



Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A weighted MVDR beamformer based on SVM learning for sound source localization[☆]

Daniele Salvati^{*}, Carlo Drioli, Gian Luca Foresti

University of Udine, Department of Mathematics, Computer Science and Physics, via delle Scienze 206, Udine-33100, Italy

ARTICLE INFO

Article history:

Received 21 December 2015

Available online xxx

Keywords:

Sound source localization

Microphone array

Weighted minimum variance distortionless response

Support vector machine

Reverberant environment

ABSTRACT

A weighted minimum variance distortionless response (WMVDR) algorithm for near-field sound localization in a reverberant environment is presented. The steered response power computation of the WMVDR is based on a machine learning component which improves the incoherent frequency fusion of the narrowband power maps. A support vector machine (SVM) classifier is adopted to select the components of the fusion. The skewness measure of the narrowband power map marginal distribution is showed to be an effective feature for the supervised learning of the power map selection. Experiments with both simulated and real data demonstrate the improvement of the WMVDR beamformer localization accuracy with respect to other state-of-the-art techniques.

© 2016 Published by Elsevier B.V.

1. Introduction

Sound source localization using microphone arrays is of considerable interest in an increasing number of applications: teleconferencing systems [23], audio surveillance [26], autonomous robots [5], animal ecology [19], musical control interfaces [25], hearing aid [11], volcanology research [24], and medical intervention [18].

The steered response power (SRP) algorithms, which are based on maximizing the power output of a beamformer, are a robust class of methods used to estimate the sound source position in space. Typically, broadband SRP is computed in the frequency-domain by calculating the response power on each frequency bin and by fusing the narrowband SRP with incoherent [1,10,27] or coherent [9,31,32] averaging with respect to frequency.

For increasing the spatial resolution of the broadband SRP, usually a normalization of narrowband power maps is computed before the fusion of the maps. The well-studied SRP algorithm based of phase transform (PHAT) [10] considers only the phase information for computing the normalization. In [27], it is shown that a post-filter normalization of each narrowband power map substantially improves the spatial resolution of the minimum variance distortionless response (MVDR) [3] beamformer, which is more robust against noise if compared to other algorithms. Hence, the normalization provides significant advantages in reverberant

environments since it allows a better identification of direct path and reflections. Unfortunately, the normalization has the disadvantage of emphasizing the noise in those frequencies in which the signal-to-noise ratio (SNR) is low, resulting in large errors that may provide an inaccurate final frequency data combination.

In this paper, we consider the near-field sound localization problem of a single source in reverberant environments. This scenario can be of interest in videoconferencing applications [23], in which the estimation of sound coordinates can be used to automatically steer a videocamera towards an active speaker; or in human-computer interaction systems, in which the localization is used in a signal enhancement beamformer for speech recognition or dictation system; or even in multimedia interactive systems for performing arts, in which a performer can interact with a computer by using the space-time information of an acoustic source including voice, a musical instrument, or a sounding object to control a creative expressive domain [25].

We present a weighted MVDR (WMVDR) broadband beamformer, which is based on a normalized MVDR (NMVDR) [27] and on a support vector machine (SVM) [6] classifier, which is trained to classify the narrowband power maps into two classes: constructively contributing maps vs. disruptively contributing ones. The idea of a weighted MVDR was proposed in [28], in which a machine learning approach for selecting narrowband power maps was introduced, using a radial basis function network (RBFN) classifier and the marginal distribution of the narrowband power maps as input. In contrast to [28], we propose the use of a SVM learning component and of statistical features of the marginal distributions of the narrowband power maps as input, in order to remove the

[☆] This paper has been recommended for acceptance by Egon L. van den Broek.

^{*} Corresponding author. ✉

E-mail address: daniele.salvati@uniud.it (D. Salvati).

dependency from the size of the analysis region and to drastically reduce the dimensionality of the input vector. We investigate the use of three statistical features related to the marginal distribution: skewness, kurtosis, and crest factor. We show that the best performance is obtained with the skewness measure and that the use of the SVM outperforms the RBFN. In the experimental section, we provide extensive acoustic source localization experiments based on both synthetic and real data.

If compared to other supervised learning approaches [2,8,15,16,20], in which classifiers are used to directly map the acoustic cues onto a position in the search space, in the proposed scheme the machine learning component complements the SRP method, thus providing an incremental contribution to the performance of the SRP-based approaches.

2. The normalized MVDR beamformer

In this paper, we will make use of standard notational conventions. Vectors and matrices are written in boldface with matrices in capitals.

We consider an unknown sound source that is active at time k in a reverberant room of dimension $G = G_x \times G_y \times G_z$ with Cartesian coordinates $\mathbf{r}_s(k) = [x_s(k), y_s(k), z_s(k)]^T$, and we assume the source to be in the near-field. We can write the positions of the m th microphones as $\mathbf{r}_m = [x_m, y_m, z_m]^T$, $m = 1, 2, \dots, M$, where M is the number of microphones. In the short-time Fourier transform domain the m th reverberant signal can be expressed as

$$X_m(f, k) = H_m(f)S(f, k) + V_m(f, k) \quad (1)$$

where f is the frequency bin index, $S(f, k)$ is the source signal at frequency f and time k , $V_m(f, k)$ is the uncorrelated noise signal, and $H_m(f)$ is the time-invariant acoustic transfer function from the source to the microphone m . We assume that the analysis window L is sufficiently long to capture most of the room impulse response such that the multiplicative transfer function approximation holds.

The output of a beamformer at time k is given by a linear combination of the data

$$Y(f, k, \mathbf{r}_g) = \mathbf{w}^H(f, k, \mathbf{r}_g)\mathbf{x}(f, k) \quad (2)$$

where the superscript H represents the Hermitian (complex conjugate) transpose, the signal vector is $\mathbf{x}(f, k) = [X_1(f, k), X_2(f, k), \dots, X_M(f, k)]^T$, and the weight vector for steering and filtering the data on a position $\mathbf{r}_g = [x_g, y_g, z_g]^T \in G$, which is a candidate position for searching the source, is $\mathbf{w}(f, k, \mathbf{r}_g) = [W_1(f, k, \mathbf{r}_g), W_2(f, k, \mathbf{r}_g), \dots, W_M(f, k, \mathbf{r}_g)]^T$.

The power spectral density (PSD) of the beamformer output is given by

$$P(f, k, \mathbf{r}_g) = E[|Y(f, k, \mathbf{r}_g)|^2] = \mathbf{w}^H(f, k, \mathbf{r}_g)\Phi(f, k, \mathbf{r}_g)\mathbf{w}(f, k, \mathbf{r}_g) \quad (3)$$

where $\Phi(f, k) = E[\mathbf{x}(f, k)\mathbf{x}^H(f, k)]$ is the cross-spectral density matrix and $E[\cdot]$ denotes mathematical expectation.

The MVDR beamformer [3] is a well-known beamforming technique which is aimed at minimizing the energy of noise and sources coming from different directions, while keeping a fixed gain on the desired position. The MVDR filter relies on the solution of the following minimization problem

$$\mathbf{w}_c(f, k, \mathbf{r}_g) = \underset{\mathbf{w}(f, k, \mathbf{r}_g)}{\text{argmin}} \mathbf{w}^H(f, k, \mathbf{r}_g)\Phi(f, k, \mathbf{r}_g)\mathbf{w}(f, k, \mathbf{r}_g) \quad (4)$$

subject to $\mathbf{w}^H(f, k, \mathbf{r}_g)\mathbf{a}(f, \mathbf{r}_g) = 1$

where $\mathbf{a}(f, \mathbf{r}_g)$ is the steering vector corresponding to a space position \mathbf{r}_g . The steering vector depends on the time difference of arrival (TDOA) of the spherical wavefront between microphones taking into account the signal attenuation. We can write the TDOA

between microphone i and j as

$$\tau_{i,j}^{\mathbf{r}_g} = \frac{\|\mathbf{r}_i - \mathbf{r}_g\| - \|\mathbf{r}_j - \mathbf{r}_g\|}{c} \quad (5)$$

where $\|\cdot\|$ denotes Euclidean norm and c is the speed of sound. In the near-field, the steering vector takes the form

$$\mathbf{a}(f, \mathbf{r}_g) = [1, \chi_2 e^{\frac{j2\pi f \tau_{1,2}^{\mathbf{r}_g}}{L}}, \dots, \chi_M e^{\frac{j2\pi f \tau_{1,M}^{\mathbf{r}_g}}{L}}]^T \quad (6)$$

where $\chi_m = \|\mathbf{r}_1 - \mathbf{r}_g\| / \|\mathbf{r}_m - \mathbf{r}_g\|$. Solving (4) using the method of Lagrange multipliers, we obtain

$$\mathbf{w}_c(f, k, \mathbf{r}_g) = \frac{\Phi^{-1}(f, k)\mathbf{a}(f, \mathbf{r}_g)}{\mathbf{a}^H(f, \mathbf{r}_g)\Phi^{-1}(f, k)\mathbf{a}(f, \mathbf{r}_g)} \quad (7)$$

In real applications, the inverse of the cross-spectral density matrix can be calculated using the Moore–Penrose pseudoinverse Φ^+ [21]. Moreover, if Φ is ill-conditioned, the spatial spectrum might be deteriorated by steering vector errors and discrete sampling effects [4]. Diagonal loading (DL) [7] is a regularization technique that mitigates the performance degradations of the MVDR beamformer. The SRP of the beamformer output with MVDR filter and DL becomes

$$P(f, k, \mathbf{r}_g) = \mathbf{w}_c^H(f, k, \mathbf{r}_g)(\Phi(f, k) + \xi\mathbf{I})\mathbf{w}_c(f, k, \mathbf{r}_g) = \frac{1}{\mathbf{a}^H(f, \mathbf{r}_g)(\Phi(f, k) + \xi\mathbf{I})\mathbf{a}(f, \mathbf{r}_g)} \quad (8)$$

where \mathbf{I} is the identity matrix, and the data-dependent DL factor is given by $\xi = \text{tr}[\Phi(f, k)]\Delta/M$, where Δ is the loading constant, and $\text{tr}[\cdot]$ denotes the sum of the elements on the main diagonal of the cross-spectral density matrix.

The power output of the broadband SRP using the NMVDR [27] is given by

$$N(k, \mathbf{r}_g) = \sum_{f=0}^{L-1} \frac{P(f, k, \mathbf{r}_g)}{\max_{\mathbf{r}'_g} [\mathbf{p}_{\mathbf{r}'_g}(f, k)]} \quad (9)$$

where $\mathbf{p}_{\mathbf{r}'_g}(f, k) = [P(f, k, \mathbf{r}'_1), \dots, P(f, k, \mathbf{r}'_g), \dots]$ is the narrowband power map for all candidate positions $\mathbf{r}'_g \in G$ and $\max[\cdot]$ denotes the maximum value.

3. The weighted MVDR beamformer

Both MVDR and NMVDR have the disadvantage that in noisy or reverberant conditions some of the narrowband power maps in the fusion may exhibit an energy peak corresponding to a wrong position in the search space, thus providing a misleading contribution to the fusion map. To avoid using this disruptive information, we introduce the WMVDR, in which the weighting factors are here modeled by an SVM classifier. An important advantage of the SVM with respect to some comparable previous techniques [29,30], is that it requires the identification of a smaller number of parameters and does not relies on any prior information or heuristic assumptions.

The power output of the WMVDR is expressed as

$$U(k, \mathbf{r}_g) = \sum_{f=0}^{L-1} \gamma_f \frac{P(f, k, \mathbf{r}_g)}{\max_{\mathbf{r}'_g} [\mathbf{p}_{\mathbf{r}'_g}(f, k)]} \quad (10)$$

where γ_f are binary variables, which take values 0 or 1. The SVM classifier is trained on known source positions G_t . Given a reference source that is fixed in a training position $\mathbf{r}_t(k) \in G_t$, the estimated source position using the NMVDR beamformer and only the information related to frequency f is

$$\hat{\mathbf{r}}_t(k, f) = \underset{\mathbf{r}'_g}{\text{argmax}} [\mathbf{n}_{\mathbf{r}'_g}(f, k)] \quad (11)$$

145 where $\mathbf{n}_{r'_g}(f, k) = [N(f, k, \mathbf{r}'_1), \dots, N(f, k, \mathbf{r}'_g), \dots]$ is the normalized
146 narrowband power map for all the desired positions $\mathbf{r}'_g \in G$. The
147 contribution to the localization error related to frequency f is

$$\Omega(f, k, \mathbf{r}_t) = \|\mathbf{r}_t(k) - \widehat{\mathbf{r}}_t(f, k)\|. \quad (12)$$

148 The SVM classifier is trained to remove those narrowband compo-
149 nents which contribute negatively to the localization. Namely, the
150 i th training set output $\bar{\gamma}_i$ of the SVM is set as

$$\bar{\gamma}_i = \begin{cases} 0, & \text{if } \Omega(f, k, \mathbf{r}_t) > \eta \\ 1, & \text{if } \Omega(f, k, \mathbf{r}_t) \leq \eta \end{cases} \quad (13)$$

151 where η is a given threshold.

152 We consider three statistical measures of the marginal distribu-
153 tion as possible input features of the classifier: the skewness, the
154 kurtosis, and the crest factor. The use of the proposed statistical
155 features allows the use of a small input vector size, which is di-
156 mensionally independent from the analysis area and from the spa-
157 tial resolution. Theoretically, in free-noise and anechoic conditions
158 the narrowband power map is characterized by a strong impulse
159 peak in the position where the source is active. In real applications,
160 the noise and reverberation modify the response power, hence we
161 use the statistics of the input features along x , y , and z axes for a
162 more robust learning. The marginal distribution of a narrowband
163 power map with the NMVDR along the x -axis is

$$I_f(x) = \int_y \int_z N(f, k, \mathbf{r}_g) dy dz, \quad \forall x \in G_x. \quad (14)$$

164 The marginal distributions along y and z can be derived analog-
165 ously. The skewness is a measure of the symmetry of a distri-
166 bution, and it is defined for a generic distribution $I_f(i)$ as

$$S_i = \frac{E[(I_f(i) - \mu_i)^3]}{(E[(I_f(i) - \mu_i)^2])^{3/2}}, \quad i = x, y, z \quad (15)$$

167 where μ_i is the mean of $I_f(i)$. The kurtosis is a descriptor of the
168 shape of a distribution, and is defined as

$$\kappa_i = \frac{E[(I_f(i) - \mu_i)^4]}{(E[(I_f(i) - \mu_i)^2])^2}, \quad i = x, y, z. \quad (16)$$

169 The skewness and kurtosis are related respectively to the peak po-
170 sition and to the peakedness of a distribution. The crest factor is
171 the ratio of the largest absolute value to the root mean square
172 value of a distribution, and is defined as

$$l_i = \frac{|I_f(i)|_\infty}{\sqrt{\frac{1}{R_i} \sum_{j=1}^{R_i} |I_f(i_j)|^2}}, \quad i = x, y, z \quad (17)$$

173 where R_i is the dimension of $I_f(i)$. The crest factor indicates how
174 extreme the peak is in the narrowband power map. The input vec-
175 tors for the three features are $\mathbf{i}_{sk} = [s_x, s_y, s_z]^T$, $\mathbf{i}_{ku} = [\kappa_x, \kappa_y, \kappa_z]^T$,
176 and $\mathbf{i}_{cf} = [l_x, l_y, l_z]^T$. The input vector for each feature has 3 and
177 2 components for 3D and 2D localization respectively. When all
178 features are considered the input vector takes the following form
179 $\mathbf{i}_a = [\mathbf{i}_{sk}^T, \mathbf{i}_{ku}^T, \mathbf{i}_{cf}^T]^T$.

180 The SVM produces a non-linear classification boundary in the
181 input space by constructing a linear hyperplane in a transformed
182 version of the input space [6]. The SVM supervised model is then
183 defined as

$$\gamma'_f = \text{sgn}\left(\sum_{i=0}^Q \alpha_i \bar{\gamma}'_i \psi(\bar{\mathbf{i}}_i, \mathbf{i}(f)) + b\right) \quad (18)$$

184 where γ'_f takes values $\{1, -1\}$, Q is the training sample size,
185 $\psi(\bar{\mathbf{i}}_i, \mathbf{i}(f))$ is the inner-product kernel for the i th training input
186 vector $\bar{\mathbf{i}}_i$ and the input vector $\mathbf{i}(f)$ for the narrowband power map
187 at frequency f , $\bar{\gamma}'_i$ is the i th target value, computed as $\bar{\gamma}'_i = \bar{\gamma}_i(\bar{\gamma}_i +$

188 $1) - 1$ so that it takes values $\{1, -1\}$, $\alpha_i \geq 0$, and b is a real con-
189 stant. The weighting factors are transformed with $\gamma_f = (\gamma'_f + 1)/2$
190 to obtain values $\{1, 0\}$. The inner-product is used to construct the
191 optimal hyperplane in the feature space. Common types of inner-
192 product kernels are: linear, quadratic, polynomial, radial basis func-
193 tion (RBF), multilayer perceptron (MLP). The parameter α_i can be
194 found by solving the following convex maximization quadratic pro-
195 gramming problem

$$\begin{aligned} \max \quad & \sum_{i=0}^Q \alpha_i - \frac{1}{2} \sum_{i,j=0}^Q \alpha_i \alpha_j \bar{\gamma}'_i \bar{\gamma}'_j \psi(\bar{\mathbf{i}}_i, \bar{\mathbf{i}}_j) \\ \text{subject to} \quad & \sum_{i=0}^Q \alpha_i \bar{\gamma}'_i = 0, \quad 0 \leq \alpha_i \leq \lambda, \quad i = 1, 2, \dots, Q \end{aligned} \quad (19)$$

196 where λ is a user specified parameter and provides a trade-off
197 between the distance of the support vectors from the separating
198 margin and the training error. In this paper, we use the sequential
199 minimal optimization [22] algorithm for solving Eq. (19). By taking
200 any support vector $\bar{\mathbf{i}}_j$ with $\alpha_i < \lambda$, the parameter b can be calcu-
201 lated by

$$b = \bar{\gamma}'_j - \sum_{i=0}^Q \alpha_i \bar{\gamma}'_i \psi(\bar{\mathbf{i}}_i, \bar{\mathbf{i}}_j). \quad (20)$$

202 Finally, the SRP with the WMVDR-SVM filter can be written as

$$U(k, \mathbf{r}_g) = \sum_{f=0}^{L-1} \left(\frac{\gamma'_f + 1}{2} \right) \frac{P(f, k, \mathbf{r}_g)}{\max_{\mathbf{r}'_g} [P_{\mathbf{r}'_g}(f, k)]} \quad (21)$$

204 where γ'_f is estimated using Eq. (18). The sound source localization
205 is estimated using the WMVDR beamformer by picking the maxi-
206 mum value on the fusion map

$$\widehat{\mathbf{r}}_s(k) = \underset{\mathbf{r}'_g}{\text{argmax}} [\mathbf{u}_{\mathbf{r}'_g}(k)] \quad (22)$$

207 where $\mathbf{u}_{\mathbf{r}'_g}(k) = [U(k, \mathbf{r}'_1), U(k, \mathbf{r}'_2), \dots, U(k, \mathbf{r}'_g), \dots]$ is the power
208 map for all the searching positions $\mathbf{r}'_g \in G$.

209 The major computational demand of MVDR comes from the
210 matrix inversion operation of the Hermitian matrix Φ , which
211 requires $\mathcal{O}(M^3)$ flops for each narrowband beamforming. Our
212 WMVDR requires an additional cost for the SVM classification,
213 which depends on the kernel function used. A SVM with the RBF
214 kernel has $\mathcal{O}(Q_{SV}d)$ complexity, where Q_{SV} is the number of sup-
215 port vectors and d is the dimension of the input vector, while a
216 SVM with the linear kernel complexity is $\mathcal{O}(d)$. The complexity of
217 WMVDR with the SVM-RBF is thus $\mathcal{O}(LM^3 + Q_{SV}d)$, and hence the
218 additional cost is related to the number of support vectors Q_{SV} . In
219 our experiments, in which the audio buffer size was $L = 2048$, the
220 array had 8 microphones and the Q_{SV} was on average of around
221 6000 kernels, the localization algorithm was run in real time on
222 standard computer systems equipped with a i5-i7 CPU and 8GB
223 RAM.

4. Experiments

225 In this section, a performance analysis of 2D sound source
226 localization in simulated and real reverberant rooms is reported.
227 In all experiments, the SVM was trained with an USASI noise
228 signal, which roughly simulates the energy distribution of speech
229 and music. We compare the localization performance of sound
230 signals using the root mean square error (RMSE) of the proposed
231 WMVDR-SVM, the WMDR-RBFN [28], the NMVDR [27], the MVDR
232 [3], and the SRP-PHAT [10]. The SVM and RBFN have been imple-
233 mented using the Matlab Statistics and Machine Learning Toolbox,

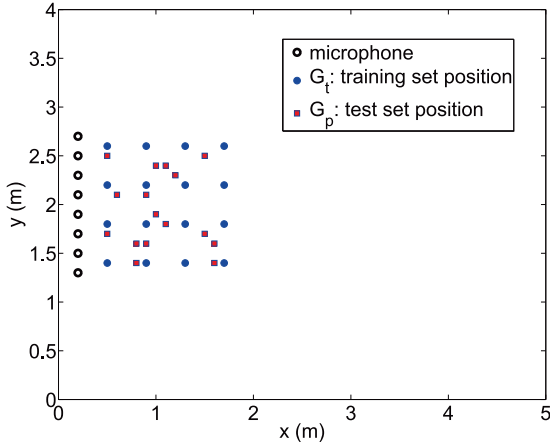


Fig. 1. The simulated room setup with the G_t and G_p positions for the training and the testing phase respectively.

Table 1

Percentage (%) of rejected maps at variation of RT_{60} using an USASI noise signal during the training phase.

RT_{60} (s)	0.2	0.3	0.4	0.5	0.6	0.7	0.8
(%)	48.19	52.14	58.47	63.11	65.35	67.89	68.11

234 whereas for the MVDR filter we used our own implementation.
235 The RMSE is calculated in the following way

$$RMSE = \sqrt{\frac{\sum_{i=1}^B [(x_s - \hat{x}_s(i))^2 + (y_s - \hat{y}_s(i))^2]}{B}} \quad (23)$$

236 where B is the total number of analysis blocks, and $[\hat{x}_s(i), \hat{y}_s(i)]$ are
237 the estimated Cartesian coordinates of the source for the analysis
238 block i .

239 4.1. Synthetic data

240 A reverberant room of 5 m × 4 m × 3 m simulated with the
241 image-source model [17] was used. Several Monte Carlo experi-
242 ments were performed. The room setup is shown in Fig. 1 con-
243 sidering a uniform linear array (ULA) of 8 microphones and two
244 set of source positions: the training set positions (G_t) and the test
245 set positions (G_p). The distance between microphones in the ULA
246 was 0.2 m and the spatial resolution of the 2D grid for searching
247 the source was 0.1 m. The sampling frequency was 44.1 kHz and
248 the window size L was 2048 samples. Thus, the bandwidth of each
249 narrowband frequency bin was 21.53 Hz. For all MVDR-based al-
250 gorithms the loading constant Δ was set to 0.001. The tests were
251 conducted with different averaging SNR levels, obtained by adding
252 mutually independent white Gaussian noise (WGN) to each chan-
253 nel.

254 4.1.1. Training phase

255 An USASI noise signal was used as source for the SVM learn-
256 ing in the training set positions with a reverberation time (RT_{60})
257 of 0.5 s, a SNR of 20 dB, and a parameter η of 0.5 m. This choice is
258 motivated by the analysis, during the training phase, of the num-
259 ber of constructively and disruptively contributing maps depicted
260 in Table 1, since the goal is to have the same number of correct
261 and reject maps as much as possible, but considering however a
262 significant level of reverberation. A RT_{60} of 0.5 s is a good com-
263 promise. This room condition was used in all simulated experiments
264 for the SVM learning.

Table 2

Classification error (%) when performing the validation on the training data set with different inner-product kernels.

	RBF	Linear	MLP	Polynomial	Quadratic
Correct Map	33.07	37.67	50.99	71.43	77.68
Reject Map	53.18	61.77	51.09	20.98	16.74

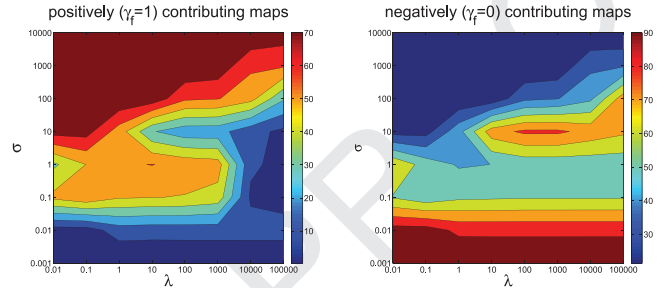


Fig. 2. Cross-validation rate (%) of constructively contributing maps and disruptively contributing ones.

4.1.2. Testing phase

265 In the first experiment, the SVM training using the skewness
266 measure was performed on USASI sound sources positioned in the
267 training set positions with different inner-product kernels. We can
268 observe the results in Table 2, which report the percentage of
269 classification error for the positively ($\gamma_f = 1$) and for the nega-
270 tively ($\gamma_f = 0$) contributing maps. The RBF kernel provides the
271 best performance since it provides the lower error for the cor-
272 rect maps, which is clearly the primary goal when attempting at
273 selecting only the correct information, and this confirms the rea-
274 sonable first choice of this kernel when the relation between class
275 labels and features is nonlinear [14]. The scaling factor σ of RBF
276 was set to 1 as in [28]. The RBF kernel was thus adopted for the
277 SVM classifier in subsequent tests. Next, an analysis of the SVM
278 and RBF parameters was conducted. The results of a grid-search
279 procedure on λ and σ using a cross-validation [14] is shown in
280 Fig. 2. High cross-validation rate of positively contributing maps
281 corresponds to low cross-validation rate of negatively contribut-
282 ing ones, and vice versa. An optimal set of parameters is thus
283 obtained by balancing the two contributes and by setting $\sigma = 1$
284 and $\lambda = [1, 10, 100, 1000]$. Hence, a good choice is given by setting
285 $\sigma = 1$ and $\lambda = 1$ with a cross-validation rate of 58,38% ($\gamma_f = 1$)
286 and 48,94% ($\gamma_f = 0$). This setup set was used in the localization
287 performance tests.

288 A set of experiments were then conducted for evaluating dif-
289 ferent operating conditions. The performance was evaluated with
290 a male and a female speech signal from the TSP Speech Database¹.
291 An optimal frequency range between 80 Hz and 8000 Hz, since it
292 is a typical spectrum range of speech signals, was used. Table 3
293 shows the localization performance using as input the skewness,
294 the kurtosis, the crest factor, all the three features together,
295 and the marginal distribution as in [28]. The WMVDR-RBFN using
296 the skewness measure (WMVDR-RBFN-S) was also considered. We
297 can see that the best performance is achieved by the WMVDR-SVM-
298 RBF using the skewness measure, which was adopted in subse-
299 quent tests. A localization evaluation for different reverberant
300 conditions and a SNR of 20 dB is showed in Figs. 3 and 4. As we
301 can observe, the WMVDR-SVM-RBF outperforms all other algorithms.
302 Then, Figs. 5 and 6 show the RMSE using spatially white noise
303 conditions with a RT_{60} of 0.3 s. We note the reduction of local-
304 ization performance of the WMVDR-SVM-RBF in a low noise con-
305 dition in which the RMSE tends to that of the NMVDR. Furthermore, we
306 can

¹ <http://www-mmsep.ece.mcgill.ca/Documents/Data>.

Table 3
RMSE (m) of localization performance with synthetic data and a RT_{60} of 0.5-s, a SNR of 20 dB, and a η of 0.5 m.

	WMVDR-SVM-RBF				
	Crest Factor	Kurtosis	Skewness	Cr. F.+Kurt.+Sk.	Marg. Distr.
Male Speech G_t	1.751	0.886	0.559	0.718	0.613
Female Speech G_t	1.921	1.012	0.875	0.952	0.882
Male Speech G_p	1.587	0.583	0.404	0.487	0.470
Female Speech G_p	1.611	0.999	0.755	0.788	0.782
	WMVDR-RBFN	WMVDR-RBFN-S	NMVDR	MVDR	SRP-PHAT
Male Speech G_t	0.816	1.154	1.063	1.572	1.020
Female Speech G_t	0.988	1.342	1.533	1.836	1.530
Male Speech G_p	0.760	1.067	1.123	2.003	1.127
Female Speech G_p	0.869	1.286	1.483	1.729	1.508

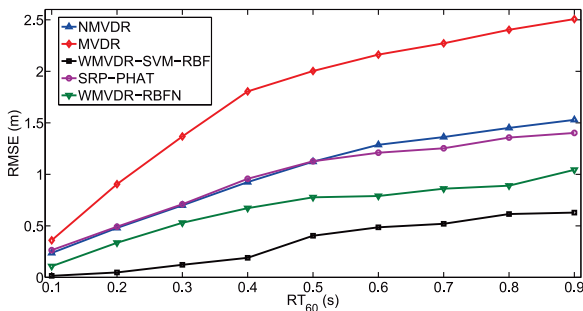


Fig. 3. Localization performance of a male speech in G_p and SNR = 20 dB.

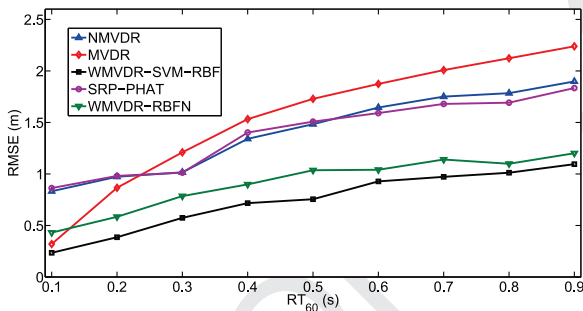


Fig. 4. Localization performance of a female speech in G_p and SNR = 20 dB.

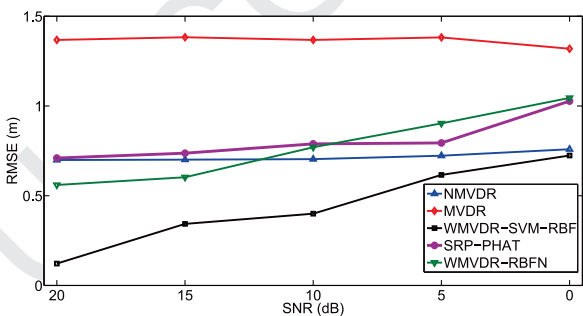


Fig. 5. Spatially white noise: localization performance of a male speech in G_p and $RT_{60} = 0.3$ -s.

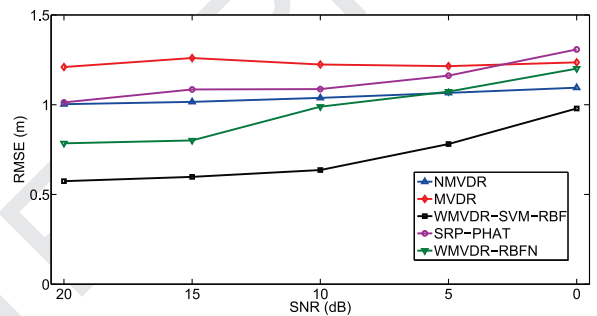


Fig. 6. Spatially white noise: localization performance of a female speech in G_p and $RT_{60} = 0.3$ -s.

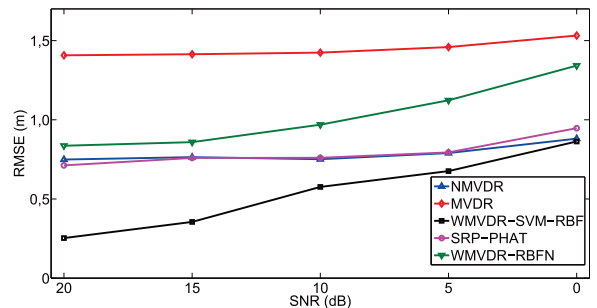


Fig. 7. Diffuse noise: localization performance of a male speech in G_p and $RT_{60} = 0.3$ -s.

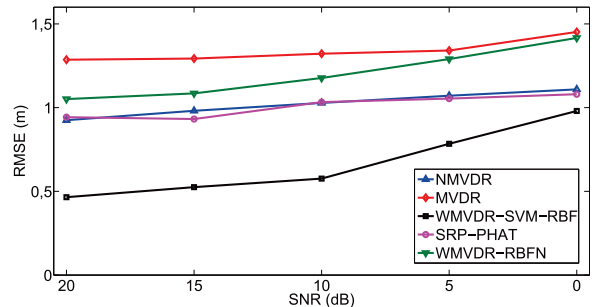


Fig. 8. Diffuse noise: localization performance of a female speech in G_p and $RT_{60} = 0.3$ -s.

307 observe in Figs. 7 and 8 a good performance of the WMVDR-SVM-
 308 RBF in a diffuse noise field [13] with a SNR range of 5–20 dB and
 309 a RT_{60} of 0.3 s. The localization performance and a rejection map
 310 analysis during the training phase of the WMVDR-SVM-RBF with
 311 respect to parameter η is showed in Fig. 9. We observe a similar
 312 RMSE performance for η in the range [0.3; 0.9].

313 Next two experiments were conducted to evaluate the impact
 314 of the room size and of the array geometry on the algorithm

performance. The SNR and the RT_{60} were set to 20 dB and 0.5 s
 respectively. Specifically, a localization evaluation of a male speech
 signal in the test position with a smaller (3.5 m \times 3 m \times 2.5 m)
 and a larger (10 m \times 8 m \times 4 m) room, with respect to the
 trained room, is depicted in Table 4. We can see that the pro-
 posed WMVDR-SVM-RBF can be used in different room sizes. In
 particular, when the size of a room is larger we have a better

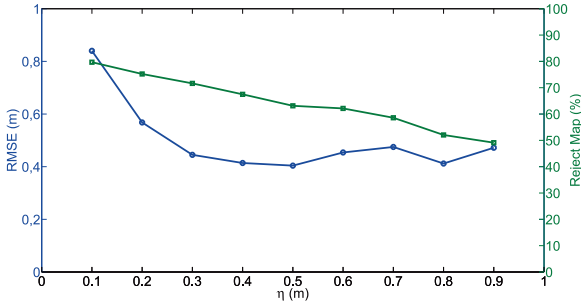


Fig. 9. Round marker: localization performance of a male speech in G_p with a SNR=20 dB and a RT_{60} =0.5-s. Square marker: rejection map during the training phase.

Table 4 RMSE (m) of localization performance with synthetic data for different room sizes using a male speech signal in the test position G_p .

Room Size	WMVDR-SVM-RBF	WMVDR-RBFN-S	NMVDR	MVDR	SRP-PHAT
(3.5 m × 3 m × 2.5 m)	0.549	1.070	0.877	1.273	0.847
(10 m × 8 m × 4 m)	0.317	1.556	1.458	0.909	1.598

Table 5 RMSE (m) of localization performance with synthetic data for randomly array geometries using a male speech signal.

WMVDR-SVM-RBF	WMVDR-RBFN-S	NMVDR	MVDR	SRP-PHAT
1.466	2.195	1.342	1.715	1.401

Table 6 RMSE (m) of localization performance with different synthetic data in the test position G_p and a frequency range of analysis of 80-16,000-Hz.

	WMVDR-SVM-RBF	WMVDR-RBFN-S	NMVDR	MVDR	SRP-PHAT
Male Speech	0.483	1.118	0.782	1.394	0.774
Gunshot	0.018	1.077	0.265	2.371	0.265
Flute	0.573	1.294	0.775	0.822	0.962

performance due to the reduction of early reflection energy at microphones. This fact is also observable as a better localization of the MVDR if compared to that of NMVDR and of SRP-PHAT. In this case, the normalization that improves the SRP resolution provides a smaller advantage in the identification of direct path and reflections. On the contrary, when the room size decreases, the energy of reflections at microphones is larger and the localization performance decreases as a consequence. Table 5 shows instead the results with different array geometries. In this experiment, the microphones and the source were randomly located with a uniform distribution in each trial so that the minimum distance between walls and microphones was 0.1 m. We have considered the same room of the training phase that is depicted in Fig. 1. As we can observe in Table 5, the SVM classifier is not able to improve the localization performance.

Finally, an analysis of the generalization with respect to the acoustic characteristics of the source was conducted. In this experiment, a RT_{60} of 0.3-s, a spatially white noise of 15 dB, and a frequency range between 80-Hz and 16,000-Hz for computing the beamforming were used. We have used a male speech signal, an impulsive gunshot signal, and a flute musical instrument signal². Table 6 shows the improvement of localization performance with different target sound signals for the proposed WMVDR-SVM-RBF.

² <http://theremin.music.uiowa.edu/>

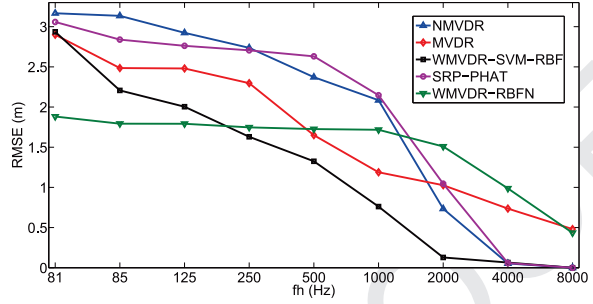


Fig. 10. Localization performance with variable bandwidth $[f_l, f_h]$ of a WGN signal. The SNR was 10 dB, the $RT_{60} = 0.5$ -s, and f_l was set to 80-Hz.

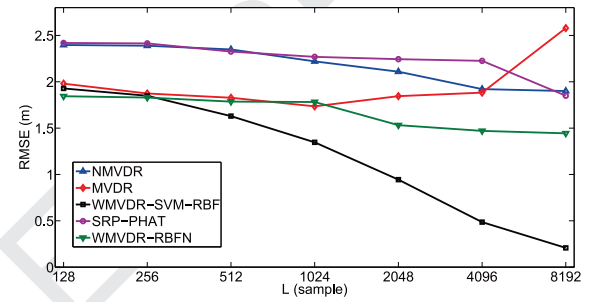


Fig. 11. Localization performance with variable FFT size L of a WGN signal with a bandwidth [80,1000] Hz. The SNR was 5 dB and the $RT_{60} = 0.6$ -s.

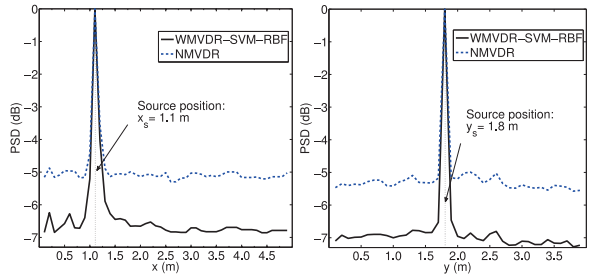


Fig. 12. The PSD of the NMVDR and the proposed WMVDR-SVM-RBF.

We note a minor improvement for the flute signal, due to its harmonic spectrum. Next, Fig. 10 shows the performance in relation to signal bandwidth. The source signal in this case was obtained by processing a WGN with a bandpass filter $H_{[f_l, f_h]}$, where f_l and f_h are the lower and upper frequency limit respectively. The experiments were conducted by using $f_l = 80$ Hz as lower limit and different upper limit frequencies ranging from 81-Hz to 8000-Hz. We note that when the signal becomes narrowband the WMVDR-SVM-RBF performance degrades to that of MVDR. We can observe also the noise emphasis problem due to the normalization with narrowband sources for the SRP-PHAT and the NMVDR [12,27]. Last, Fig. 11 shows the performance for a WGN signal with a frequency band of [80, 1000] Hz, at variation of the block size L for computing the fast Fourier transform (FFT). The WMVDR-SVM-RBF performs better when the spectral resolution is increased, and when $L < 1024$ samples the SVM is ineffective due to the reduced number of narrowband bins. Fig. 12 depicts a PSD along x and y axis of the proposed method and of the NMVDR for a speech signal in a free-noise anechoic condition. We can see the effect of removing the incorrect narrowband power maps keeping a high resolution on the source position and providing a larger attenuation of the power in the other directions.

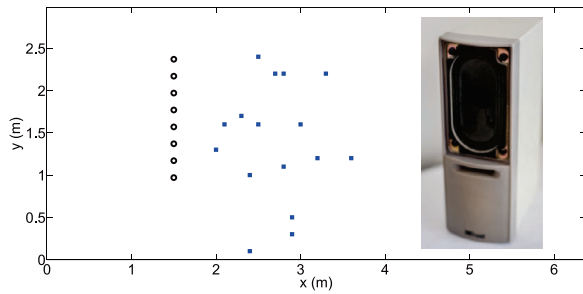


Fig. 13. The real room setup and the loudspeaker used in experiments.

Table 7

RMSE (m) of localization performance with real data.

L (sample)	WMVDR-SVM-RBF	WMVDR-RBFN-S	NMVDR	MVDR	SRP-PHAT
2048	0.896	1.208	1.279	1.610	1.169
8192	0.087	0.354	0.323	1.436	0.242

4.2. Real data

The experiments were performed in a room with a RT_{60} of 0.6 s and setup as in Fig. 13. The distance between microphones was 0.2 m, the sampling frequency was 44.1 kHz, and the window size L was 2048 and 8192 samples. Sixteen source positions have been considered. A speech signal from a male speaker was reproduced with a loudspeaker placed in each position. The loudspeaker, depicted in Fig. 13, has a small oval driver with a size of 9.5 cm \times 5 cm, a frequency response of 90–20,000 Hz, and a RMS power of 1 W. In each test position, the loudspeaker was directed toward the center of the array. An USASI noise signal was used as source for the SVM learning in the training set positions of the simulated room setup depicted in Fig. 1 with a RT_{60} of 0.5 s, a SNR of 20 dB, and the parameter η set to 0.5 m. Since the training room and the real testing room have different size, the WMVDR-RBFN-S was used. The results reported in Table 7 confirm the improvement of the localization accuracy for the proposed WMVDR-SVM using an USASI noise signal for a SVM training phase in a simulated reverberant environment.

5. Conclusions

A WMVDR beamformer based on a SVM classifier with a RBF kernel has been presented. It improves the localization accuracy in a single source scenario without point-source interferences by using the skewness measure of marginal distributions and by selecting only the narrowband power maps that positively contribute to the broadband fusion. We showed that a training phase using an USASI noise signal in a simulated room allows the machine learning to select the useful acoustic information and to discard the corrupted information with different sound signals, room sizes, and room conditions both with synthetic and real data. We showed that improved performance is achieved for different reverberant conditions and a SNR up to 5 dB. When using the SVM learning in the NMVDR algorithm, however, it is required that the geometry of the array is kept similar in the training and in the testing or operating phase, and that a sufficiently high frequency resolution is used in the FFT analysis step.

References

- [1] M.R. Azimi-Sadjadi, A. Pezeshki, N. Roseveare, Wideband DOA estimation algorithms for multiple moving sources using unattended acoustic sensors, *IEEE Trans. Aerosp. Electron. Syst.* 44 (4) (2008) 1585–1599.

- [2] D.A. Blauth, V.P. Minotto, C.R. Jung, B. Lee, T. Kalker, Voice activity detection and speaker localization using audiovisual cues, *Pattern Recognit. Lett.* 33 (4) (2012) 373–380.
- [3] J. Capon, High resolution frequency-wavenumber spectrum analysis, *Proc. IEEE* 57 (8) (1969) 1408–1418.
- [4] Y.L. Chen, J.-H. Lee, Finite data performance analysis of MVDR antenna array beamformers with diagonal loading, *Prog. Electromagn. Res.* 134 (2013) 475–507.
- [5] Y. Cho, D. Yook, S.C.H. Kim, Sound source localization for robot auditory systems, *IEEE Trans. Consum. Electron.* 55 (3) (2009) 1663–1668.
- [6] C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.* 20 (3) (1995) 273–297.
- [7] H. Cox, R. Zeskind, M. Owen, Robust adaptive beamforming, *IEEE Trans. Acoust. Speech Signal Process.* 35 (10) (1987) 1365–1376.
- [8] A. Czyzewski, Automatic identification of sound source position employing neural networks and rough sets, *Pattern Recognit. Lett.* 24 (6) (2003) 921–933.
- [9] E. Di Claudio, R. Parisi, WAVES: weighted average of signal subspaces for robust wideband direction finding, *IEEE Trans. Signal Process.* 49 (10) (2001) 2179–2190.
- [10] J.H. DiBiase, H.F. Silverman, M.S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer.
- [11] S. Doclo, M. Moonen, T.V. den Bogaert, J. Wouters, Reduced bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids, *IEEE Trans. Audio Speech Lang. Process.* 17 (1) (2009) 38–51.
- [12] K.D. Donohue, J. Hannemann, H.G. Dietz, Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments, *Signal Process.* 87 (7) (2007) 1677–1691.
- [13] E.A.P. Habets, S. Gannot, Generating sensor signals in isotropic noise fields, *J. Acoust. Soc. Am.* 122 (6) (2007) 3464–3470.
- [14] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *A Practical Guide to Support Vector Classification*, Technical Report, National Taiwan University, 2003.
- [15] H. Kayser, J. Anemuller, A discriminative learning approach to probabilistic acoustic source localization, in: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2014, pp. 99–103.
- [16] B. Laufer, R. Talmon, S. Gannot, Relative transfer function modeling for supervised source localization, in: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.
- [17] E. Lehmann, A. Johansson, Prediction of energy decay in room impulse responses simulated with an image-source model, *J. Acoust. Soc. Am.* 124 (1) (2008) 269–277.
- [18] Y. Li, K.C. Ho, M. Popescu, A microphone array system for automatic fall detection, *IEEE Trans. Biomed. Eng.* 59 (5) (2012) 1291–1301.
- [19] D.J. Mennill, M. Battiston, D.R. Wilson, J.R. Foote, S.M. Doucet, Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour, *Methods Ecol. Evol.* 3 (4) (2012) 704–712.
- [20] T. Nishino, K. Takeda, Binaural sound localization for untrained directions based on a gaussian mixture model, in: *Proceedings of the European Signal Processing Conference*, 2008, pp. 1–5.
- [21] C. Pan, J. Chen, J. Benesty, Performance study of the MVDR beamformer as a function of the source incidence angle, *IEEE/ACM Trans. Audio Speech Lang. Process.* 22 (2014) 67–79.
- [22] J.C. Platt, *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Technical Report, Microsoft Research, 1998.
- [23] F. Ribeiro, C. Zhang, D.A. Florencio, D.E. Ba, Using reverberation to improve range and elevation discrimination for small array sound source localization, *IEEE Trans. Acoust. Speech Signal Process.* 18 (7) (2010) 1781–1792.
- [24] C.R. Rowell, D. Fee, C.A.S.K. Arnoult, R.S. Matoza, P.P. Firstov, K. Kim, E. Makhmudov, Three-dimensional volcano-acoustic source localization at Karymsky Volcano, Kamchatka, Russia, *J. Volcanol. Geotherm. Res.* 283 (2014) 101–115.
- [25] D. Salvati, S. Canazza, Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments, *IEEE Signal Process. Lett.* 20 (6) (2013) 507–510.
- [26] D. Salvati, S. Canazza, Incident signal power comparison for localization of concurrent multiple acoustic sources, *Sci. World J.* 2014 (2014) 1–13.
- [27] D. Salvati, C. Drioli, G.L. Foresti, Incoherent frequency fusion for broadband steered response power algorithms in noisy environments, *IEEE Signal Process. Lett.* 21 (5) (2014) 581–585.
- [28] D. Salvati, C. Drioli, G.L. Foresti, Frequency map selection using a RBFN-based classifier in the MVDR beamformer for speaker localization in reverberant rooms, in: *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 3298–3301.
- [29] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, Comparing support vector machines with gaussian kernels to radial basis function classifiers, *IEEE Trans. Signal Process.* 45 (11) (1997) 2758–2765.
- [30] D.J. Sebal, J.A. Bucklew, Support vector machine techniques for nonlinear equalization, *IEEE Trans. Signal Process.* 48 (11) (2000) 3217–3226.
- [31] H. Wang, M. Kaveh, Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources, *IEEE Trans. Acoust. Speech Signal Process.* 33 (4) (1985) 823–831.
- [32] Y. Yoon, L.M. Kaplan, J.H. McClellan, TOPS: new DOA estimator for wideband signals, *IEEE Trans. Signal Process.* 54 (6) (2006) 791–802.