Attentional mechanism based on a microphone array for embedded devices and a single camera

Antonio Martinez-Colon, Jose M. Perez-Lorenzo, Fernando Rivas, Raquel Viciana-Abad, and Pedro Reche-Lopez

Multimedia & Multimodal Processing Research Group
Telecommunication Engineering Department - University of Jaén
{amcolon, jmperez, rivas, rviciana, pjreche}@ujaen.es

Abstract. This work presents an attentional mechanism with the capability of detecting the localization of a speaker for interaction purposes, based on audio and video information. The localization is computed in terms of azimuth and elevation angles, to be used as input values for controlling mobile systems such as a pan-tilt videocamera or a robotic head. For this purpose the SRP-PHAT algorithm has been implemented with a commercial array of microphones for embedded devices, in order to estimate the localization of a sound source in the surroundings of the array. In order to improve the limitations of the SRP-PHAT algorithm in the estimation of the z coordinate, the elevation angle is corrected via video information by using Haar cascade classifiers for face detection. Simulations and experiments show the accuracy of the system, as well as the application for controlling a pan-tilt videocamera in a real scenario with speakers and ambient noise.

Keywords: attentional mechanism, audio source localization, microphone array, face detector

1 Introduction

In order to obtain an effective interaction between humans and machines or physical agents, the machines have to be able to localize people in its surroundings [1]. Moreover, machines have to pay attention to the human voice and look at the possible speakers. In a dynamical situation where the speakers are in movement the physical agents should be able to track them. The development of a good attentional mechanism has many applications in a great variety of engineering and research fields such as social robotics, video-conference systems and speaker diarization for meetings.

In this paper, an attentional mechanism has been developed based on audio and visual information. These data are used in order to localize possible human speakers. In our approach an acoustical azimuthal localization is performed by means of SRP-PHAT algorithm [2] whereas the visual localization is performed with a camera using the OpenCV Haar-cascade face and eyes detector [3] to set the elevation angle. In our proposal an array with only 4 microphones geometrically closed among them is used. The maximum separation among microphones

is 8.6 cm which allows its utilization in compact systems of videoconference as well as in social robotics.

Related work on person tracking in the field of social robots includes [4–7]. In [4] an array of 8 microphones mounted on a rectangular prism of dimensions 50 cm × 40 cm × 36 cm is used for sound source localization on a mobile robot. Nakamura et al. use an extension of MUSIC algorithm in a robot-embedded 8-channel circular microphone array, along with thermal and distance cameras [5]. In contrast to these works, our proposal is based on an array with a lower number of microphones and with a smaller size. Also, binaural architectures are found in the literature. As examples, Ferreira et al. propose a framework for perception by fusing vision, vestibular sensing and binaural auditory system [6]. Viciana-Abad et al. make a bioinspired proposal using a pair of microphones and a stereo vision camera system [7]. However, in binaural approaches, the audio system may present problems in the accuracy of the detection in noisy environments and in the front-back ambiguity. In our work, the use of a planar array instead of a linear one allows to detect the direction of arrival of any speaker 360° around the array without ambiguity. Respect to the applications for videoconference systems and diarization, most of them use in general microphones that are distributed within the room and very far away each others [8–10], frequently with a great number of microphones [11].

The rest of this paper is organized as follows. Section 2 presents the fundamentals of SRP-PHAT, follows with the implementation of the algorithm, and ends with simulations with a MATLAB software. Next, Section 3 describes briefly the visual features that are included in the system. Section 4 presents the results in the audio localization task with two different arrays, being one of them a commercial one for embbeded devices, and comparing the implemented algorithm also with a commercial one. At the end of the Section some results with the whole system are also shown. Finally, conclusions and future work are described in Section 5.

2 Audio source localization

2.1 Introduction to SRP-PHAT

An array of microphones can be defined as a set of N microphones, where every microphone is located at an unique position. Then, in a simple model, it can be supposed that the acoustic waves originated at a sound source follow a direct path to every microphone along N lines simultaneously. The orientation of these lines defines the direction of the propagation vectors (Fig. 1) and the time of arrival of the waves at every microphone will depend on these vectors. Methods for the localizacion of acoustical sources using an array of microphones can be divided into two groups: indirect and direct. The indirect methods are typically based on two steps. First, the *Time Difference of Arrival* (TDOA) is estimated for each of the existing pairs of microphones within the array. Then, based on these time delay estimation (TDE) values, the localization of the acoustic source can be estimated if the geometry of the microphones array is well known. On the

other hand, the direct methods are able to compute both TDOA and localization estimations in a single step. For that purpose, direct methods perform a scanning of localizations in the surroundings of the microphones array, and the most likely position is selected as the estimation of the acoustical source. Both type of approaches, indirect and direct, are also often based on the General Cross Correlation (GCC) technique [12]. If an array of microphones is used, the GCC for two signals captured by a pair of microphones (k, l) is defined as:

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \psi_{kl}(w) M_k(w) M_l^*(w) e^{jw\tau} dw \tag{1}$$

being τ a time delay, $M_k(w)$ and $M_l(w)$ are the Fourier Transform of the microphone signals $m_k(t)$ and $m_l(t)$ respectively, * denotes the complex conjugation, and $\psi_{kl}(w)$ is known as the generalized frequency weighting. The choice of this weighting function may influence on the performance of the method under adverse acoustic conditions [13], and the Phase Transform (PHAT) weighting is one of the most widely used, which is defined as:

$$\psi_{kl}(w) = \frac{1}{|M_k(w) M_l^*(w)|} \tag{2}$$

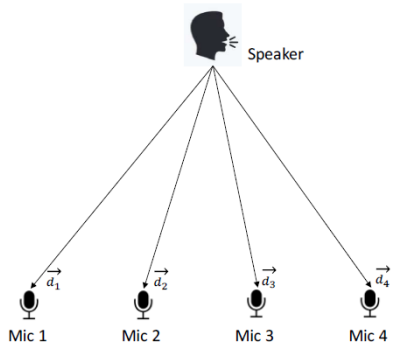

Fig. 1. Propagation vectors for an array of 4 microphones

One of the best known direct algorithms for the localization of sound sources is the Steered Response Power-Phase Transform (SRP-PHAT) [14] [15], which uses the PHAT weighting for the computation of the GCC for every pair of microphones inside the array. The SRP-PHAT algorithm is commonly interpreted as a beamforming method which is used to distinguish the spatial properties of a target signal from a background noise. For this purpose, the position of a candidate source that maximizes the output of a steered delay-and-sum beamformer is searched. The SRP-PHAT algorithm has been proved to be a robust method

under adverse conditions, although its computational cost may be a problem for real time applications.

2.2 Implementation

In this subsection the implementation made for the SRP-PHAT algorithm is described. The steered response power (SRP) at the surroundings of the array can be expressed as:

$$P_n(\mathbf{x}) = \int_{nT}^{(n+1)T} \Big| \sum_{l=1}^{N} w_l m_l(t - \tau(\mathbf{x}, l)) \Big|^2 dt \tag{3}$$

being $\mathbf{x}$ the spatial point where the SRP is computed, n the time frame of length T, N the number of microphones, $m_l(t)$ denotes the signal output for a given microphone l, w_l is a weight, and $\tau(\mathbf{x}, l)$ is the propagation time of the direct path from the point $\mathbf{x}$ towards the microphone l.

Removing some terms of fixed energy, the part of $P_n(\mathbf{x})$ that is variable with $\mathbf{x}$ can be expressed as [2]:

$$P'_n(\mathbf{x}) = \sum_{k=1}^{N} \sum_{l=k+1}^{N} R_{m_k m_l}(\tau_{kl}(\mathbf{x})) \tag{4}$$

being $R_{m_k m_l}$ the GCC for the pair of microphones (k, l) as defined in (1), and $\tau_{kl}(\mathbf{x})$ is the Inter-Microphone Time-Delay Function (IMTDF), which can be expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{||\mathbf{x} - x_k|| - ||\mathbf{x} - x_l||}{c} \tag{5}$$

being c the speed of sound, and x_k, x_l the points where microphones k and l are respectively located.

The SRP-PHAT algorithm evaluates (4) over a grid of the space in order to find a maxima, and typically it can be implemented following next steps:

1. Grid Definition. A spatial grid is defined with a given resolution. Then, the theoretical delays between all the points of the grid and every pair of the microphones that can be formed are precomputed. This is done only once at the beginning of the application, and these precomputed delays depend on the geometry of the array. In this work, the size of the grid cells has been set to 0.1 m.
2. GCC computation. For the selected audio frame, the GCC for every pair of microphones is computed following (1).
3. For every point in the grid, the cross-correlation values are accumulated following (4).
4. Selection of the source position. The location $\mathbf{x}$ of the active source is selected as the point of the grid with a maximum value.

The cost computation in the implementation of the algorithm is a key aspect in applications that must interact with a user. Different types of optimization exist in the literature, and in this work a set of equispaced delays is used in step 2 and 3. By this way, instead of computing the GCC for the delays resulting from every combination of grid points and pairs of microphones, it is computed for a set of equispaced delays, with the limit values depending on the size of the array:

$$-\frac{d}{c} \leq \tau \leq \frac{d}{c} \tag{6}$$

being d the diameter of the array. Then, when accumulating the values of cross-correlations in (4), for the estimation of (1) given a delay τ_{kl}, the value is computed as the linear interpolation of the cross-correlation values corresponding to the two closest delays from the set. From now on, the number of equispaced delays will be denoted as n_τ, and its selection has been made based on a set of simulations as explained in SubSection 2.3. For the application of the algorithm given a stream of audio, Hann windows of 25 ms length have been used over frames of 2 s (see Fig. 2). The SRP-PHAT is computed for every window and accumulated through the frame, and the position of the source is selected as the mode of the accumulated values.

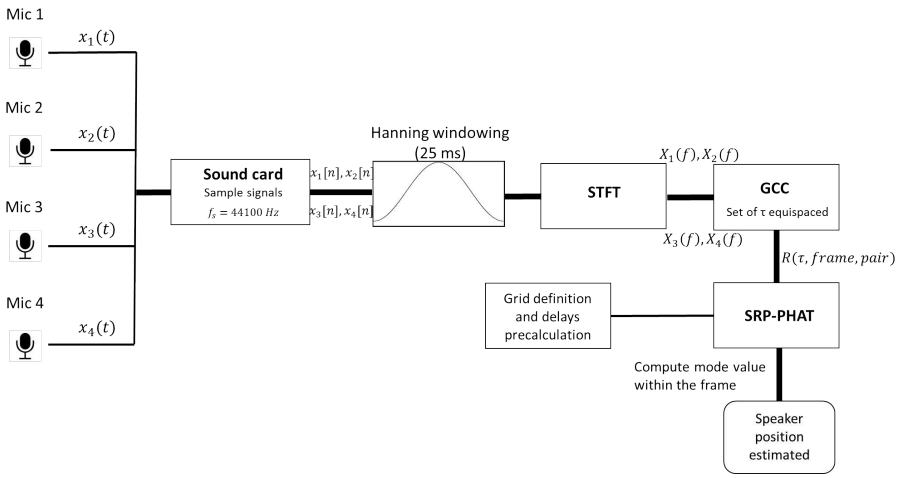

Fig. 2. SRP-PHAT applied to frames of 2 s length with windows of 25 ms.

2.3 MATLAB Simulations

The simulation software RoomSim for MATLAB [16] has been used to decide the initial configuration of the system. This simulator is a useful tool since it allows

to generate the impulse response between an audio source and a microphone in a simulated room with a shoe-box shape. The simulator allows to locate both the source and the microphone in any coordinate (x, y, z) of the space. Therefore, a simulation of an audio signal captured by every microphone of an array can be obtained by modifying the microphone coordinates through the different positions within the microphone array (Fig. 1), and afterwards computing the convolution of every obtained impulse response with an anechoic audio signal.

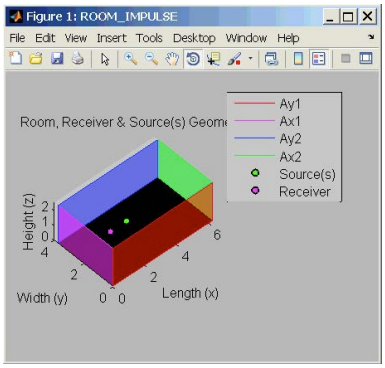

Fig. 3. RoomSim simulator

Based on these simulations, different configurations for the SRP-PHAT implementation have been tested. Since the number of equiespaced delays affect the computation cost of the algorithm, sets with different values for n_τ have been tested: 181, 362 and 543. Simulations have shown that increasing the number above 181 does not improve significantly the accuracy on the detection, and however the computational cost is incremented. Some of the simulations results regarding the detection accuracy are shown in Table 1. The selected configuration has been a squared array of 4 microphones located in the plane $z = 0$ and with a distance between the microphones of 0.3 m. On the other hand, a voice source has been located at the coordinates $(4.0, 4.5, 1.7)$ m. The grid has been limited to a size of (5.5 x 9.0 x 3.3) m with a resolution of 0.1 m for every dimension. It can be seen that the accuracy in the detection is the same for the different sets with a value of n_τ above 181. Also, the simulations show that the accuracy for x and y coordinates is high, but on the contrary, the accuracy in the z coordinate is quite poor. In consequence, the estimation for the azimuth angle, which depends on x and y coordinates, is suitable for directing the attentional mechanism, but for the elevation angle, which depends on y and z coordinates, an alternative way to improve the estimation must be found.

Table 1. Accuracy in the estimation of a simulated speaker at $(4.0, 4.5, 1.7)$ m.

n_τ	x	y	z
181	3.9	4.7	0.2
362	3.9	4.7	0.2
543	3.9	4.7	0.2

3 Face detection

In order to overcome the limitation on the estimation for the elevation angle, the integration of visual information has been used. Since the purpose of the application is to direct the attention towards people speaking, a face detection is applied over the video streaming captured by the camera (Fig. 4). While the control of the azimuth position is done mainly based on the audio information, the visual information is used in order to control the elevation angle, as well to adjust the azimuth value in order to center on the image the face of the person who is detected. For this purpose, Haar cascade classifiers [3] for the detection of both frontal faces and eyes features are applied with OpenCV library.

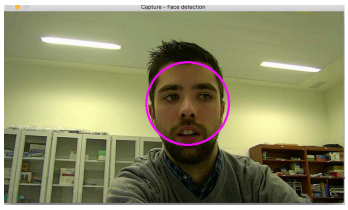

Fig. 4. Face Detection with OpenCV library

4 Results

Initially, an array of four omnidirectional AKG C 417 PP microphones with the same geometry as the one simulated in SubSection 2.3 has been used in a real environment. An experiment has been carried out to test the accuracy in the detection of the azimuth angle where a person has been speaking in different positions respect to the array, at an average distance of 1.75 meters. Also, for a more realistic situation, typical noisy sound sources such as an air conditioner or computer keyboards were active in the room. In order to establish the ground truth for the azimuth values, a laser based angle measurer has been used. While the parameters of the SRP-PHAT implementation have been the same as in the simulations, additionaly a configuration with a lower value for n_τ has been tested

for a lower response time of the attentional mechanism. Table 2 summarizes the results in the estimation for the speaker position, while he/she has been moving around the array. It can be seen that for $n_\tau = 80$ the accuracy of the results is slightly lower, but still valid for the purpose of localizing the speaker, since the error is at most of 2.7°.

Table 2. Accuracy in the estimation of the azimuth angle with a squared array of 0.3 m length. The values represent the average of three measures for each position. The speaker was located at an average distance of 1.75 meters from the microphone array.

Ground Truth	Squared array, $n_\tau = 181$	Squared array, $n_\tau = 80$
0°	0°	0°
30°	29.5°	28.5°
60°	60.5°	59.0°
90°	90.0°	90.0°
120°	118.8°	118.6°
150°	149.2°	149.0°
180°	180.0°	180.0°
210°	208.4°	208.0°
240°	238.4°	237.3°
270°	270.0°	270.0°
300°	299.3°	298.3°
330°	329.3°	329.3°

At a second stage, the algorithm has been implemented with a commercial microphone circular array for embedded devices from XMOS Company (see Fig. 5), that enables developers and equipment manufacturers to add far-field voice capture to consumer electronics and IoT products, and has been qualified by Amazon for the Alexa Voice Service [17]. The array incorporates 7 PDM microphones, 6 of them forming a circle with a diameter of 8.6 cm, and 1 microphone is located at the center. While the device driver allows to capture the raw audio from all the microhopones via USB, for this work only the audio signal captured by 4 of them have been used, with the positions depicted in Table 3. In the implementation of this system, for keeping a lower computational cost, the value for n_τ has been kept to 80. Besides, another XMOS device (xCORE VocalFusion Speaker kit) that incorporates an on-board computation of the azimuth angle of speaker sources has been used for comparison purposes, since it is based on a similar circular array. By this way, the results of the azimuth detection algorithm has been compared with a state-of-the-art commercial product with the same hardware.

A similar evaluation to the previous one has been carried out for these setups, being the results shown in Table 4. It can bee seen that the SRP-PHAT based algorithm is more accurate in the localization task than the on-board built Vo-

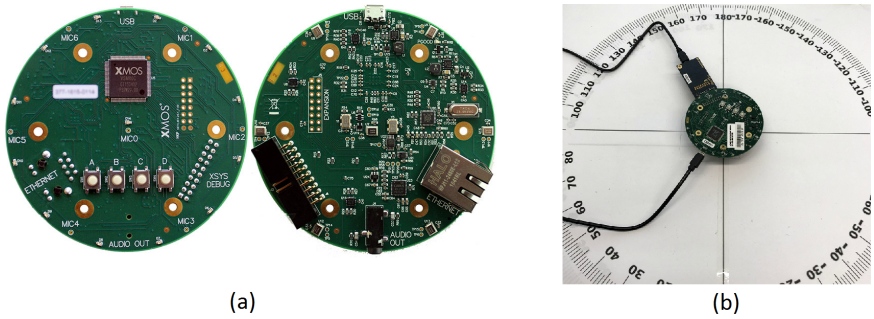

Fig. 5. a) Commercial circular array of 7 PDM microphones from XMOS Company b) Experiments setup

calFusion one. Fig. 6 depicts the error of the estimations respect to the ground truth.

Table 3. (x,y) coordinates of the locations of the 4 microphones used within the XMOS circular array. The origin (0,0) is the center of the array.

	$x(cm)$	$y(cm)$
Microphone #1	3.75	2.15
Microphone #2	-3.75	2.15
Microphone #3	3.75	-2.15
Microphone #4	-3.75	-2.15

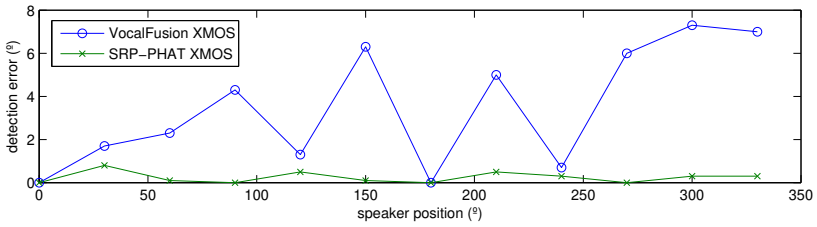

Fig. 6. Absolute value of the azimuth estimation error vs speaker position. The values represent the average of three measures for each position.

Despite the accuracy in the estimation of the azimuth angle, the estimation in the z coordinates was poor in the experiments, and thus the elevation angle is not valid for the attentional mechanism, similarly to the results of the previous

Table 4. Accuracy in the estimation of the azimuth angle. The values represent the average of three measures for each position. The speaker was located at an average distance of 1.75 meters from the microphone array.

Ground Truth	VocalFusion	SRP-PHAT
0°	0°	0°
30°	31.7°	30.8°
60°	62.3°	60.1°
90°	94.3°	90°
120°	121.3°	120.5°
150°	143.7°	150.1°
180°	180°	180°
210°	215°	209.5°
240°	240.7°	240.3°
270°	264°	270°
300°	292.7°	299.7°
330°	323°	329.7°

simulations. This fact makes necessary to use an alternative way to estimate these values, which in this work it has been done through the analysis of video frames, searching for the faces of the speakers. Respect to the video setup, a Pan-Tilt-Zoom network AXIS videocamera has been used (Fig. 7), which can be controlled via HTTP and incorporates a video streaming server.

Fig. 7. AXIS V5914 videocamera

Several experiments in a live situation have been carried out, where it has been tested that the camera is able to aim to the person that is speaking in a room with several speakers. The pan movement is autonomously controlled based on the audio localization, while the face detection is used for an adjustment of both pan and tilt angles in order to center the speaker face in the video frames (Fig. 8). To avoid a continuous readjustment of the camera trying to center the face, a threshold value has been set as a displacement limit for the coordinates of consecutive detected faces before updating the camera. This value has been set experimentally to 20 pixels for a comfortable visualization at the client side of the video streaming. Also, the average computation time for the SRP-PHAT audio localization takes around 1 s using a laptop with an i7 processor. Some relevant

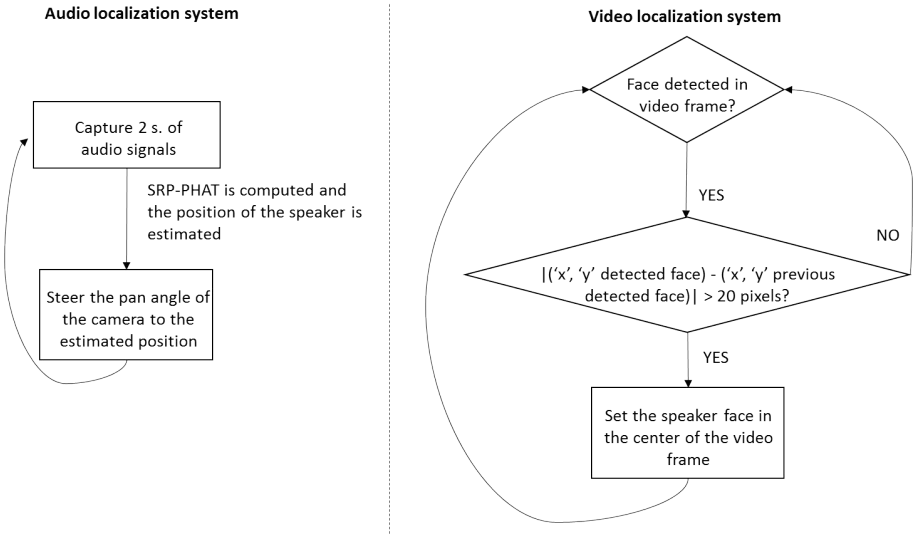

Fig. 8. Pan-Tilt Control of the videocamera

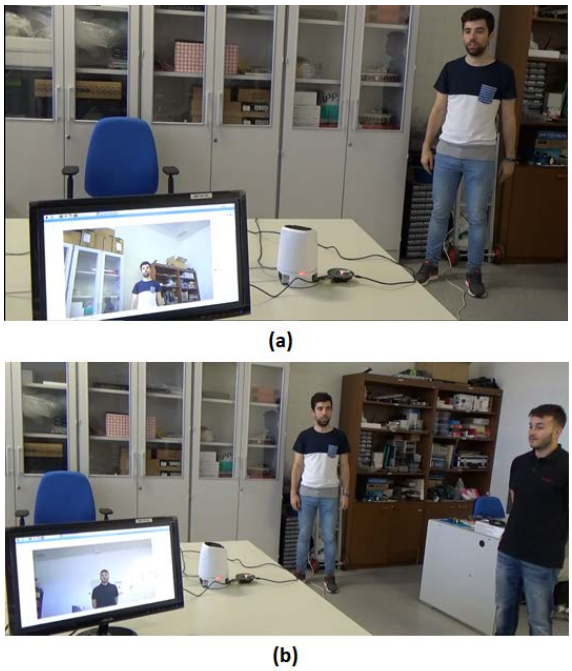

Fig. 9. Experiments. The monitor in the image allows to see the video sent from the camera server to a HTTP client. (a) There is one person speaking in the room, and the camera is pointing towards the speaker (b) A second person starts to speak, and the camera updates its orientation.

Fig. 10. Experiments. The figures depict the images of the monitor screen (a) Speaker has moved towards his left (b) The camera changes its orientation (c) Speaker moves again (d) The camera changes again its orientation

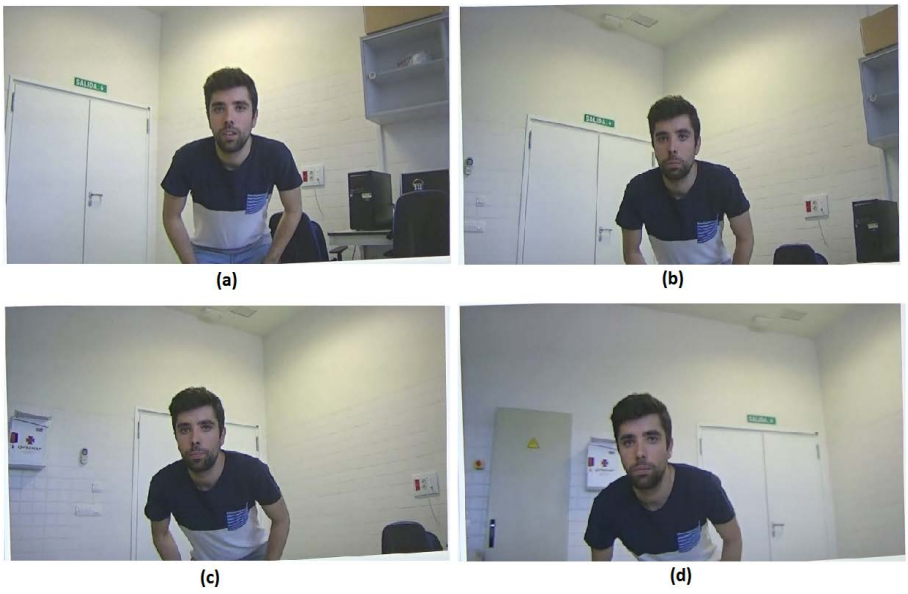

Fig. 11. Experiments. The figures depict the images of the monitor screen (a)-(d) The speaker is moving towards his right side, and the camera corrects its orientation in order to center his face

captures of the experiments are depicted through Fig. 9-11. In Fig. 9.a it can be seen a frame where the camera is pointing towards an active speaker in the room. Then, a second speaker begins to speak in Fig. 9.b and the camera changes its pointing orientation to this second speaker. Fig. 10.a - 10.d try to show the behaviour of the system when a person moves in the room while speaking. In this case, Fig. 10.a and Fig 10.c show a movement of the speaker towards his left side, and Fig. 10.b and Fig. 10.d show how the videocamera moves towards the speaker, even when the speaker is outside the field of view, since the localization is based on the audio information. Finally, Fig. 11.a - Fig. 11.d show a situation where the camera follows the speaker based on the face detection, where the system is trying to get centered the detected face on the screen.

5 Conclusions and future work

In this work an attentional mechanism based on a microphone array for embedded devices has been presented. While SRP-PHAT is a widely used technique with arrays of microphones, its application has been barely reported in arrays of reduced dimensions and with a low number of microphones, being the main contribution of this work. It has been shown that, while it can achieve a high accuracy in the x and y coordinates for estimating the azimuth angle of a sound source, the z coordinate fails in this estimation. To overcome this limitation, a video based face detection is proposed to correct the elevation angle. The experiments have shown the viability of the system in the interaction with people speaking while moving around. Further work will be focused on different items, such as the implementation of a more robust control algorithm based on probabilistic fusion of audiovisual information, the recognition of the type of audio sources in the environment, the incorporation of a zoom functionality to the pan/tilt movement, the application within a hardware for a robotic head, or the integration of the algorithm in a robotic framework such as Robocomp [18].

Acknowledgements

This work has been supported by Economy and Competitiveness Department of the Spanish Government and European Regional Development Fund under the project TIN2015-65686-C5-2-R (MINECO/FEDER, UE).

References

1. Fong, T., Nourbakhsh I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems. 42 (3), pp. 143–166 (2003).
2. DiBiase, J.: A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. Ph.D. Thesis. Brown University. (2000).

3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. 1, 511–518.
4. Valin, J.M.; Michaud, F.; Rouat, J.; Letourneau, D.: Robust sound source localization using a microphone array on a mobile robot. International Conference on Intelligent Robots and Systems (IROS 2003). 2, pp. 1228–1233 (2003).
5. Nakamura, K., Nakadai, K., Asano, F., Ince, G.: Intelligent Sound Source Localization and its application to multimodal human tracking. IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 143–148 (2011).
6. Ferreira, J., Lobo, J., Bessiere, P., Castelo-Branco, M., Dias, J.: A Bayesian framework for active artificial perception. IEEE Transactions on Cybernetics. 43 (2), pp. 699–711 (2013).
7. Viciana-Abad, R., Marfil, R., Perez-Lorenzo, J.M., Bandera, J.P., Romero-Garces, A., Reche-Lopez, P.: Audio-Visual Perception System for a Humanoid Robotic Head. Sensors. 14 (6), pp. 9522–9545 (2014).
8. Do, H., Silverman, H.F., Yu, Y.: A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 07. 1, 121–124 (2007).
9. Do, H., Silverman, H.F.: A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction(CFRC). 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 295–298 (2007).
10. Marti, A., Cobos M., Lopez, J. J.: Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2592–2595 (2011).
11. Silverman, H., Yu, Y., Sachar, J., Patterson, W.: Performance of real-time source-location estimators for a large-aperture microphone array. IEEE Transactions on Speech and Audio Processing. 13(4), 593–606 (2005).
12. Knapp C., Carter G. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing, 1976; 24(4): 320–327.
13. Perez-Lorenzo J.M., Viciana-Abad R., Reche-Lopez P., Rivas F., Escolano J. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. Applied Acoustics 2012; 73(8): 698–712
14. J. H. DiBiase, H. F. Silverman, M. S. Brandstein: Microphone Arrays: Signal Processing Techniques and Applications. M. S. Brandstein and D. Ward, Eds. Springer-Verlag (2001).
15. Marti A. Multichannel audio processing for speaker localization, separation and enhancement. Ph.D. Thesis. Universitat Politècnica de València, 2013
16. Campbell D. R., Palomäki K. J., Brown G. J. A MAT-LAB simulation of shoebox room acoustics for use in re-search and teaching, Computing and Information Systems Journal, vol. 9, 2005
17. http://www.xmos.com/
18. Manso L.J., Bachiller P., Bustos P., Núñez P., Cintas R., Calderita L.V. RoboComp: a tool-based robotics framework. In: Ando N, Balakirsky S, Hemker T, Reggiani M, von Stryk O, editors. Simulation, Modeling, and Programming for Autonomous Robots. Berlin: Springer; 2010:251-261.