# Integration of Audio and Video Clues for Source Localization by a Robotic Head

Raffaele Parisi, Danilo Comminiello, Michele Scarpiniti, and Aurelio Uncini

DIET Dept., University of Rome "Sapienza", Rome, Italy
raffaele.parisi@uniroma1.it

**Abstract.** In this work the first step of an integration process between audio and video information for the localization of speakers in closed environments is presented. The proposed metod is based on binaural source localization followed by face recognition and tracking and was realized and implemented in a real environment. Some preliminary results demonstrated the effectiveness of this approach.

**Keywords:** Binaural source localization, face detection and tracking, audio and video integration.

## 1 Introduction

Binaural localization consists in estimating the position of a sound source in a generic environment by use of a single pair of microphones. This approach gets inspiration from biological organisms, where the auditive system works by integrating information acquired by the body, the outer ear and the inner ear [1].

Different models of binaural localization are available [2]. A popular approach is based on combined use of *Interaural Level Difference* (ILD) and *Interaural Time Difference* (ITD) [3]. These cues can separately give information about the source position in different range of frequencies and can be fruitfully combined so as to generate an effective binaural localization algorithm [3].

The exploitation of audio signals is just one side of a localization system based on proper integration of audio and video clues. As a matter of fact, in biology the two senses of hearing and vision cooperate in order to augment the information acquired on the surrounding environment. Of course this is a fundamental task, both for hunting and for escaping from hunters.

Some works exist that deal with the fusion of audio and video signals at different levels and for different applications [4] [5] [6] [7] [8]. As far as we know, there are not works explicitely dealing with the topic of integration of binaural audio signals and video signals.

In this paper some preliminary results toward effective integration of audio and video signals in a robotic head are described. Fig. 1 shows the robotic head that was realized in the ISPAMM Laboratory of the DIET Dept. at the University of Rome "La Sapienza". The device is equipped with two omnidirectional microphones and two cameras. Two stepper motors can rotate the head



S. Bassis et al. (eds.), *Recent Advances of Neural Networks Models and Applications*,
Smart Innovation, Systems and Technologies 37, DOI: 10.1007/978-3-319-18164-6_15

and move the eyes. These stepper motors are controlled using the Arduino Uno board. The Arduino Uno is a very common microcontroller board: it has 14 digital input/output pins that can be used to control some external devices, and a USB connection, used to load the control software from a personal computer.

In this preliminary setup, two main tasks were implemented:

1. a binaural source localization procedure. The joint ILD/ITD estimation was employed to localize the speaker in terms of angular distance from the center. The main issue of this approach is the presence of reverberation, that actually reduces the accuracy of the estimation.
2. A face detector/tracking procedure. It is possible to find and to track a human face in images captured by the cameras. In this way it is possible to track the movement of the speaker and to correct the sound localization errors due to reverberation.

Experimental results in a real environment demonstrated the effectiveness of this preliminary idea, as a first step toward a full integration of audio and video information. In the following the main steps of the developed procedure are described.

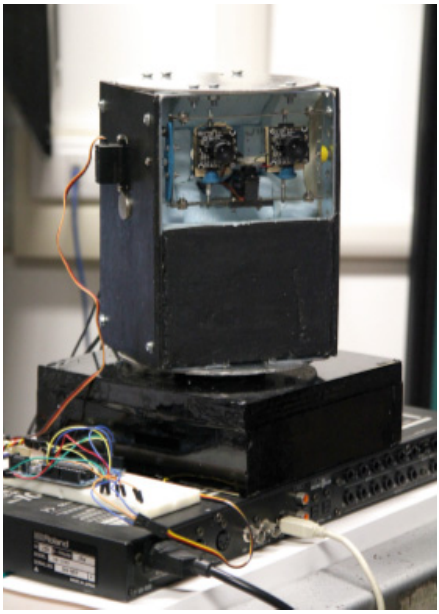

**Fig. 1.** The artificial head described in this paper

## 2 Description of the Audio System

In this section we briefly recall the main concepts of localization of audio sources by binaural processing. Binaural perception was studied by Lord Rayleigh at the

beginning of the $20th$ century [1]. From that time on several models of the human binaural system have been proposed. An estensive description was presented in [9].

Binaural localization can be realized by using the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) in a joint way. ILD is proportional to the difference in the sound levels reaching the left and right ear, while ITD is the measure of the time difference of arrival of a signal to each ear. These cues can be used to obtain information about the source position in different ranges of frequencies. In fact, independent use of ILD and ITD does not yield robust source position estimators [3], since ITD is affected by ambiguity due to an *a priori* unknown phase unwrapping factor, while ILD estimates display a significant standard deviation. Localization of sources can be realized by properly combining ILD and ITD. In the following we briefly describe a possibile approach [3].

The binaural model of received signals is

$$x_l[n] = h_l[n] * s[n] + \eta_l[n], \tag{1}$$

$$x_r[n] = h_r[n] * s[n] + \eta_r[n], \tag{2}$$

where $l$ and $r$ refer to the left and right ear respectively. In this equation $h_i[n]$ $(i = l, r)$ is the impulse response, $s[n]$ is the source signal while $\eta_i[n]$ represent an additive uncorrelated noise term. In the following description noise will be considered negligible, a simplifying assumption which is true in many practical situations.

As in [3], ILD and ITD for the generic $n$-th time-frame are

$$ILD^n(\omega, \theta, \phi) = 20log_{10} \left| \frac{X_r^n(\omega, \theta, \phi)}{X_l^n(\omega, \theta, \phi)} \right|, \tag{3}$$

$$ITD^n(\omega, \theta, \phi) = \frac{1}{\omega} \left( \angle \frac{X_r^n(\omega, \theta, \phi)}{X_l^n(\omega, \theta, \phi)} + 2\pi p \right). \tag{4}$$

In these equations $\omega$ is frequency, $\theta$ and $\phi$ are the elevation and azimuth angles respectively, $X_r^n(\omega, \theta, \phi)$ and $X_l^n(\omega, \theta, \phi)$ are the Short Time Fourier Transforms (STFTs) of the right and left ear signals and $p$ is the *phase unwrapping factor*, which is unknown *a priori* and needs to be estimated.

The new joint ILD and ITD localization method [3] is based on comparison between the particular estimated pair (ILD, ITD) and a reference set of pairs contained in a data lookup matrix. This matrix is constructed by exploiting the fact that Head Related Transfer Functions (HRTFs) are stationary and can be used in calculating two different ITD and ILD reference sets that depend on azimuth and frequency alone. Equations (3) and (4) in this case can be written as

$$ILD(\omega, \phi) = 20log_{10} \left| \frac{HRTF_r(\omega, \phi)}{HRTF_l(\omega, \phi)} \right|, \tag{5}$$

$$ITD(\omega, \phi) = \frac{1}{\omega} \left( \angle \frac{HRTF_r(\omega, \phi)}{HRTF_l(\omega, \phi)} + 2\pi p \right). \tag{6}$$

In these equations $HRTF_r$ and $HRTF_l$ are the HRTF functions on the right and left ears respectively. By assumption the value of the unwrapping factor $p$ does not change dramatically across azimuth [3]. Smoothing across azimuth with a constant $Q$ filter was performed on the ILD lookup set in order to better represent the limits of human interaural level difference perception. More specifically, a Gaussian filter was employed, as indicated in the CIPIC database [10].

Comparison between the ILD and ITD lookup sets and the estimated ILD and ITD allows to estimate the azimuth of the sound source. In particular ILD is exploited to find the correct value of the unwrapping factor $p$ and to select the azimuth value minimizing the difference between the ITD-only and ILD-only estimates. This $p$-estimation procedure was repeated for each available time frame. A time average across frames was performed and the results graphed. The final azimuth estimations selected were those displaying a minimum in the difference function that was consistent across frequencies.

As an example, fig. 2 shows the results obtained in simulations with the source placed at different azimuth angles. Joint exploitation of ILD and ITD allows to obtain an azimuth estimate which is correct over the whole frequency band and for different positions of the source.

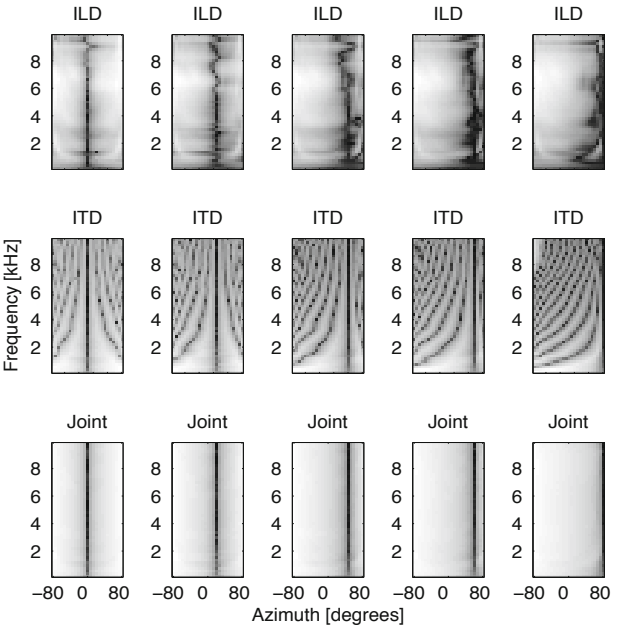


**Fig. 2.** Source azimuth estimate in an anechoic room and Gaussian noise: ILD, ITD and joint ILD-ITD methods. Columns from left to right refer to source azimuth angles of 0°, 20°, 45°, 60° and 80° respectively. Darkest pixels are lowest in value.

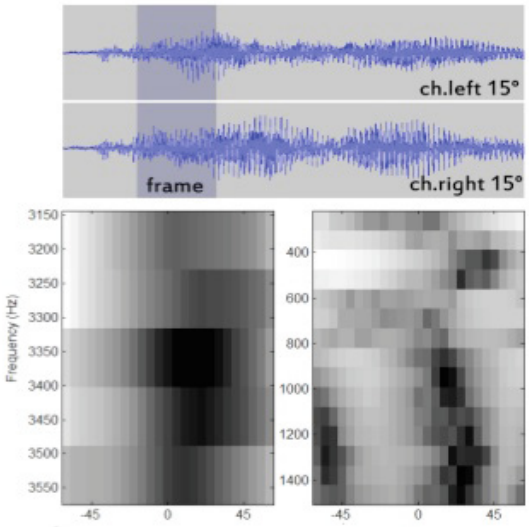


**Fig. 3.** Source azimuth estimate in a real room and a female speaker placed at the azimuth angle of 15°: ILD and ITD estimates in different ranges of frequencies.

Fig. 3 shows the results obtained in a real environment with a female speaker, speaking from an azimuth angle of 15°, in terms of the ILD and ITD estimates in different ranges of frequencies. Slight reverberation is present. It is clear that in the presence of reverberation [11], commonly assumed in closed environments, proper prefiltering techniques should be adopted [12] [13] [14]. An example is cepstral prefiltering [15].

## 3 Description of the Video System

In this preliminary study, the video information was used for localizing and tracking the head of a speaker, after she/he has been localized by using the audio information. The main task in this process is the localization of the face of the speaker in the image acquired.

### 3.1 Face Detection

The face recognition task was realized by using the *Viola-Jones method* [16]. This technique was one of the first methods introduced for detecting the presence of objects in images and it is currently used for the detection of faces. It is based on classification of specific features rather than on the intensity values of the image pixels. Namely the steps of the classification process are:

1. extraction of *Haar features*. Haar features are basically determined by computing the sum and/or the differences of the pixels within two rectangular regions of the image.

2. Construction of the *integral image*. The integral image is an intermediate representation of the original image. Namely, the generic point $(x, y)$ of the integral image is defined as the sum of the pixels above and to the left of $(x, y)$.
3. *AdaBoost.* The AdaBoost (short for *Adaptive Boosting*) is a machine learning meta-algorithm used to improve the performance of learning algorithms [17]. It is based on the combination of various weak classifiers in order to obtain a final robust classifier and it is employed in the Viola-Jones method
4. *Chain classifier.* The Viola-Jones method is based on a cascade of AdaBoost classifiers in order to classify portions of images. As a consequence of this processing phase, the performance of the detection task is increased, while reducing the computation time required.

### 3.2 Face Tracking

Once the region containing the face has been detected, the next step is to move the image of the face to the central position of the video image. This task can be realized by a feedback loop where a pair of proportional controls is employed to progressively reduce the difference between the position of the detected face and the center of the video image. To this goal, the *tilt* and *pan* angles of the head are used. Figure 4 shows the scheme of the head control unit.

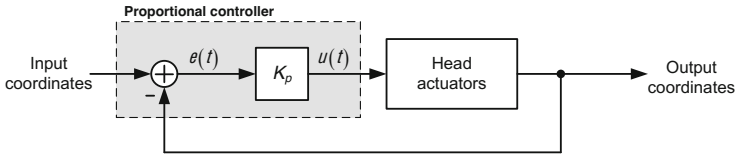


**Fig. 4.** The head control loop

## 4 Experiments

The head was equipped with two omnidirectional microphones AKG C562M. Signals were acquired through an Edirol UA-1000 acquisition board. Figure 5 shows the configuration of the testbed, with five possible positions of the source.

The control of the servomotors was realized by an Arduino board[1]. The face-tracking algorithm was written in C++ by using the functions available at the OpenCV website[2]. Figure 6 shows in detail the Arduino board used for processing of the video part.

The face recognition and tracking algorithm was used to localize the face of the speaker after the audio localization task and to move it to the center of the

[1] `www.arduino.cc`
[2] `www.opencv.org`

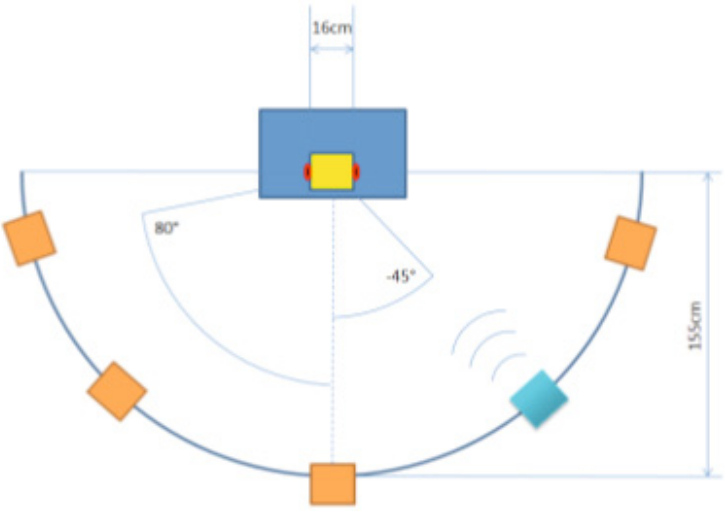


**Fig. 5.** Testbed configuration

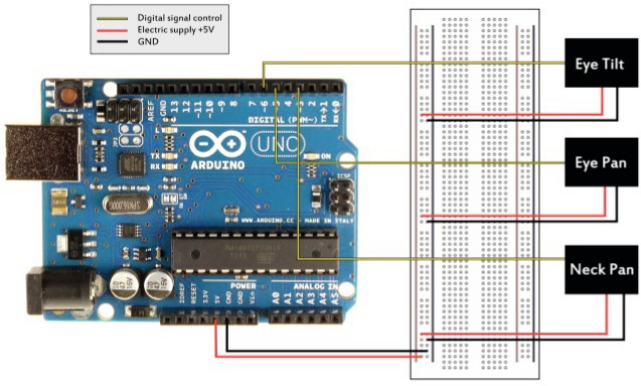


**Fig. 6.** The Arduino board and its connections

**Fig. 7.** A single frame of the localization process

image. As an example of the experimental results, fig. 7 shows a single frame of the output video. The artificial head can be seen in the upper part of the image, together with the speaker. In the lower part, it is shown the image as acquired by the camera mounted on the head, after the face of the speaker has been moved to the center.

## 5 Conclusion

In this paper a possible cooperation between binaural audio and video signals was described. The objective was the localization and tracking of an audio source moving in a closed environment. Some preliminary experiments demonstrated the quality of the proposed solution, also in a real environment. Further research will be devoted to pursue full integration of both audio and visual information.

## References

1. Rayleigh, L.: On our perception of sound direction. Phil. Mag. 13, 214–232 (1907)
2. Blauert, J.: Spatial Hearing - The Psychophysics of Human Sound Localization. MIT Press (1996)
3. Raspaud, M., Viste, H., Evangelista, G.: Binaural source localization by joint estimation of ILD and ITD. IEEE Trans. on Audio, Speech and Language Processing 18(1), 68–77 (2010)
4. Monaci, G., Jost, P., Vandergheynst, P., Mailé, B., Lesage, S., Gribonval, R.: Learning multimodal dictionaries. IEEE Trans. on Image Processing 16(9), 2272–2283 (2007)

5. Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., Sun, X., Pinto, N., Zhang, Z.: Boosting-based multimodal speaker detection for distributed meeting videos. IEEE Trans. on Multimedia 10(8), 1541–1552 (2008)
6. Schmalenstroeer, J., Haeb-Umbach, R.: Online diarization of streaming audio-visual data for smart envirnments. IEEE Journ. of Selected Topics in Signal Processing 4(5), 845–856 (2010)
7. Naqvi, S.M., Wang, W., Khan, M.S., Barnard, M., Chambers, J.A.: Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking. IET Signal Processing 6(5), 466–477 (2012)
8. Minotto, V.P., Jung, C.R., Lee, B.: Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmms. IEEE Trans. on Multimedia 16(4), 1032–1044 (2014)
9. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis - Principles, Algorithms, and Applications. IEEE Press, Wiley Interscience (2006)
10. Algazi, V.R., Duda, R.O., Thompson, D.M., Avendano, C.: The CIPIC HRTF database. In: 2001 IEEE Workshop on Applications of Digital Signal Processing to Audio and Acoustics (2001)
11. Kuttruff, H.: Room Acoustics, 4th edn. Taylor & Francis (2000)
12. Stéphenne, A., Champagne, B.: A new cepstral prefiltering technique for estimating time delay under reverberant conditions. Signal Processing 59(3), 253–266 (1997)
13. Parisi, R., Gazzetta, R., Di Claudio, E.: Prefiltering approaches for time delay estimation in reverberant environments. In: Proceedings of ICASSP, vol. 3, pp. III-2997–III-3000 (2002)
14. Zannini, C.M., Parisi, R., Uncini, A.: Binaural sound source localization in the presence of reverberation. In: Proc. of the 17th International Conference on Digital Signal Processing (July 2011)
15. Parisi, R., Camoes, F., Scarpiniti, M., Uncini, A.: Cepstrum prefiltering for binaural source localization in reverberant environments. IEEE Signal Processing Letters 19(2), 99–102 (2012)
16. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. of Computer Vision 57(2), 137–154 (2004)
17. Freund, Y.Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)