using a microphone array and a joint HMM with speech spectrum and angle of arrival," in *Proceeding of the IEEE International Conference on Multimedia and Expo*, 2006, pp. 1381–1384.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5745–5749.

[15] Ing-Jr Ding and Jia-Yi Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Computers & Electrical Engineering*, vol. 62, pp. 719–729, 2017.

[16] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[17] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[18] S. Doclo, M. Moonen, T. V. 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, no. 1, pp. 38–51, 2009.

[19] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[20] D. Salvati, C. Drioli, and G. L. Foresti, "A low-complexity robust beamforming using diagonal unloading for acoustic source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 609–622, 2018.

[21] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.

[22] D. Salvati, C. Drioli, and G. L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2016.

[23] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[24] P. Kabal, "TSP speech database," Tech. Rep., McGill University, Montreal, Quebec, 2002.

ISBN 978-90-827970-1-5