ELSEVIER

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Robotics and Autonomous Systems 42 (2003) 69–88

Robotics and Autonomous Systems

www.elsevier.com/locate/robot

Algorithms for acoustic localization based on microphone array in service robotics

Enzo Mumolo[a,*], Massimiliano Nolich[a], Gianni Vercelli[b]

[a] DEEI, Università di Trieste, Via Valerio 10, 34127 Trieste, Italy
[b] DISA, Università di Genova, Corso Montegrappa 39, 16137 Genova, Italy

Received 8 September 2000; received in revised form 12 August 2002
Communicated by F.C.A. Groen

Abstract

This paper deals with the development of acoustic source localization algorithms for service robots working in real conditions. One of the main utilizations of these algorithms in a mobile robot is that the robot can localize a human operator and eventually interact with him/herself by means of verbal commands. The location of a speaking operator is detected with a microphone array based algorithm; localization information is passed to a navigation module which sets up a navigation mission using knowledge of the environment map. In fact, the system we have developed aims at integrating acoustic, odometric and collision sensors with the mobile robot control architecture. Good performance with real acoustic data have been obtained using neural network approach with spectral subtraction and a noise robust voice activity detector. The experiments show that the average absolute localization error is about 40 cm at 0 dB and about 10 cm at 10 dB of SNR for the named localization. Experimental results describing mobile robot performance in a talker following task are reported.

Keywords: Service robotics; Acoustic localization; Microphone array; Spectral subtraction; Neural network

1. Introduction

Nowadays the role of acoustic perception in autonomous robots and intelligent building applications is increasingly important. In the context of Service Robotics, the request to integrate acoustic devices into mobile robotized systems is growing, not only for developing smart voice-activated user interfaces, but also to perform special tasks such as source localization/tracking and acoustic-visual coordination for target localization in security tasks. For example, in [24] the position of a speaker is detected from him/her voice for teleconference applications, while in [18] a robot capable to detect acoustic source positions and to move towards an acoustic source performing obstacle avoidance is presented. Acoustic perception is an important way to acquire environmental information when integrated and coordinated with other perceptual systems (ultrasonic range systems, cameras, proximity sensors, etc.). The main advantage is, of course, the possibility of greater interaction with humans through the vocal channel; moreover, localizing a human speaker enables the robot to plan a navigation mission taking into account the presence of the human operator, in addition to the other objects in the working environment. Using speaker verification algorithms, other

[*] Corresponding author. Tel.: +39-40-676-3861;
fax: +39-40-676-3460.
E-mail addresses: mumolo@univ.trieste.it, mumolo@units.it (E. Mumolo), mnolich@units.it (M. Nolich), vercelli@unige.it (G. Vercelli).

0921-8890/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0921-8890(02)00325-1

functions can be devised besides the orientation, tracking [29] and approaching maneuvers. Most of the work done so far use sound localization algorithm separately from other sensing systems; recently some attempts towards the integration of different sensory systems (e.g. audio-visual) have reached significant results [1].

In this paper, we present a neural network algorithm to perform localization of an acoustic source in a indoor environment for mobile robots.[1] This work starts from an analysis of the available relevant localization approaches for broadband acoustic signals. The approaches presented in [7,19,22,25] are based on microphone arrays and compute the acoustic source position by triangulation of some estimations of the direction of arrival of the acoustic wave. Such estimations are obtained from the phase differences between couples of microphones of the array. Since our service robot moves in real indoor environments, with an high degree of noise, one objective of this study is to work-out a practical algorithm (from a real-time point of view) with high performance at low SNR; for this reason a particular attention in the algorithm development was given to guarantee high noise robustness. The algorithm described in this paper has been first simulated and then implemented on a system formed by an experimental mobile robot (called Snoopy) equipped for indoor navigation in unstructured environments, a computer responsible for the microphone array acquisition and processing (called Woodstock), and a supervision host. All three elements communicate by means of TCP/IP based protocols and Wavelan Ethernet link, while the internal communication within Snoopy is based on the Echelon field bus. The acquisition of the signals coming from the microphone array is performed by a dedicated board, DSP based, specifically designed for this project and installed on the ISA bus on Woodstock. The experimental results described in the following have been obtained by algorithms simulations and a voice database acquired in the environment chosen for the experiments.

This work has been motivated by a wider project in service robotics: the main goals of this project were the definition of a cognitive architecture for Service Robots and the realization of an advanced pre-competitive mobile robot to be used in surveillance as well as for transportation tasks. Several cognitive architectures have been considered, aiming at overtaking the hierarchical model used in industrial robots, by means of agent-based models which implement senso-motorial behavior in a hybrid integrated system. Many important reference architectures related to such kind of robots are currently adopted (Subsumption Architecture [9], 3T [5], AUra [2], etc.), where the attention is more focused on cognitive aspects. Since the task performed by Snoopy is to track and follow the vocal commands spoken by a human operator, its cognitive model is reactive. That is, the mobile robot is moved according to a purely reactive action, which is a programmed reaction to sensor data. The sensors used in this case are microphones connected in an array configuration, bumpers which are used to detect collisions and odometry which track the current position of the robot. Hence, a major part of the system is devoted to the integration of the sensor data with the mobile robot reactive agents. Execution of reactive agents must be performed in real-time, so time control is necessary. A real-time scheduler is thus a fundamental component of the programming environment, providing both execution and management of periodic and aperiodic tasks. Concurrent agents run on a distributed platform and communication among them is performed transparently with respect to the network, which is based upon different media, i.e. cable and wireless. The movement of Snoopy from its current position to an acoustic target is performed using the ξ-model [27]. The ξ-model is a biologically inspired non-linear model of trajectory generation algorithm, which allows the expert "ξ-trajectory Generator" to generate a smooth real-time trajectory from the robot's current position–orientation pair (p_R, θ_R) to a given target (p_T, θ_T). The main purpose of this model is to overcome the instability problem in motion generation when the robot must react to a rapidly changing environment. The practical effect of the ξ-model is shown in Fig. 17.

The paper is organized as follows: Section 2 describes the main research performed so far for localization. Section 3 describes the proposed acoustic localization algorithm, based on spectral subtraction and neural networks and some experimental results at various level of SNRs are reported in

[1] The system architecture of the mobile robot here described has been developed at the University of Trieste within the National Project CERTAMEN [17].

Section 5. Conclusions and final remarks are reported in Section 6.

2. Related work in acoustic localization algorithms

Acoustic source localization in indoor environments, which is of interest in service robotics, is classically performed estimating the time delay of arrival (TDOA) between couples of acoustic sensors, also known as interaural time difference (ITD). The acoustic sensing subsystem we use on our service robot is a linear microphone array, which is a widely used technology adopted to solve the direction finding problem.

In the comparison of the considered algorithms, several factors are taken into account, such as the propagation speed of the sound waves, the positions of the microphones and the geometric characteristics of the environment.

2.1. *The direction finding problem*

Let us consider a simple set-up composed of an acoustic source and an array of microphones, as depicted in Fig. 1.

If the acoustic source produces $r(t)$, the signal acquired by the ith microphone can be modeled as follows:

$$s_i(t) = \alpha_i r(t - \tau_i) + n_i(t), \tag{1}$$

where α_i is an attenuation factor due to the propagation of the acoustic signal travelling from the source to the microphone, τ_i is the propagation time and $n_i(t)$ includes all the noise present in the room, including reverberation and multiple echoes. Moreover, let us call d_{ij} the distance between the i and j microphone pair and $\delta_{ij} = \tau_j - \tau_i$ is the delay between the arrival time of the acoustic wavefront to the i and j microphones. Under simplifying assumptions—basically a far-field hypothesis, i.e. $S_{ri}, S_{rj} \gg d_{ij} \gg \lambda$, where S_{ri}, S_{rj} are the distances between the source r and the microphones i and j, respectively, and λ is the predominant wavelength of the signal—it is easy to determine from the time delay the angles Ψ of arrival of the acoustic source. This is described in Fig. 1, where from $\Delta_s = d_{ij}\cos(\theta) \simeq S_{rj} - S_{ri} = v(\tau_j - \tau_i) = v\delta_{ij}$, v being the speed of the sound, it is clearly possible to estimate the θ angle. Using simple geometric considerations it is then possible to estimate the (x, y) coordinates of the acoustic source, as described in Section 2.2.2.

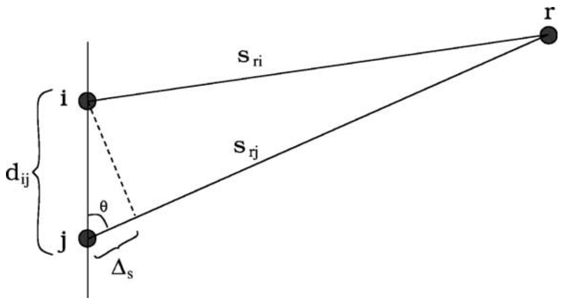

Fig. 1. Geometry for the direction finding problem.

The techniques considered in this work for estimating the time delay between the signals received by the microphones are suitable for broadband acoustic signal, in particular, for speech signals. In the following sections some known techniques are briefly described, namely:

- cross-correlation based approaches;
- adaptive filtering approaches;
- neural approaches.

2.2. *Acoustic localization using cross-correlation*

These techniques compute the TDOA between couples of microphones, and then the position of the acoustic source is estimated from it.

2.2.1. *Time delay estimation using cross-correlation*

Basically, the techniques for the time delay estimation are based on the maximization of the cross-correlation between a couple of signals. Thus, the simplest way is to find the delay τ which maximizes the cross-correlation function

$$R_{ik}(\tau) = E\{s_i(t)s_k(t + \tau)\}, \tag{2}$$

which, giving the model (2), becomes

$$R_{ik}(\tau) = \alpha_i\alpha_k R_{rr}(\tau - \delta_{ik}) + R_{n_i n_k}(\tau), \tag{3}$$

where R_{rr} is the autocorrelation of the noise source $r(t)$. However, the formulation of (3) is quite sensitive to the noise, and further improvements are needed

[22]. The first idea is to introduce some kind of filtering. The simplest approach is to normalize the cross-correlation with respect to the signal energy, which leads to the Normalized Cross-Correlation method (NCC) [28]:

$$R_{ik}(\tau) = \frac{\int_{-\infty}^{+\infty} s_i(u)s_k(u+\tau)\,\mathrm{d}u}{\sqrt{\int_{-\infty}^{+\infty} s_i^2(u)\,\mathrm{d}u}\sqrt{\int_{-\infty}^{+\infty} s_k^2(u)\,\mathrm{d}u}}. \tag{4}$$

Better performance can be obtained by filtering in the spectral domain. More precisely, a spectral weighting filter $\psi(f)$ [19] can be introduced as follows:

$$R_{ik}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f)G_{ik}(f)\,\mathrm{e}^{\mathrm{j}2\pi f\tau}\,\mathrm{d}f. \tag{5}$$

The function reported in (5) is called Generalized Cross-Correlation (GCC). Various choices of the weighting function are possible. For instance, the $\psi(f)$ function can be derived with a Maximum Likelihood formulation leading to the TDOA algorithm as described in [8], called in the following AML (Approximated Maximum Likelihood):

$$\hat{\psi}_{\mathrm{ML}} = \frac{|S_1(\omega)|\cdot|S_2(\omega)|}{|N_1(\omega)|^2\cdot|S_2(\omega)|^2+|N_2(\omega)|^2\cdot|S_1(\omega)|^2}. \tag{6}$$

Another approach that is derived from a Maximum Likelihood estimator is the Cross-Power Spectrum Phase (CSP) estimator [23]:

$$\psi_{\mathrm{CSP}}(f) = \frac{1}{|G_{ik}(f)|}, \tag{7}$$

and its generalization, the Modified Cross-Power Spectrum Phase formulation (MCSP) [25]:

$$\psi_{\mathrm{MCSP}}(f) = \frac{1}{|G_{ik}(f)|^{\rho}}, \tag{8}$$

where $0 < \rho \leq 1$.

2.2.2. Source position estimate from time delay

The localization of the source is determined from the knowledge of the time delay between microphone pairs.

The estimation of the localization from the time delay is obviously a non-linear problem. However, by introducing some approximations it is possible to derive simple methods to solve the problem.

Let us consider a linear four-element microphone array as a two-microphone pairs arrangement. It can be assumed that the estimated angles are taken at the mean points in each microphone pair, as shown in Fig. 2. Using simple geometric considerations, the (x, y) coordinates of the acoustic source are then computed as:

$$x = D\frac{\sin(\psi_1)\,\sin(\psi_2)}{\sin(\psi_1-\psi_2)}, \qquad y = -D\frac{\sin(\psi_1)\,\cos(\psi_2)}{\sin(\psi_1-\psi_2)}, \tag{9}$$

where D is the distance between the mean points of the two microphone pairs. Despite the approximations used in the derivation, and given that the delay time δ_{ij}

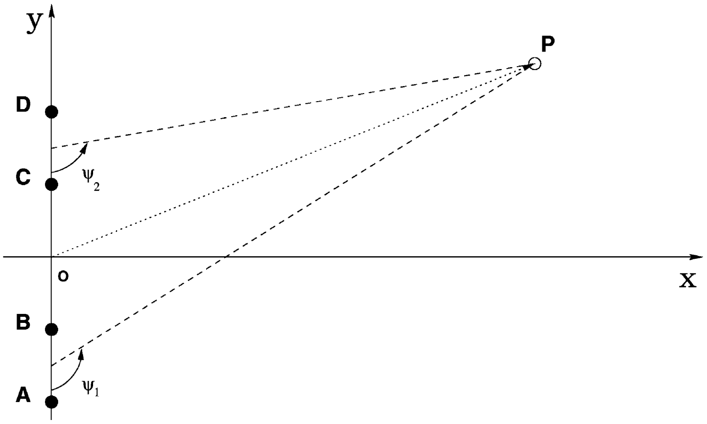

Fig. 2. Geometry for the localization problem.

is estimated with good accuracy, this approach works reasonably well in many circumstances.

More complex approaches have been described. In [6] a linear intersection estimator, which is a closed form method for the localization of a source position given sensor array time delay estimate information, is used together a nine elements orthogonal microphone array. In [25] a non-directed gradient descent search is performed over all possible locations to find a best match for a source location based on these time delays. The latter work used two four-element orthogonal microphone arrays. It is also possible to use a neural approach, as described in [21].

2.3. Acoustic localization using adaptive filtering approaches

This type of estimator is described in Fig. 3, called in the following LMS, where $s_i(n), s_j(n)$ are the signals coming from a microphone pair and $e_{ij}(n)$ is the error.

The output of an adaptive filter of length L is given by:

$$y_{ij}(n) = W_{ij}^{\mathrm{T}}(n) X_{ij}(n), \tag{10}$$

where $X_{ij}(n) = [s_j(n), s_j(n-1), \ldots, s_j(n-L+1)]^{\mathrm{T}}$ is the input vector which represents the state of the filter and $W_{ij}(n)$ is the coefficient vector. The coefficient vector is adaptively updated for tracking the dynamic of the system. A common rule for updating is LMS [15], which is described in (11)

$$W_{ij}(n+1) = W_{ij}(n) + \mu e_{ij}(n) X_{ij}^{*}(n). \tag{11}$$

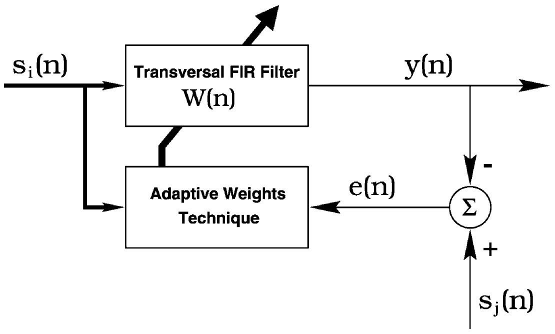

Fig. 3. Block diagram of the adaptive filtering approach.

Clearly, if the coefficients are determined in such a way that the mean squared error is minimized, then the coefficients of the filter reach an arrangement such that the $s_i(n)$ signal is delayed by the time delay between the two microphones. The time delay between the two microphones can be determined by looking at the position of the greatest coefficient of the filter.

Starting from the TDOA of two pairs of microphones in a linear microphone array set-up, we can estimate the position with the techniques described in Section 2.2.2.

2.4. Acoustic localization using neural networks

A neural network approach is particularly appealing for localization both for near and far field assumptions [3] because of its approximation capabilities [15]. Several neural network based techniques have been described. A neural network based speaker localization technique was presented in [3] for bearing estimation: in this approach the dominant frequencies of incoming signals are used to compute the correspondent CSP coefficients in the pre-processing phase of the neural network. In [10] a set of neural network is used to find the azimuth of one or several sources impinging on a linear array of equally spaced sensors, while in [11] a time delay neural network is used to bearing estimation. In [12] the far-field localization problem is addressed using recurrent neural networks; in this case, instead of considering time as a second dimension of the input space, time is implicitly coded in the structure of a recurrent neural network.

3. A neural network based algorithm using spectral subtraction

The neural network approach described in the following section is derived from our previous work [21]; a robust VAD has been added in the pre-processing phase of the input signals. In order to reduce the dimensionality of the problem, it is important to feed the network with pre-processed data; obviously the GCC is the first candidate.

The block diagram of the overall algorithm is described in Fig. 4. The algorithm is divided into a signal pre-processing phase and a neural network phase.

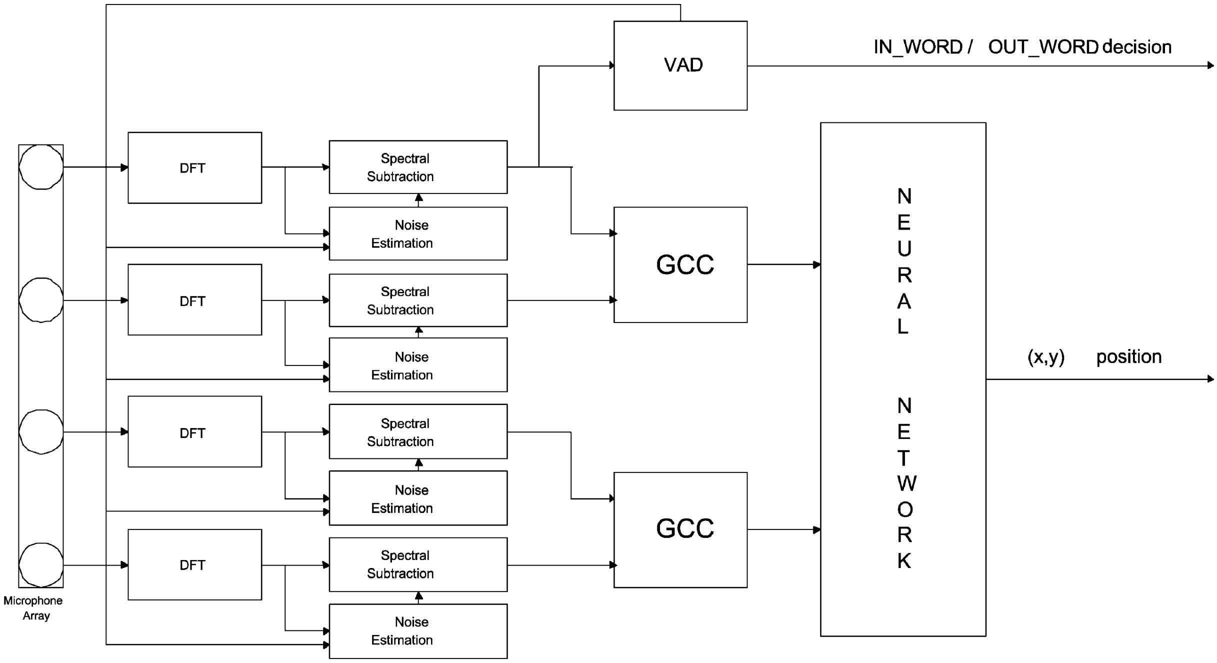

Fig. 4. Block diagram of the neural network based algorithm presented.

3.1. Signal pre-processing

The signals coming from the microphone array are divided in frames of 575 samples (corresponding to an analysis frame of 23 ms) with an overlap of 64 samples (2.56 ms). In the following the blocks depicted in Fig. 4 are described.

3.1.1. Spectral subtraction

One of the simplest techniques for noise reduction is based on power spectrum subtraction [4]. Let us suppose that $P_s(i)$ and $P_n(i)$ are the power in the ith harmonic of speech and noise, respectively, and $P_n(i)$ is supposed to be known by a noise estimation technique. If $P_s(i)$ were known, then the spectrum of the speech could have been obtained simply as $S_s(i) = \sqrt{P_s(i)}$. However, $P_s(i)$ is not known and it must be estimated from the observation X_i and from $P_n(i)$. Since X_i is complex Gaussian with variance $\sigma^2(i)$, its real and imaginary parts are Gaussian with variance $\sigma^2(i)/2$. Hence, the probability density function for X_i is:

$$p(X_i) = \frac{1}{\pi[P_s(i) + P_n(i)]} \mathrm{e}^{-|X_i|^2/(P_s(i)+P_n(i))}. \tag{12}$$

By maximizing $p(X_i)$ with respect to $P_s(i)$, the maximum likelihood estimate of $P_s(i)$ is

$$\hat{P}_s(i) = |X_i|^2 - P_n(i). \tag{13}$$

The estimation $\hat{S}_s(i)$ of the speech spectra is obtained by appending to $\hat{P}_s(i)$ the input phase as shown in (14):

$$\hat{S}_s(i) = \sqrt{\hat{P}_s(i)}\frac{X_i}{|X_i|} = \sqrt{\frac{|X_i|^2 - P_n(i)}{|X_i|^2}}X_i. \tag{14}$$

3.1.2. VAD

The role of voice activity detection (VAD) is fundamental in any localization algorithm. A generic VAD has two main goals: to trigger the localization process, which should be activated only when a vocal signal is detected by the microphone array, and to select the meaningful part of the input signal to be processed by the localization algorithm.

However, the signal is corrupted by many types of noise which occur in the environment. For this reason the VAD should be highly reliable in the presence of noise. A block diagram of the VAD algorithm, which is sub-band based, is described in Fig. 5.

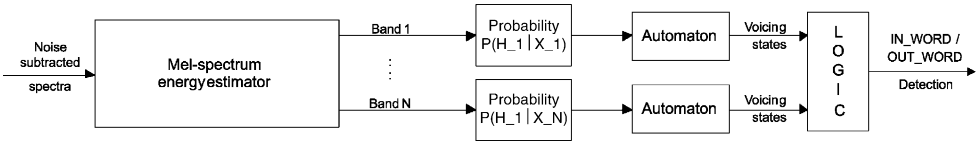

Fig. 5. Block diagram of the VAD algorithm.

As shown in Fig. 5, the VAD is formed by the following blocks:

- Mel-spectrum estimator. The Mel scale is a nonlinear scale, motivated by perceptual studies of the frequency response of the human auditory system [14]. The Mel scale reflects the fact that it is more meaningful to consider more frequencies in the low spectrum rather than in the high spectrum. The Mel-spectrum estimator thus computes the energy in the pitch-related bands of a bandpass filter bank distributed uniformly on a Mel frequency scale.
- Voicing probabilities estimation. This is the probability that speech is present in the signal and is based on statistical considerations, derived from McAulay and Malpass [20].

  Let us briefly summarize the theory behind. In the frequency domain, the noisy speech signal can be described as

$$X_i(k) = A_i(k)\,\mathrm{e}^{\mathrm{j}\phi_i(k)} + N_i(k), \tag{15}$$

where k denotes the frame number, i the number of sub-band, A and ϕ represent amplitude and phase of the voice, respectively, and N represents the noise spectrum. Our problem is to determine whether a given signal consists of noise only or speech plus noise; therefore the following binary hypothesis model is appropriate:

- ○ Event H_0 (only noise): $X_i(k) = |N_i(k)|$.
- ○ Event H_1 (speech plus noise): $X_i(k) = |A_i(k)\,\mathrm{e}^{\mathrm{j}\phi_i(k)} + N_i(k)|$.

By the Bayes rule, the 'a posteriori' probability for the presence of speech, given a measured value of the signal spectrum envelope V, is:

$$p(H_1|V) = \frac{p(V|H_1)p(H_1)}{p(V|H_1)p(H_1) + p(V|H_0)p(H_0)}. \tag{16}$$

Accordingly, the 'a posteriori' probability for the absence of speech is:

$$p(H_0|V)=\frac{p(V|H_0)p(H_0)}{p(V|H_1)p(H_1)+p(V|H_0)p(H_0)}. \tag{17}$$

Without other knowledge, the probabilities $p(H_1)$ and $p(H_0)$ can be assumed to be equally likely (i.e. both equal to 0.5). The probability density functions $p(V|H_0)$ and $p(V|H_1)$ are the a priori probability functions of the spectral envelope under hypothesis H_0 and H_1, respectively. Under hypothesis H_0, the envelope pdf $p(V|H_0)$ is a Rayleigh pdf, since the signal is only noise and the noise is assumed complex Gaussian with zero mean and variance $\sigma_i(k) = E[|N_i(k)|^2]$. Hence:

$$p(V_i(k)|H_0) = \frac{2X_i(k)}{\sigma_i(k)}\mathrm{e}^{-X_i^2(k)/\sigma_i(k)}. \tag{18}$$

Under hypothesis H_1, the envelope pdf is Rician:

$$p(V_i(k)|H_1) = \frac{2X_i(k)}{\sigma_i(k)}\mathrm{e}^{-((X_i^2(k)+A_i^2(k))/\sigma_i(k))}I_0 \times \left[\frac{2X_i(k)A_i(k)}{\sigma_i(k)}\right], \tag{19}$$

where I_0 is the modified Bessel function of the first kind. Hence, substituting $p(V|H_0)$, $p(V|H_1)$, $p(H_1)$ and $p(H_0)$ into (9), and denoting with X_i the spectral measurements, we obtain:

$$p(H_1|X_i(k)) = \frac{\mathrm{e}^{-\xi_i}\,I_0\left[2\sqrt{\xi_i X_i^2(k)/\sigma_i(k)}\right]}{1+\mathrm{e}^{-\xi_i}\,I_0\left[2\sqrt{\xi_i X_i^2(k)/\sigma_i(k)}\right]}, \tag{20}$$

where $\xi_i = A_i^2(k)/\sigma_i(k)$ is the 'a priori' SNR and $p(H_1|X_i(k))$ is the 'a posteriori' probability for the presence of speech in sub-band i for the frame k. Usually the value of the a priori SNR is held fixed for all the sub-bands and for each frame to a value in the range [5..10].

- Finite-states automaton. The voicing probability $p(H_1|X_i(k))$ of (20) is the parameter used for endpointing, and *StartThr*, *StopThr* are the thresholds that have to be applied to the level of this parameter to detect beginning and end of the vocalization. It is worth noting that since the probability is between 0 and 1, the thresholds are independent on the level of the amplitude of the signal, and therefore their setting is rather simple. Suppose moreover that *ThrBeg*, *ThrEnd* are the temporal thresholds that apply to the parameter in order to confirm the possible beginning and the possible ending of the vocalization. Suppose, finally, that *ThrTest* is a third temporal threshold that aims at recognizing whether a spike occurring during the ending phase is an artefact or denotes a speech onset. With these definition, the transitions among the various states of vocalization can be described by a simple five states automaton (namely `WAIT_WORD`, `TEST_START`, `IN_WORD`, `TEST_END`, `STOP_WORD`). The heuristic represented by this automaton is the following. An input frame is detected as initial frame of a word if $p(H_1|X_i(k))$ remains above *StartThr* for at least *ThrBeg* frames. If the probability goes below threshold within *ThrBeg* frames, then it is assumed that a spike has occurred and the search for the initial point starts again. As regards the final point, the rule is to detect the ending frame of a word when the probability remains below *StopThr* for at least *ThrEnd* frames. If the spike lasts less than *ThrTest* frames, then it is ignored. Otherwise, a transition to the `STOP_WORD` state is performed. Typical values of the mentioned threshold can be around 0.3 for *StartThr* and *StopThr*, 2 and 12 for *ThrBeg* and *ThrEnd*, respectively, and 2 for *ThrTest*.

  The automaton produces a set of states which are used both for noise estimation and for the final decision logic.
- Decision logic. The logic works as follows: a starting point is detected when at least N_1 sub-bands are in the state `IN_WORD`; similarly a frame is the ending frame of a word when at least N_2 sub-bands are in the state `STOP_WORD`. N_1 and N_2 are in the range [1..4].

An example of the VAD operation is shown in Fig. 6: in the upper panel a 20 dB utterance is shown, in the middle the 0 dB corrupted utterance is plotted

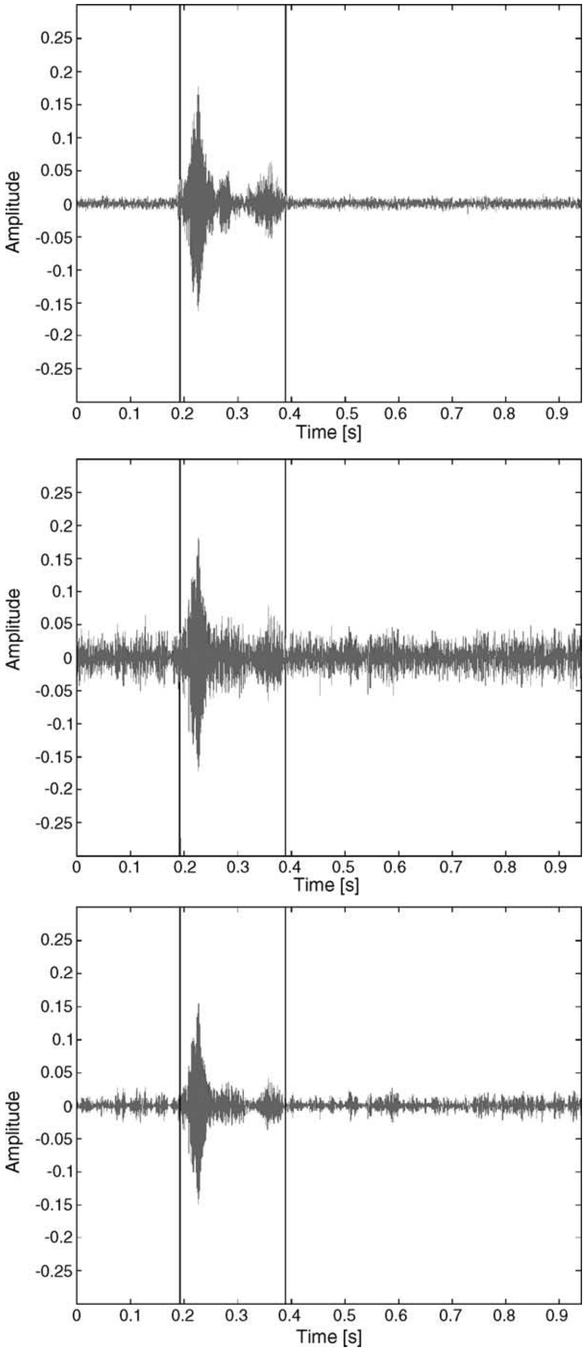

Fig. 6. Clean Signal on top panel, noisy signal on center panel and the signal after noise reduction in the bottom panel. The vertical lines point out the speech signal as revealed by the Voice Activity Detector.

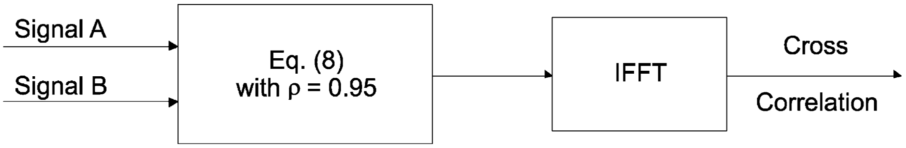

Fig. 7. GCC block diagram.

while in the lower panel the same utterance after noise-removing with the end-points marked on it is depicted.

3.1.3. *Noise estimator*

The algorithm for noise estimation makes an initial noise estimate, i.e. in a period at the very beginning where speech is not yet uttered, and adapts the noise estimation as the signal goes on, in order to track non-stationary noise signals in periods of no speech activity. The estimator calculates the weighted sum of the current spectral magnitude values X_i with the previous noise estimation in each sub-band i, leading to an exponential estimation. The computation is described as:

$$\hat{N}_i(k) = \alpha X_i(k) + (1-\alpha)\hat{N}_i(k-1), \tag{21}$$

where $X_i(k)$ is the signal spectral magnitude of the ith sub-band in the kth frame and $\hat{N}_i(k)$ is the estimated noise magnitude in the ith sub-band. The α coefficient denotes how much of the previous history should be taken into account to obtain the current noise estimate and it is responsible of the noise tracking capabilities of the algorithm. Of course, noise estimation is made in noisy sections thus, for each sub-band, the presence of noise is detected with the finite-states automaton

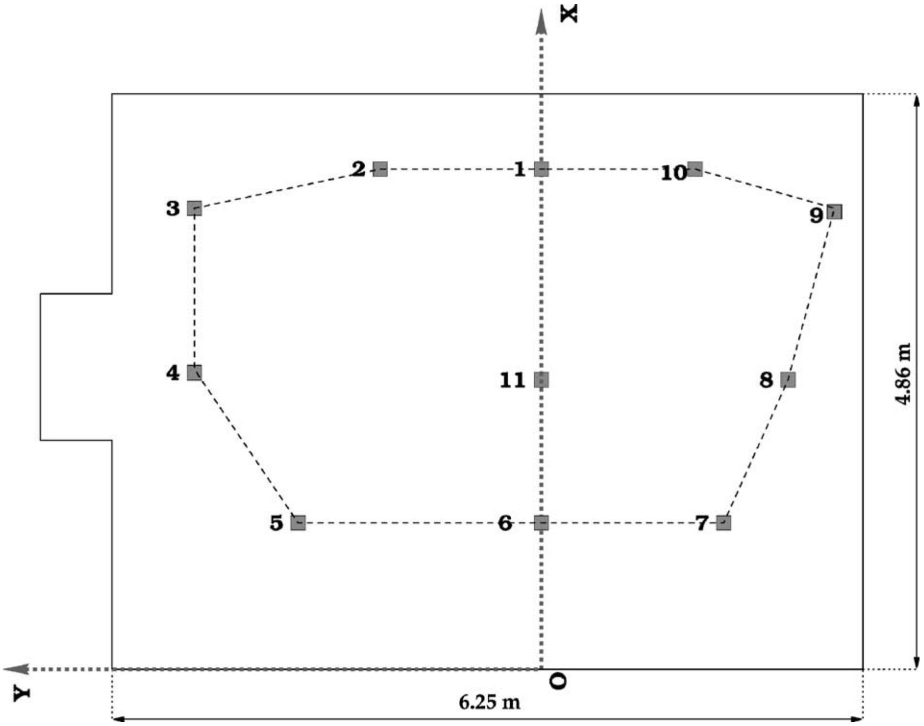

Fig. 8. Plane view of the test environment.

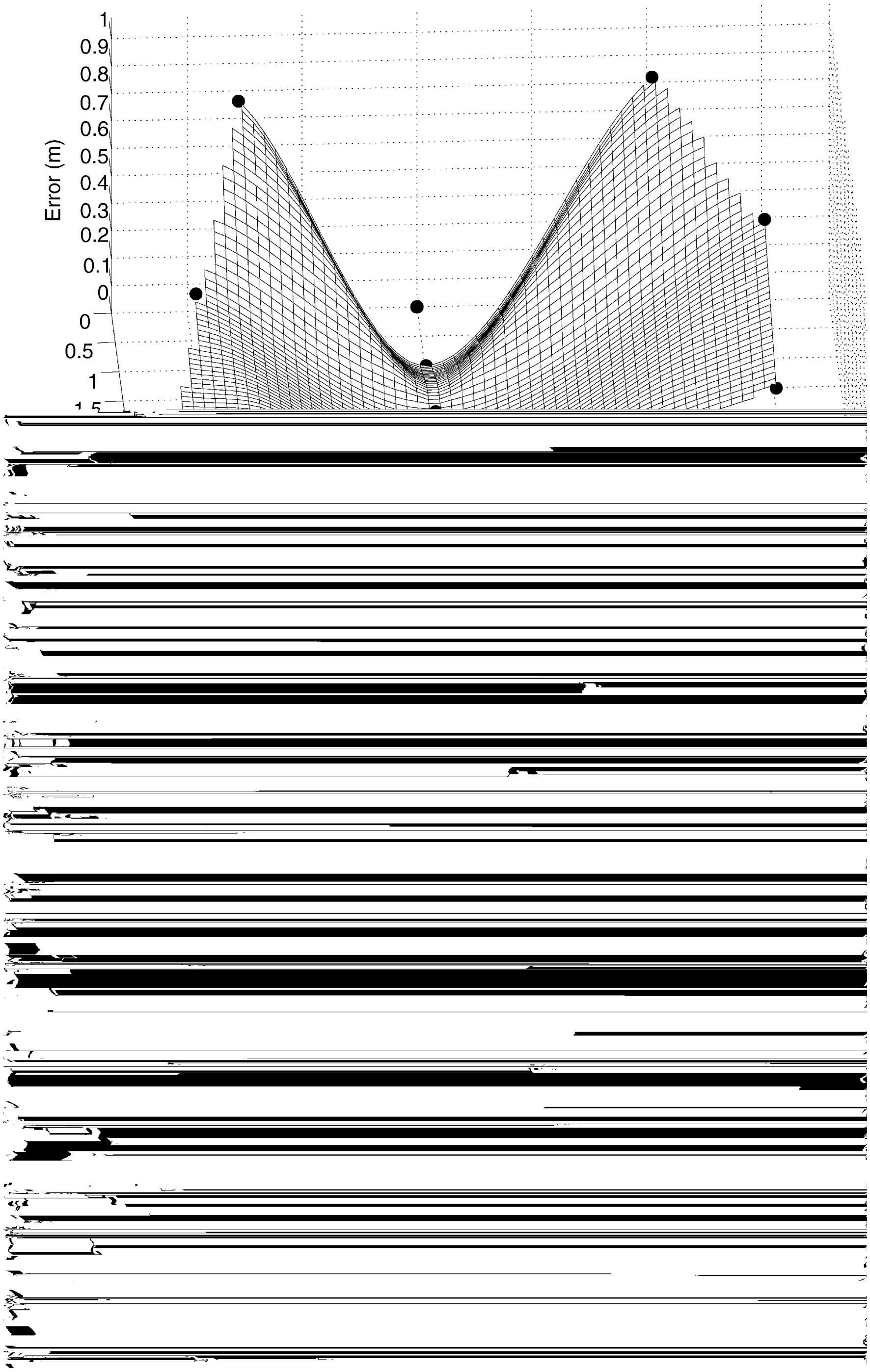

Fig. 9. Graph of the absolute error (above) and plane view of the environment using AML plus geometric localization and the signal mvasa.

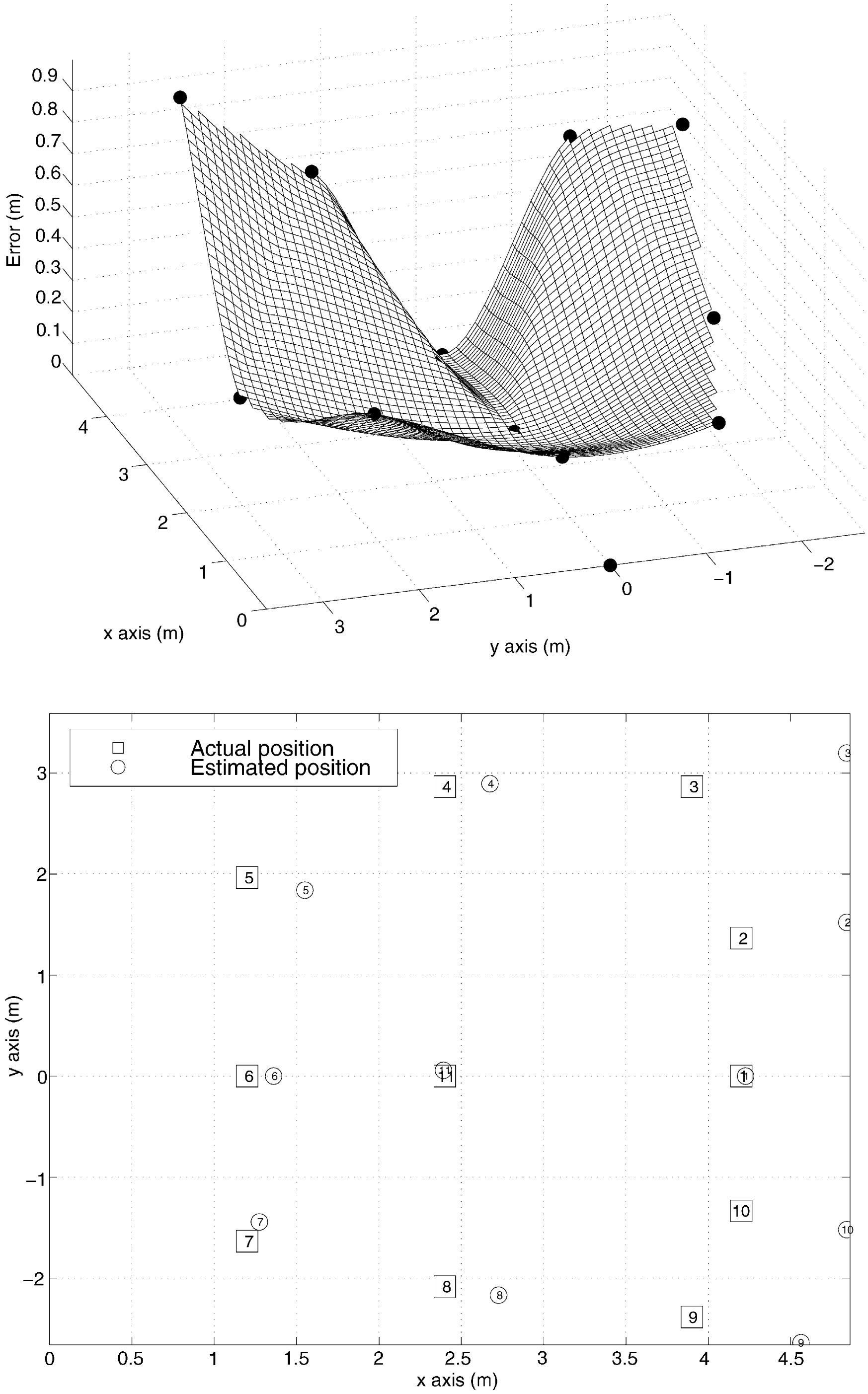

Fig. 10. Graph of the absolute error (above) and plane view of the environment using CSP plus geometric localization and the signal mvasa.

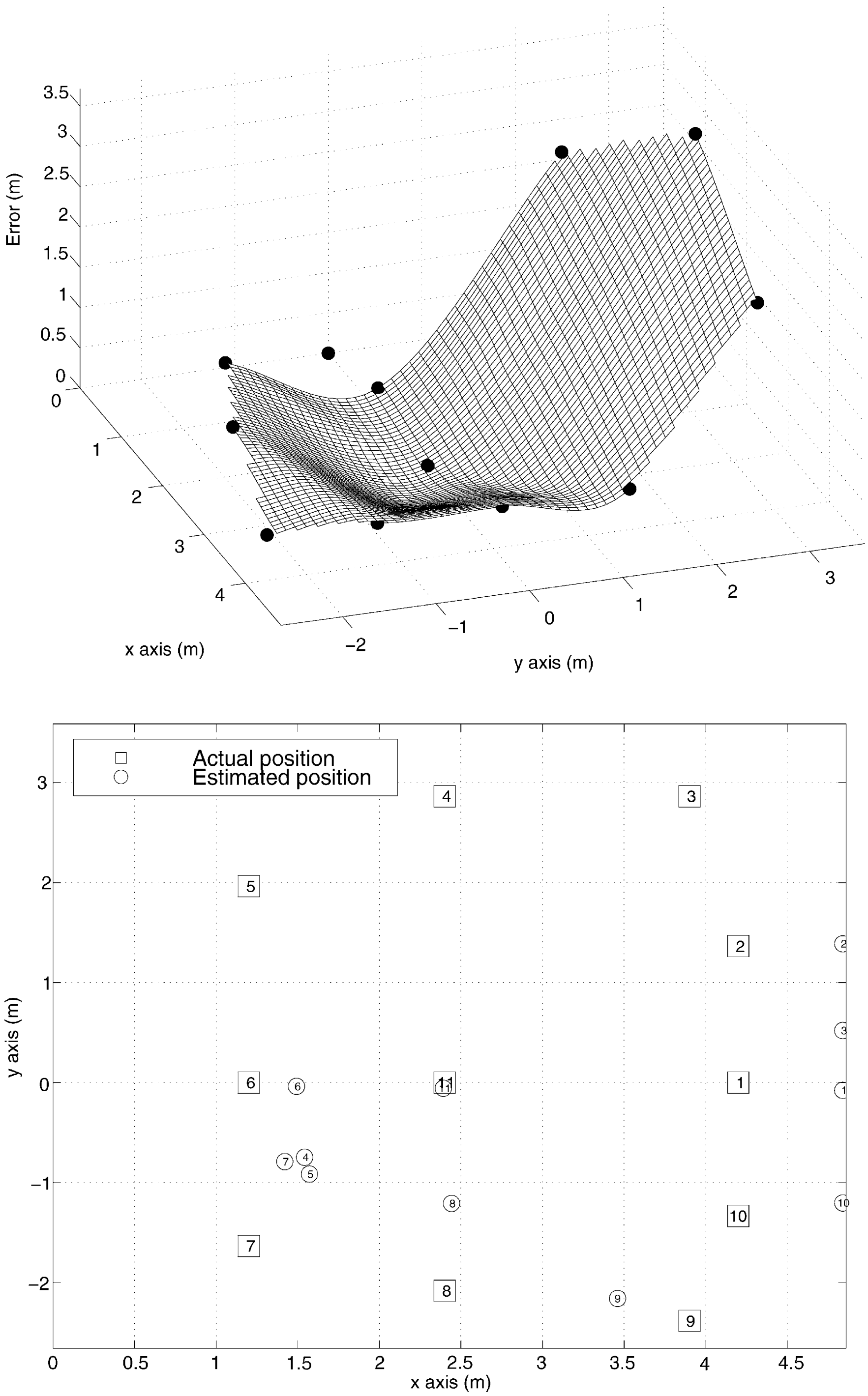

Fig. 11. Graph of the absolute error (above) and plane view of the environment using LMS plus geometric localization and the signal mvasa.

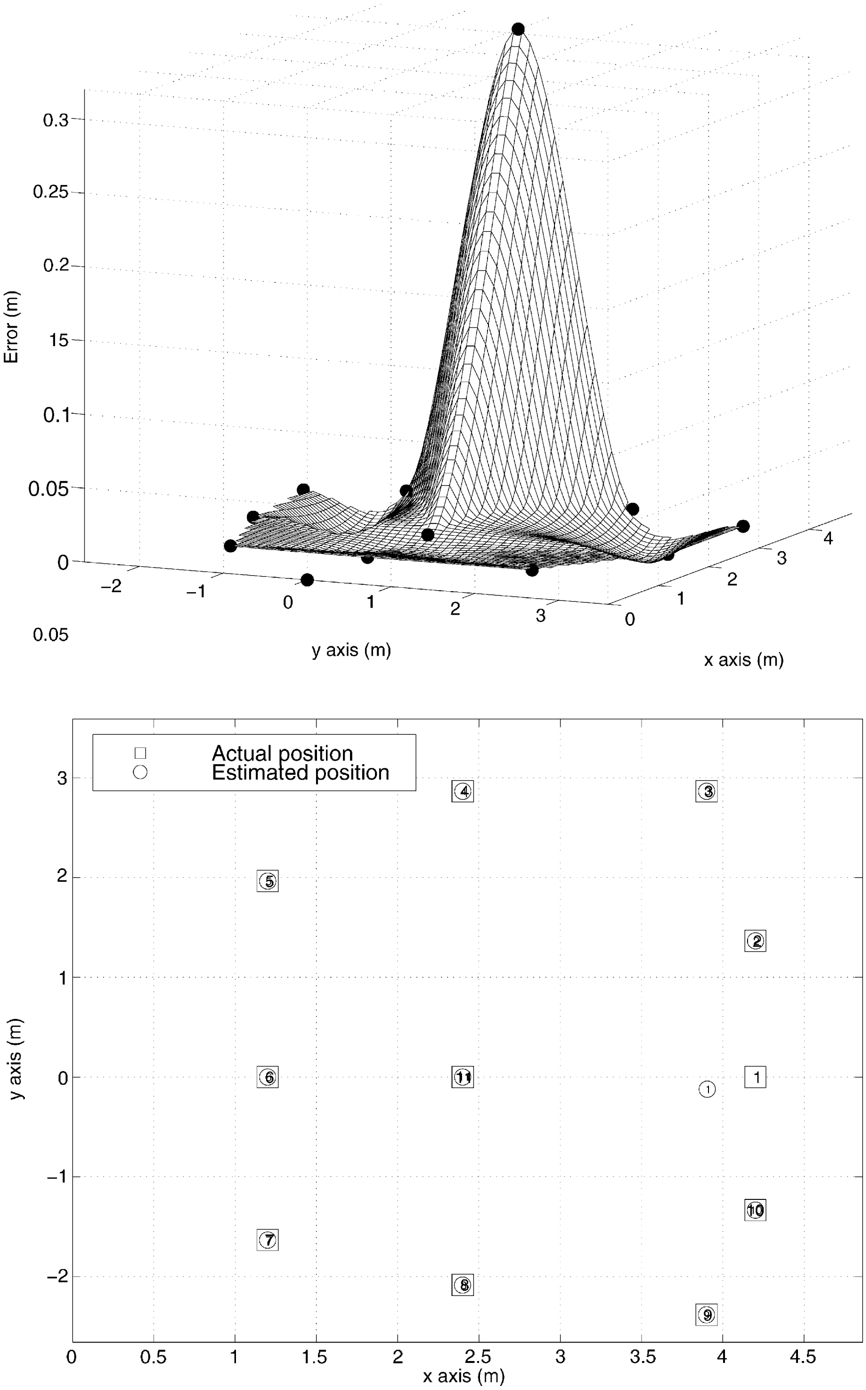

Fig. 12. Graph of the absolute error (above) and plane view of the environment using the presented neural network algorithm and the signal mvasa.

described above. The adaptation described in (21) in fact is performed in correspondence to the automaton states `WAIT_WORD` and `TEST_END`.

3.1.4. GCC

This block, described in Fig. 7, has the goal of computing the MCSP of the incoming signals. The outputs of this block is the cross-correlation between a couple of signals coming from the microphones. The maximum time delay with the microphone configuration considered in this work is about 0.9 ms. Therefore the inputs of the neural network were 64 samples of cross-correlation (or ±1.2 ms) obtained from each of the two GCC computed on the analysis frame.

3.2. Neural network

One of the most important parameter of a neural network is the number of its hidden nodes. This is responsible of the trade-off among convergence, complexity and performance of the network. The neural network model adopted was a Multi-Layer Perceptron (MLP) [16] with three hidden layers, obtained merging two simple neural networks devoted to the subproblems of calculating the time delay and estimating the (x, y) position from the time delay, respectively. The starting weights were the trained ones of the two separate simple neural networks. The resulting neural network has three hidden layers, 128 inputs and 2 output, the (x, y) coordinates of the source. Several optimization techniques, namely backpropagation with momentum, the Levemberg–Marquardt approach, Newton-based approaches, have been tested for training the neural network, and the best results were obtained with Quickprop [13] and Rprop [26], which are fast local Newton-based optimization techniques. Therefore, it is worth noting that both the time delay and the source position estimates are made with neural networks.

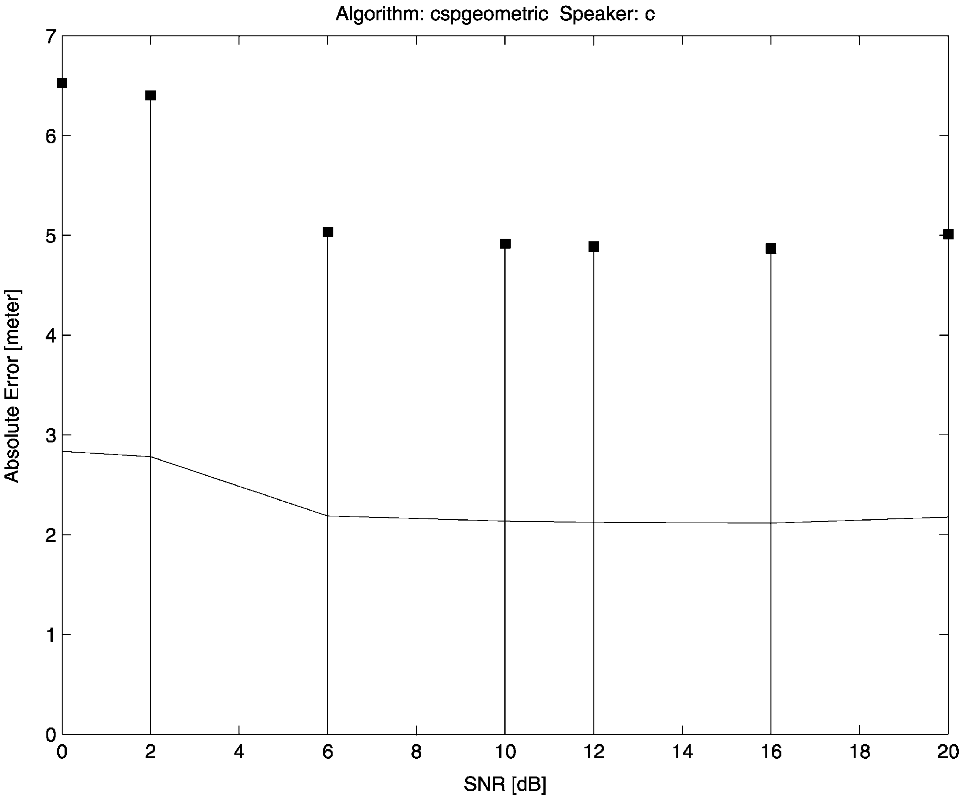

Fig. 13. Localization mean performances of CSP algorithm and geometric localization for several SNR ratios for the speaker c.

4. Experimental set-up

The environment chosen for the experiments was the small room shown in Fig. 8. A four-elements linear microphone array, connected to a DSP board based on TSM320C50 DSP capable of acquiring four synchronous channel up to the frequency of 25 kHz were fixed on the wall of the room (at the origin of the y-axes of Fig. 8). The DSP board was mounted on a PC (Woodstock) connected via radio-link to the mobile robot. A cooling system plus a number of computers present in the room generated fan noise, so that speech uttered in the room and acquired by the microphone array had an SNR of 20 dB on the average. The fan noise was also acquired and digitally added to the recorded phrases in order to achieve the desired SNR values.

For testing the algorithms a database of signals was acquired with a sampling rate of 25 kHz. The following Italian phrases, typical in the context of human–robot interaction, were read 10 times by two speakers, and acquired in digital format:

- *Vieni qui (Come here)*—acronym vq,
- *Vai al sito A (Go to the A site)*—acronym vasa,
- *Prendi l'oggetto B (Take the B object)*—acronym plob.

With this procedure, a speech database of 60 phrases was obtained. The database, played with a loudspeaker put at a height of 1.6 m in the 11 points of Fig. 8 with an acoustic intensity of about 60 dB, corresponding to a normal human conversation, were acquired by the microphone array. The microphones of the array were low cost omni-directional devices (Sony ECM-F8) with a bandwidth of 12 kHz. After the acquisition with the microphone array, a new database of 2640 phrases (60 phrases × 11 points × 4 microphones) was obtained for each microphone configuration.

While for some algorithms, such as AML, CSP and LMS, no training is necessary and therefore the

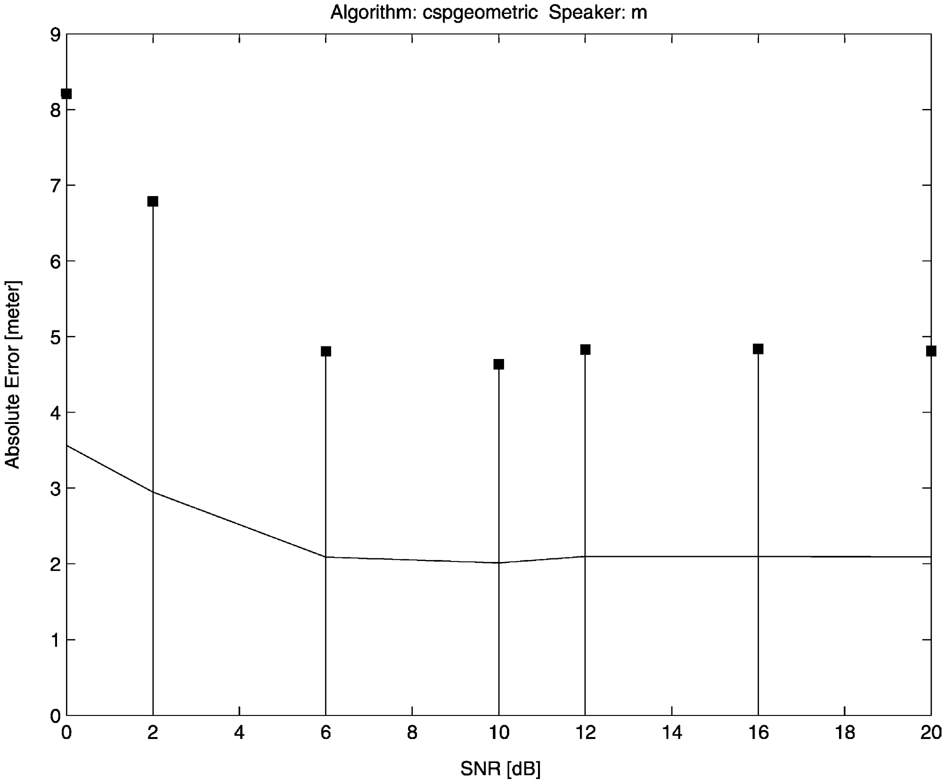

Fig. 14. Localization mean performances of CSP algorithm and geometric localization for several SNR ratios for the speaker m.

database can be used ‘as it is’, the neural network must be trained. However, the acquired database is related only to 11 points which are not sufficient for the network to reach a good generalization for localization. For this reason the training of the neural network was realized combining artificially created signals in addition to real data acquired in the reverberant room. The artificial data were created by delaying the real signals from a given position up to the microphones constituting the array, assuming the absence of reflection or reverberation. Equally spaced position, separated 50 cm each, were selected in the plan view of the small room.

5. Experimental results

Several experiments have been performed with the algorithms presented in Section 2 and with the database described above with different configurations of the microphones. In the following results we have used signal frames of 575 samples. Experiments with our neural network based algorithm was performed as follows: 30 phrases of the baseline database plus 50 artificial signals were used for training the neural network,al 7se

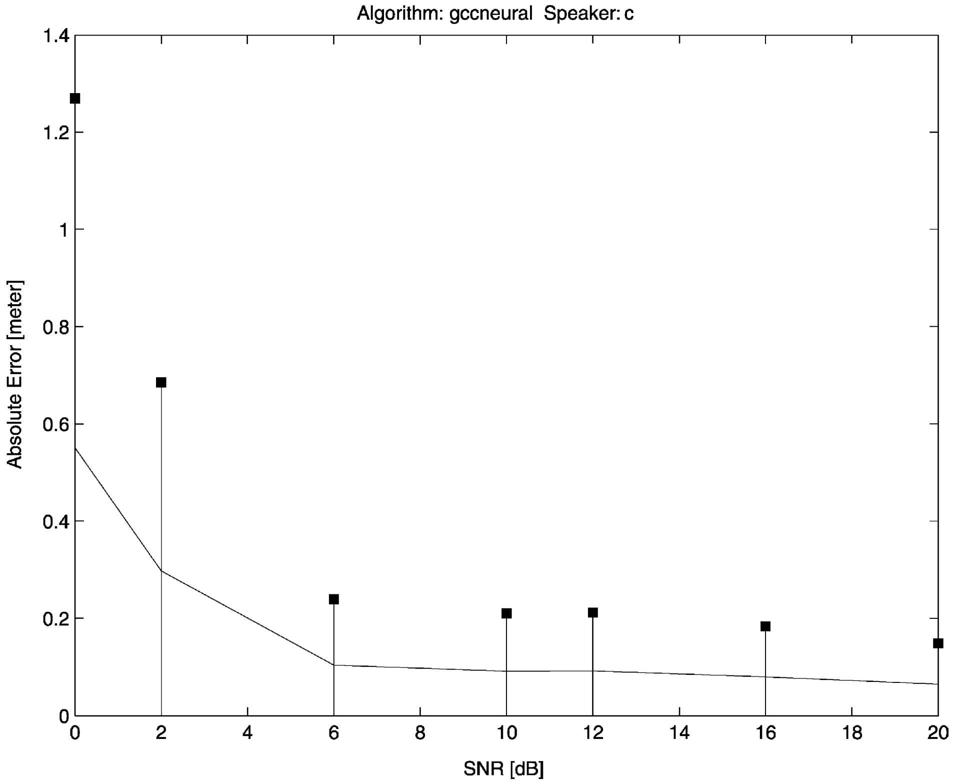

Fig. 15. Localization mean performances of the presented neural network algorithm for several SNR ratios for the speaker c.

estimated point locations. In the upper panel, a 3D plot of the absolute localization error in each point is reported.

From the presented results we can notice that the most difficult to localize points' locations are the ones numbered 3–5 of the test environment. This is due to the fact that such points are close to the door aperture (see Fig. 8) and therefore more affected than other by air turbulence. The best performing technique is the neural network based algorithm described in this paper, since its average localization error is about 10 cm at an SNR of 20 dB, while the average error of the CSP is about 200 cm. It is worth noting, however, that the CSP use a geometric approach for the source position estimate from the time delay. For this reason its performance is quite poor especially in the points which do not respect the far-field assumption and are more sensitive to the reflection of the walls.

Several investigations were then performed with the neural network algorithm at lower SNRs, from 20 dB (see Fig. 12) down to 0 dB. For this reason the fan noise were digitally added to each of the test phrases acquired by the microphone array in order to obtain phrases with SNRs of 16, 12, 10, 6, 2 and 0 dB. Such a noisy speech were fed to pre-processing block. It is worth noting that the training is performed with 20 dB data while the testing is performed with lower SNR data, so noise-reduction is fundamental.

The localization performance was given by averaging the absolute error over all points:

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i^T - x_i^{\mathcal{E}})^2 + (y_i^T - y_i^{\mathcal{E}})^2}, \qquad (22)$$

where x^T and y^T are the true coordinate, $x^{\mathcal{E}}$ and $y^{\mathcal{E}}$ are the estimated coordinates (on the basis of the algorithm) and i is the index of the points. $\mathcal{E}$ is a stochastic variable with an exponential pdf, i.e. $p(x) = \lambda\,\mathrm{e}^{-\lambda x}$. We have experimentally shown that λ is a function of the noise power, and it was estimated using

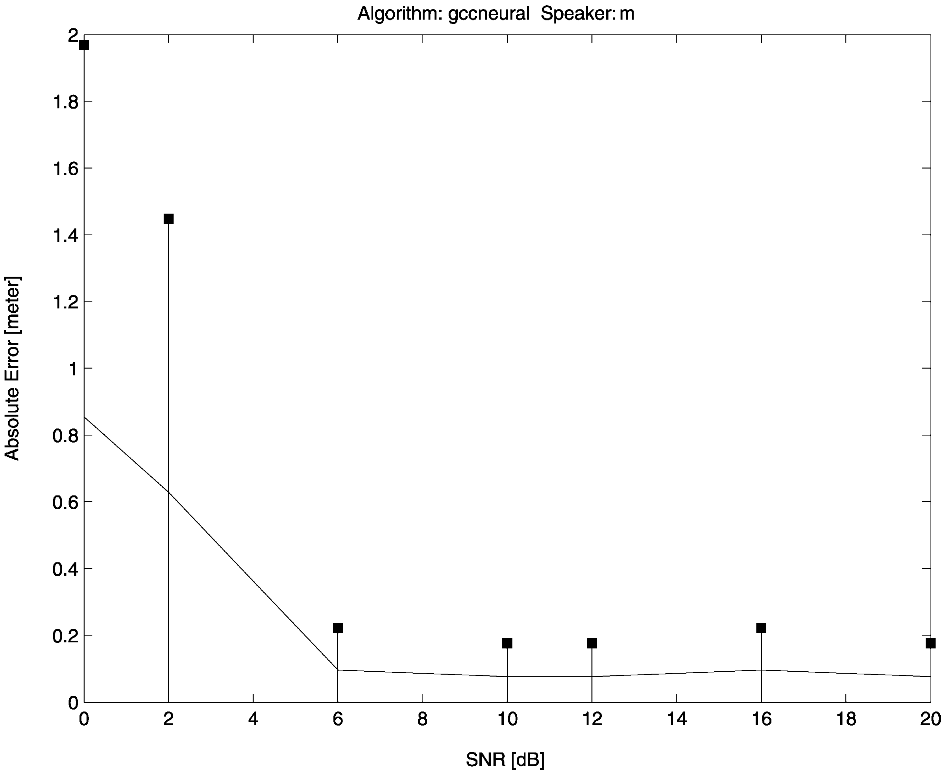

Fig. 16. Localization mean performances of the presented neural network algorithm for several SNR ratios for the speaker m.

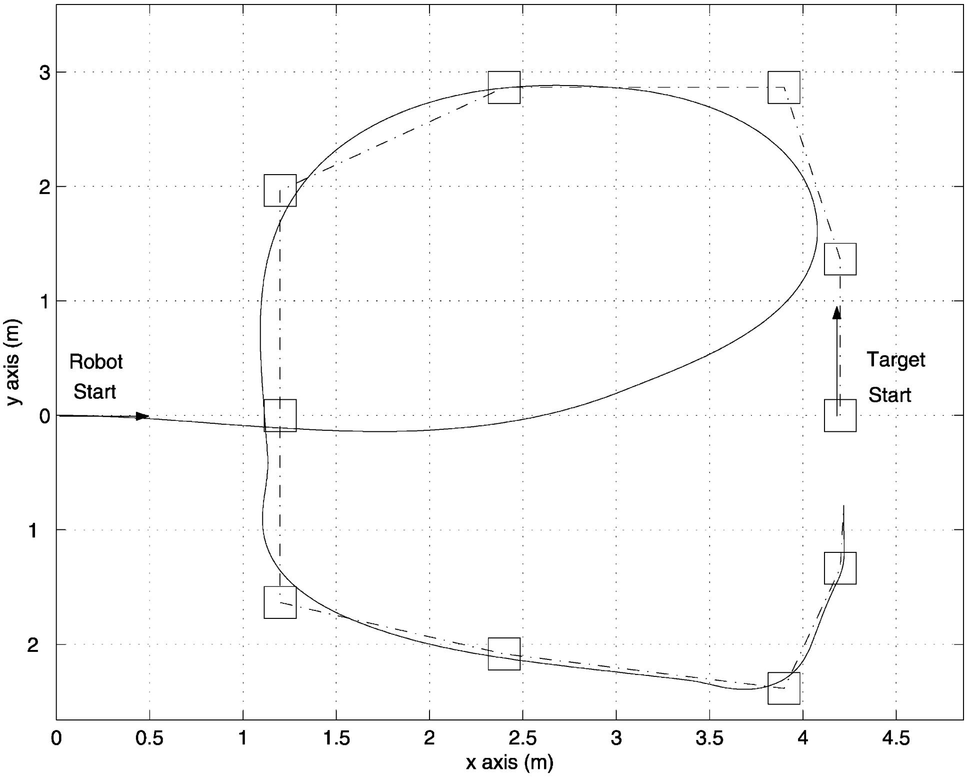

Fig. 17. Example of robot trajectory generate with the ξ-model in order to track a moving acoustic source.

classical best-fitting procedures. With the estimated λ, the probability that $\mathcal{E}$ is between 0 and x is set to 90%:

$$\text{Prob}\{0 \leq \mathcal{E} \leq x\} = \int_0^x \lambda\, \mathrm{e}^{-\lambda\zeta}\, \mathrm{d}\zeta = 0.9. \qquad (23)$$

From (23) it is straightforward to derive the value of x, which is the 90% confidence interval of $\mathcal{E}$:

$$x = -\frac{1}{\lambda}\ln(1-0.9). \qquad (24)$$

This value is shown with black squares in Figs. 13 and 14 for two speakers as obtained with the CSP algorithm, while the average value is the solid line. Similar measurements have been performed for the neural algorithm proposed in this paper. The experimental results are shown in Figs. 15 and 16. A comparison with the previous results shows the great improvement of localization accuracy obtained by the neural network approach over CSP at high SNR (10 vs. 200 cm at 20 dB) as well as at lower SNR (40 vs. 350 cm at 0 dB) and also the relatively independence of CSP against SNR. On the other hand, the CSP algorithm is much more simple to implement because no training is required.

As described above, the robot reacts to the signal acquired by the microphone array by planning an approaching navigation mission, thus making the robot follow the moving acoustic target. In Fig. 17, an example of the robot trajectory generated with the ξ-model is reported in solid line. In this experiment, a talker moved slowly at about 0.7 m/s across some points marked on the floor while uttering the number of the points, and the robot performs a talker following task. The robot stops when the distance between the robot and the target is below 20 cm.

6. Conclusions and final remarks

The results demonstrate that the neural localization algorithm described in this paper is quite robust against noise. All the algorithms presented in Sections 2 and 3 are well suited for real-time applications because of the limited computational efforts required and the short frame lengths of the acquired signals used. All of them have been tested, with different loudspeaker heights and different configurations of the microphones, with the four-elements linear array considered.

The performance of the algorithm are quite good as the localization error is about 40 cm at 0 dB. This is due to the use of spectral subtraction for noise removing and to the development of the VAD with good noise robustness.

The algorithm was integrated in a mobile robot and the trajectory in a talker-following task has been also reported.

Current development activities are concerned with the introduction of a microphone array on the robot itself, in order to obtain relative localization information rather then absolute ones. This leads to the issues of relative movements between noise sources and array; of course the same algorithms can be used if the relative movements are slow. Moreover, vocal interactions using isolated words speech recognition and speech synthesis are being integrated in the system. Finally, a surveillance application of the robot is currently being studied.

References

[1] P. Aarabi, S. Zaky, Robust sound localization using multi-source audio-visual information fusion, Information Fusion 2 (2001) 209–223.

[2] R.C. Arkin, T. Balch, AuRA: principles and practice in review, Journal of Experimental and Theoretical Artificial Intelligence 9 (2–3) (1997) 175–189.

[3] G. Arslan, F.A. Sakarya, A unified neural-network-based speaker localization technique, IEEE Transactions on Neural Networks 11 (2000) 997–1002.

[4] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on Acoustics, Speech, and Signal Processing 27 (2) (1979) 113–120.

[5] R.P. Bonasso, R.J. Firby, E. Gat, D. Kortenkamp, D. Miller, M. Slack, Experiences with an architecture for intelligent, reactive agents, Journal of Experimental and Theoretical Artificial Intelligence 9 (2–3) (1997) 237–256.

[6] M.S. Brandstein, J.E. Adcock, J.H. DiBiase, H.F. Silverman, A closed-form method for finding source locations from microphone-array time-delay estimates, in: Proceedings of the ICASSP'95, Vol. 17, Detroit, MI, May 1995, pp. 3019–3022.

[7] M.S. Brandstein, J.E. Adcock, H.F. Silverman, A practical time-delay estimator for localizing speech sources with a microphone array, Computer Speech and Language 9 (1995) 153–169.

[8] M.S. Brandstein, H.F. Silverman, A practical methodology for speech source localization with microphone arrays, Computer Speech and Language 11 (2) (1997) 91–126.

[9] R. Brooks, A robust layered control system for a mobile robot, IEEE Journal of Robotics and Automation 2 (1) (1986) 14–23.

[10] B. Colnet, J.-P. Haton, Far field array processing with neural networks, in: Proceedings of the ICASSP'94, Vol. 2, 1994, pp. 281–284.

[11] B. Colnet, J.D. Martino, Bearing estimation with time-delay neural networks, in: Proceedings of the ICASSP'95, Vol. 5, 1995, pp. 3583–3586.

[12] B. Colnet, J.D. Martino, Source localisation with recurrent neural networks, in: Proceedings of the ICASSP'96, Vol. 6, 1996, pp. 3073–3076.

[13] S.E. Fahlman, An empirical study of learning speed in back-propagation networks, Technical Report CMU-CS-88-162, Carnegie Mellon University, Pittsburgh, PA, September 1988.

[14] D. Greenwood, Critical bandwidth and consonance: in relation to cochlear frequency-position coordinates, Hearing Research 54 (1991) 164–208.

[15] S. Haykin, Adaptive Filter Theory, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

[16] S. Haykin, Neural Network A Comprehensive Foundation, 2nd ed., Macmillan, New York, 1998.

[17] http://www.robotics.laboratorium.dist.unige.it/Projects/CertNaz/index.html.

[18] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, N. Sugie, A model-based sound localization system and its application to robot navigation, Robotics and Autonomous System 27 (1999) 199–209.

[19] C.H. Knapp, G.C. Carter, The generalized correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-24 (1976) 320–327.

[20] R. McAulay, M. Malpass, Speech enhancement using a soft-decision noise suppression filter, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) 137–145.

[21] E. Mumolo, M. Nolich, G. Vercelli, Algorithms and architectures for acoustic localization based on microphone array in service robotics, in: Proceedings of the ICRA2000, Vol. 3, 2000, pp. 2966–2971.

[22] M. Omologo, P. Svaizer, Acoustic event localization using a crosspower-spectrum phase based technique, in: Proceedings of the ICASSP'94, Vol. 2, Adelaide, 1994, pp. 273–276.

[23] M. Omologo, P. Svaizer, Acoustic source localization in noisy and reverberant environment using a CSP analysis, in: Proceedings of the ICASSP'96, Vol. 2, Atlanta, 1996, pp. 921–924.

[24] D. Rabinkin, R. Renomeron, A. Dahl, J. French, J. Flanagan, M. Bianchi, A DSP implementation of source location using microphone arrays, Journal of Acoustical Society of America 99 (1996) 2503.

[25] D. Rabinkin, R. Renomeron, J. French, J. Flanagan, Estimation of wavefront arrival delay using the cross-power

spectrum phase technique, Journal of Acoustical Society of America 100 (1996) 2697.

[26] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: Proceedings of the ICNN, San Francisco, CA, 1993, pp. 586–591.

[27] V. Sanguineti, T. Tsuji, P. Morasso, A dynamical model for the generation of curved trajectories, in: S. Gielen, B. Kappen (Eds.), Artificial Neural Networks, Springer, Berlin, 1993, pp. 115–118.

[28] H.F. Silverman, S.E. Kirtman, A two-stage algorithm for determining talker location from linear microphone array data, Computer Speech and Language 6 (1992) 129–152.

[29] D.E.
Sturim, M.S. Brandstein, H.F. Silverman, Tracking multiple talkers using microphone-array measurements, in: Proceedings of the ICASSP'97, Vol. 1, April 1997, pp. 371–374.

Enzo Mumolo received a Dr. Eng. degree in electronical engineering from the University of Trieste, Italy, in 1982, where he conducted research in signal processing before joining in 1984 the Central Laboratory of Alcatel Italia in Pomezia, Rome, Italy. In 1985 he was with ITT DCD-West in San Diego, CA. In 1987 he became responsible for research activities within the Speech Processing Department of the Alcatel Italia Lab. From 1990 to 1991 he was with Sincrotrone Trieste, Italy, as head of the Electronics Group. In 1991 he joined the Computer Science Department at DEEI, University of Trieste, as Research Engineer and Assistant Professor in Operating Systems. His current research interests include non-linear systems, operating systems and robotics. Member of IEEE and ACM, he has published more than 90 papers and holds two United States patents.

Massimiliano Nolich received his Laurea degree in electronic engineering in 1999. He is currently a Ph.D. student in computer science at the Engineering Faculty of the University of Trieste, working on algorithms and software techniques for robotic platforms. He is a member of the IEEE Computer Society and of the Italian Association for Artificial Intelligence. His scientific interests are focused on robotics, artificial intelligence, signal processing and operating systems. He has written more than 10 papers.

Gianni Vercelli received his Laurea degree in electronic engineering in 1987 and his Ph.D. in computer science in 1992. He was with the University of Trieste, Italy, from 1996 to 1999, and he is currently an Assistant Professor in Computer Science and Multimedia Design at the Education Faculty of the University of Genoa. He is a member of the IEEE Computer Society and of the Italian Association for Artificial Intelligence. His scientific interests are focused on robotics and artificial intelligence, intelligent agents, and multimedia education. He has written more than 70 papers.