Audio-Visual Detection of Multiple Chirping Robots

Alexey GRIBOVSKIY [a] and Francesco MONDADA [a]
[a] *Laboratoire de Systèmes Robotiques (LSRO), Ecole Polytechnique Fédérale de Lausanne, Switzerland*

Abstract. Design, study and control of mixed animals-robots societies are the fields of scientific exploration that can bring new opportunities to study and control groups of social insects or animals and, in particular, to improve welfare and breeding conditions of domestic social animals. Our long-term objective is to develop a socially acceptable by chickens mobile robot able to interact with them using appropriate communication channels. For interaction purposes the robot has to know positions of all the birds in the experimental area and detect those uttering calls. In this paper, we present the audio-visual approach to locate the robots or animals on the scene and detect their calling activity. The visual tracking is provided by the marker-based tracker with help of the overhead camera. Sound localization is achieved by interpolation delay and sum beamforming approach using the array of the sixteen microphones. Visual and sound information are probabilistically mixed to detect the calling activity. Experimental results demonstrate that our system is capable to detect the sound emission activity of multiple moving robots with 90% probability.

Keywords. Microphone arrays, sound localization, audio-visual multi-source information fusion.

Introduction

Study of animal-robot interaction can bring new opportunities to analyse and control groups of social insects or animals and, in particular, to improve welfare and breeding conditions of domestic social animals. After the promising results of the European project Leurre, where a mixed society of cockroaches and robots was created and successfully controlled [1], we focused our interests on more complex social animals – chickens. Our goal is to develop a mobile robot socially acceptable by chickens and able to interact with them.

Vision and hearing are the two most important sources of information for birds. Hence in order to communicate with them, a robot has to effectively use the audio and visual information. One of the tasks here is to use these multimodal data to obtain positions of all birds on the experimental area and to detect among them the birds uttering calls.

A number of robotic auditory systems have been designed where the sound sources localization is achieved by using only two microphones [2]. For sound source localization these methods use an approach inspired by the animal hearing system and based on estimation of the differences in phase and in intensity level between two sensors [3].

However microphone arrays, containing a large number of sensors, provide considerably better results, as they allow to improve the resolution of the localization procedure and its robustness to ambient noise [4]. In [5] a robot equipped with an array the eight microphones localizes and tracks multiple moving speakers with the use of beamforming and particle filtering techniques. In [6] the beamforming and Kalman filters are used for the same task by the SIG2 robot with the array of eight microphones. The method of audio-visual data fusion for human tracking and speaking activity detection was proposed in [7] and tested on the HRP-2 robot equipped with two arrays of eight microphones each. In the biological studies sensor networks, where each sensor includes a microphone array, are applied to detection, recognition and localization of bird songs [8].

In this paper, we present a system for analysis of audio-visual information that serves to determine positions of animals and robots in the experimental area and to detect sound emission activity. The positions of robots and animals are detected by the marker-based visual tracker, while sound localization is provided by interpolation beamforming approach using the array of the sixteen microphones. These results are then probabilistically mixed to detect the calling activity. This system was tested to track the positions and sound emission activity of e-puck robots, playing prerecorded birds calls.

Although the visual tracking and sound localization methods, used in our system are not new and have already been employed in robotics and biological study separately, there are no reported results known to authors where such methods was combined for tracking animals and detecting their calling activity. Thus we believe that this approach provides useful and robust tool for the field of behavioral study and interactions with vertebrates.

The paper is organized as follows. Section 1 presents a brief overview of the system. Sections 2.1 explains how our sound localization system works and section 2.2 presents the visual tracker that we are using. Section 3 describes the probabilistic calls detector, followed by the experimental results in section 4.

1. System Overview

The entire system as shown in Figure 1 is composed of the following parts:

- a microphone array;
- a sound localization system based on the beamforming technique;
- a camera;
- a markers based visual tracking system;
- a probability based calls detection algorithm.

The circular microphone array is formed by the sixteen omnidirectional microphones. The input audio data are processed by the beamformer, that is focused in all possible directions around the microphone array in order to find those, where the output energy is maximal.

The visual tracking subsystem processes the input video data and detects positions of all the object of interest in the experimental area.

Then the calls detector analyzes results of two previous processes and associates the sound sources with the tracked objects.

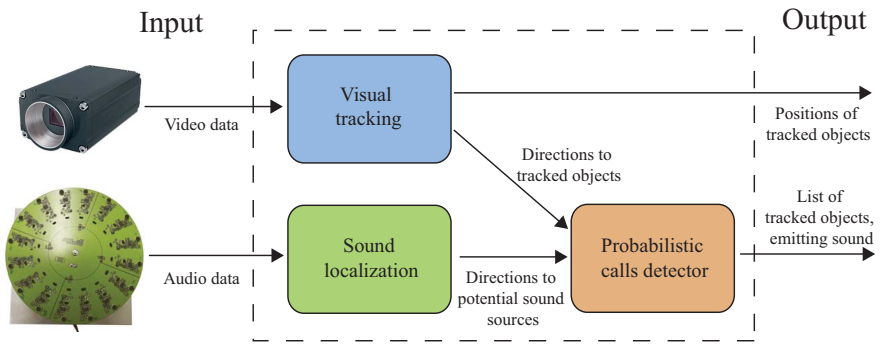


Figure 1. The overview of the system.

2. Features Extraction From Audio and Video Data

2.1. Sound localization

In our experiments we use the array of the sixteen microphones with the diameter of 15 cm (Figure 2). This number of microphones is enough to avoid the spatial aliasing effect, as it satisfies the equation:

$$M \geq \frac{4\pi r f_{max}}{v_s},$$

where M – the number of the microphones, v_s – the speed of sound (345 m/s), f_{max} – the maximal signal frequency (for chickens' calls it is about 6000 Hz [9]), and $r = 7.5$ – the radius of the array.

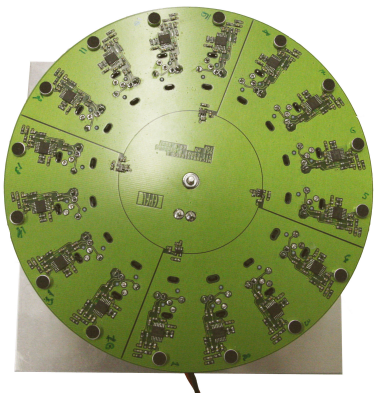


Figure 2. The array of the sixteen microphones (the view from above).

2.1.1. Sound Localization by Beamforming

In beamforming based localization systems microphone arrays are steered to multiple directions and maximums of output signals' energy are then searched. The advantage of this technique over other sound localization methods, such as approaches based on time-difference of arrival information and techniques adopting spectral analysis concepts, is that beamforming can be used for wideband signals and a multi-speaker model [10]. These characteristics make the beamforming especially attractive for mobile robotics applications [5,11,12].

The simplest type of beamforming is the delay-and-sum beamforming method. Let $x_m(t)$ be an input signal for a m-th microphone, $m = 1, 2, .., M$, where M – the number of microphones. Let θ be an arbitrary direction in a 2D space, $\theta \in [0, 2\pi)$. The output of the delay-and-sum beamforming steered to the direction θ has the following form:

$$y^{\theta}(n) = \sum_{m=1}^{M} x_m(n - \tau_m^{\theta}), \tag{1}$$

where τ_m^{θ} – a time of arrival difference between a m-th and the first microphones:

$$\tau_m^{\theta} = \frac{(\mathbf{r}_m - \mathbf{r}_1, \mathbf{e}^{\theta})}{v_s}.$$

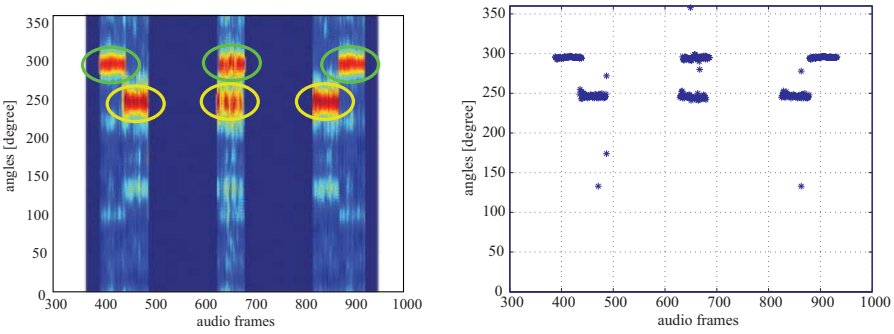
Here $\mathbf{r}_m$ – the position vector of a m-th microphone, $\mathbf{e}^{\theta}$ – a unit vector pointing in the direction θ, and $(\cdot, \cdot)$ denotes the inner product of two vectors. The output energy of the beamformer steered to the direction θ over a frame of the length N is thus given by:

$$E(\theta) = \sum_{i=1}^{N} y^{\theta}(i)^2. \tag{2}$$

If a sound source is located in a direction θ', then the beamformer's output energy $E(\theta)$ will have the maximum at the point θ'.

The main drawback of the delay-and-sum beamforming lies in the sampling procedure: to approximate time delays required for beam steering with reasonable precision signals has to be sampled at a rate much greater than the Nyquist frequency [13]. To overcome this constraint several methods were proposed [13]. One of them is the interpolation beamforming method [14] that we use in our system. In this method the temporal interpolation of the input signal is performed before the beamforming operation, this allows to sample input data at the Nyquist rate. In order to match channels and consequently increase the noise suppression and localization precision of the beamforming we use microphones' gain self-calibration procedure proposed in [15].

Figure 5a shows the result of applying the sound localization system to the sample experimental audio data. Potential directions of sound sources positions can be estimated from maximal peaks of the energy (Figure 5b). After potential sound sources are detected, the set of corresponding directions is transferred to the calls detector subsystem.



(a) The results of applying the sound localization system (ovals mark robots producing sound).

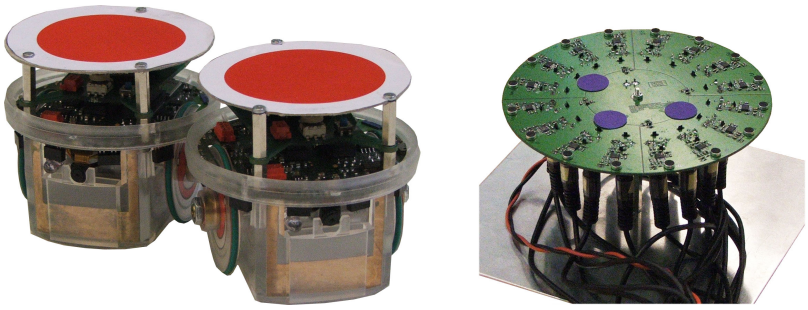
(b) The detected sound sources directions.

Figure 3. The localization of multiple sound sources using the interpolation beamforming.

2.2. Visual Tracking

For visual tracking we use the SwisTrack software [16]. SwisTrack is an open source tracking software that was initially developed for the Leurre project to track cockroaches' and robots' displacements. The advantage of this software is its flexibility: by combining different image processing algorithms it allows to track both marked and marker-less objects. The average absolute calibration error of SwisTrack, reported in [16] is 2.3mm (standard deviation 1.4mm).

For tracking purposes we equipped the two e-puck robots and the microphone array with color markers (Figure 4): each robot with one circular red marker and the microphone array with three blue markers, that allows to obtain both position and orientation of the array.

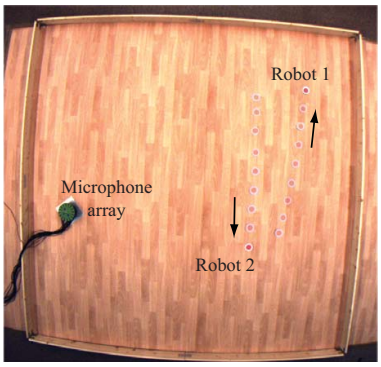


(a) Two e-puck robots with red markers.

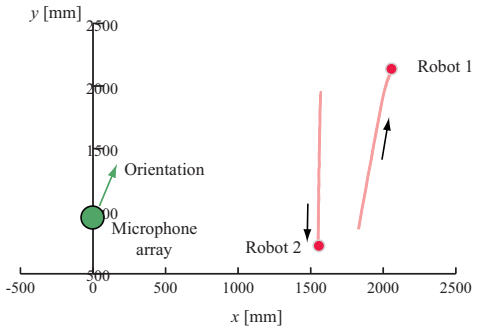
(b) The microphone array with three blue markers.

Figure 4. The marked robots and a microphone array.

The visual tracking subsystem provides us with the positions of all the robots on the arena, as well as with the position and orientation of the microphone array (Figure 5). Then robots' coordinates are converted to the angles, representing the direction to the robots with respect to the microphone array and sent to the calls detector subsystem.



(a) The experimental video data recorded by the overhead camera.



(b) The trajectories of the robots and the position and orientation of the microphone array, obtained by the visual tracker.

Figure 5. The visual tracking of the microphone array and the robots.

3. Probabilistic Calls Detector

The role of the calls detection subsystem is to superpose the acoustical events with detected robots' positions and determine for each time frame the robots emitting sound with a certain probability. Taking into account the fact, that the visual tracking algorithms has better spatial resolution than acoustic localization techniques [17], it is natural to choose the tracked robots' directions as starting points for the analysis.

The goal of the analysis is to find for each tracked robot the probability that it emits sound. We model the probability that a detected sound source corresponds to a given robot as a normal distribution with a experimentally determined variance and mean with respect to the direction of the robot [18].

4. Experimental Results

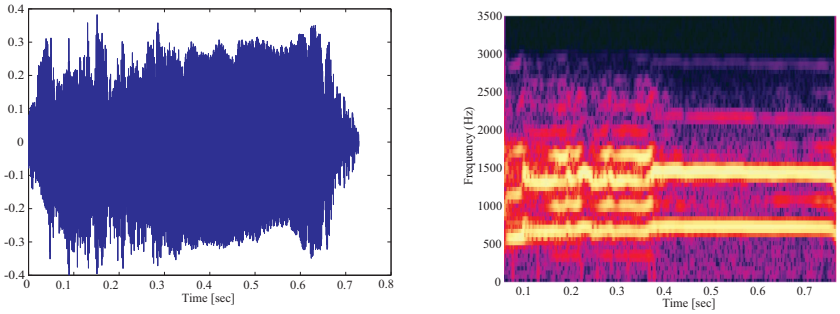
The experimental set-up is composed of the square arena with the sides of three meters in length, the overhead camera and the microphone array placed on the arena (Figure 6). The camera used in our experiments is the color Basler scout Gigabit Ethernet camera scA1000-30gc.

For experiments the two e-puck robots were programmed to emit the prerecorded chicken call (Figure 7): one robot each six seconds and another each five seconds. The emission frequency difference allows us to test robustness of the sound detection for non-overlapped, slightly overlapped and strongly overlapped calls.

All results presented in this paper were obtained through off-line computations using audio and video records of the experiments. We computed the beamformer output energy over a frame of the length $N = 1024$ samples at 44.1 kHz, scanning all directions around the microphone array with the step of 1°. Video data were recorded at 10 fps.

In the first series of experiments we estimated parameters of the normal distribution for the calls detection system. During the four minutes test one marked e-puck robot producing chicken call was placed at four different parts of the experimental arena. Figure 8 shows the histogram of the distribution of resulting differences between the target direc-

Figure 6. The experimental set-up.



(a) The call's waveform.

(b) The call's spectrum.

Figure 7. The prerecorded chicken's call used in the experiments.

tion, obtained by sound localization and visual tracking. The estimated mean is $1.0544°$ and variance is $2.4252°$. The non-zero mean can be explained by minor errors in the marker placement on the microphone array, that entails systematic bias in estimation of the microphone array orientation.

In the series of experiments we evaluated the ability of the system to detect the robots' calling activity. We used the two criteria: detection probability, which we estimate as a fraction of time frames, where the sound emission is correctly detected for the robot, from the whole number of frames of its calling activity; and a false alarm rate that is estimated as a fraction of time frames, where the sound emission is detected but the robot doesn't produce sound, from the whole number of frames of absence of its calling activity.

We performed eight experiments: four with the two stationary robots and four with the two robots, moving on the arena (Figure 9). Each experiment was one minute long. As it can be seen from the Table 1 the calls detection is fairly robust achieving around 90% of detection probability and the false alarm rate around 1%. The difference in detection probabilities between stationary and moving robots can be explained by the limited resolution of the beamformer. If two calls are emitted close to each other, the beamformer produces a single energy peak centered in-between the directions of sound sources and thus the calling activity can not be detected.

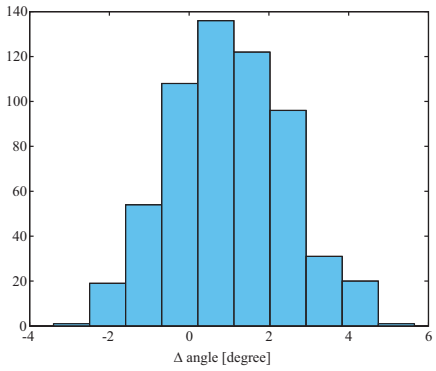
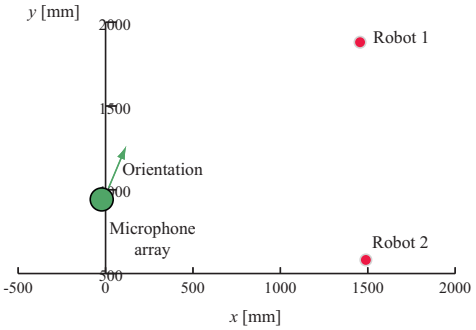
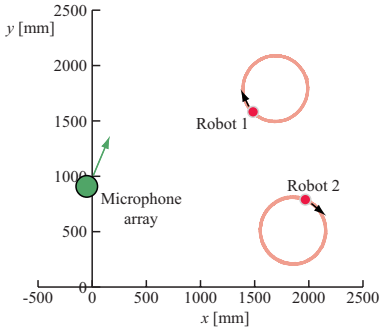


Figure 8. The sound localization error with respect to visual tracking results.



(a) The positions of the robots and the microphone array in the experiment with the stationary robots.



(b) The trajectory of the robots and the position of the microphone array in the experiment with the moving robots.

Figure 9. The location of the robots and the microphone array in the calling activity detection experiment.

Table 1. The results of calling activity detection for stationary and moving robots

Type of experiment	**Detection probability (%)**	**False alarm rate (%)**
Stationary robots	91.45	1.09
Moving robots	86.76	1.24

5. Conclusion

In this paper, we have described the system that is able to localize, using audio-visual information, the positions and calling activities of multiple moving robots with the good accuracy.

Several issues remain open for improvement. First, in the current system we work with the prerecorded data, however, our goal is to make our system work in real time. We are currently developing the twenty four channels audio acquisition board to be able to process the audio data in real time. Second, as our robot has to understand the meaning of chickens calls, the sound separation and birds' calls recognition system has to be

integrated into our system. Finally, the evaluation of our approach on real animals will be a part of future work.

Acknowledgements

This work is supported by the Swiss National Science Foundation grant no. 112150 "Mixed society of robots and vertebrates". We thank Michael Bonani (LSRO) and Daniel Burnier (LSRO) for discussions and technical support.

References

[1] J. Halloy, G. Sempo, G. Caprari, C. Rivault, M. Asadpour, F. Tache, I. Said, V. Durier, S. Canonge, J.M. Ame, C. Detrain, N. Correll, A. Martinoli, F. Mondada, R. Siegwart, and J.-L. Deneubourg. Social Integration of Robots into Groups of Cockroaches to Control Self-Organized Choices. *Science*, 318(5853):1155–1158, 2007.

[2] H.-D. Kim, K. Komatani, T. Ogata, and H. G. Okuno. Auditory and visual integration based localization and tracking of humans in daily-life environments. In Kazunori Komatani, editor, *International Conference on Intelligent Robots and Systems (IROS 2007)*, pages 2021–2027, 2007.

[3] H.G. Okuno, K. Nakadai, K.I. Hidai, H. Mizoguchi, and H. Kitano. Human-robot interaction through real-time auditory and visual multiple-talker tracking. In K. Nakadai, editor, *International Conference on Intelligent Robots and Systems*, volume 3, pages 1402–1409, 2001.

[4] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency – domain steered beamformer approach. In *IEEE International Conference on Robotics and Automation*, volume 1, pages 1033–1038, 2004.

[5] J.-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

[6] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. G. Okuno. Multiple moving speaker tracking by microphone array on mobile robot. In *European Conference on Speech Communication and Technology (Interspeech)*, 2005.

[7] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto. Robust speech interface based on audio and video information fusion for humanoid HRP-2. In F. Asano, editor, *International Conference on Intelligent Robots and Systems (IROS 2004)*, volume 3, pages 2404–2410 vol.3, 2004.

[8] V. Trifa, L. Girod, T. Collier, D.T. Blumstein, and Taylor C.E. Automated wildlife monitoring using self-configuring sensor networks deployed in natural habitats. In *International Symposium on Artificial Life and Robotics (AROB07)*, 2007.

[9] Lesley J. Rogers. *The Development of Brain and Behaviour in the Chicken*. Oxford University Press, 1996.

[10] M. S. Brandstein and D. B. Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[11] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4172– 4177, 2005.

[12] S. Argentieri, P. Danes, P. Soueres, and P. Lacroix. An experimental testbed for sound source localization with mobile robots using optimized wideband beamformers. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2536–2541, 2005.

[13] R. Mucci. A comparison of efficient beamforming algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(3):548–558, 1984.

[14] Roger G. Pridham and Ronald A. Mucci. A novel approach to digital beamforming. *The Journal of the Acoustical Society of America*, 63(2):425–434, 1978.

[15] N. Tashev. Gain self-calibration procedure for microphone arrays. *IEEE International Conference on Multimedia and Expo (ICME 2004)*, 2:983–986, 2004.

[16] N. Correll, G. Sempo, Y. Lopez de Meneses, J. Halloy, J-L. Denebourg, and A. Martinoli. Swistrack: A tracking tool for multi-unit robotic and biological systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8–8, 2006.

[17] D. Gatica-Perez, D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore. Audio-visual speaker tracking with importance particle filters. In G. Lathoud, editor, *International Conference on Image Processing (ICIP 2003)*, volume 3, 2003.

[18] C. Busso, P.G. Georgiou, and S.S. Narayanan. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, 2:II–685–II–688, 2007.