University of Groningen

A real-time system for audio source localization with cheap sensor device

Saggese, Alessia; Strisciuglio, Nicola; Vento, Mario; Petkov, Nicolai

Published in:

2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017

DOI:

[10.1109/AVSS.2017.8078461](https://doi.org/10.1109/AVSS.2017.8078461)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Saggese, A., Strisciuglio, N., Vento, M., & Petkov, N. (2017). A real-time system for audio source localization with cheap sensor device. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017* [8078461] Institute of Electrical and Electronics Engineers Inc.. <https://doi.org/10.1109/AVSS.2017.8078461>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

A real-time system for audio source localization with cheap sensor device

Alessia Saggese¹, Nicola Strisciuglio², Mario Vento¹, Nicolai Petkov²

¹ University of Salerno - DIEM, Italy

² University of Groningen - JBI, The Netherlands

asaggese@unisa.it, n.strisciuglio@rug.nl

Abstract

We propose an architecture for real-time audio source localization based on the integration of localization methodologies within a framework that employs a cheap acquisition sensor. The architecture that we present takes as input the audio signals from two calibrated microphones. Then, it computes biological-inspired features of the sound signal and estimates its direction by means of a Gaussian Mixture Model estimator. We carried out an extensive experimental analysis on four data sets, one of which we realized and made publicly available. We evaluated several characteristics of the sound localization architecture and its use in real scenarios.

1. Introduction

When we perceive sounds, we simultaneously identify their direction and are able to recognize the type of sound. From psychological studies we know that humans feel more comfortable when the sound source (e.g. a speaker) can be located accurately. Generally, in the case of surveillance applications, we can say that a system that knows the source of a sounds provides better information to improve the safety of the environment.

In the last years, applications of audio source localization received a growing interest from research and industry in the field of intelligent audio surveillance with the aim of localizing the sound source of hazardous events. Audio analytic systems can be deployed together with and are a complementary tool to existing video surveillance infrastructures. Many IP surveillance cameras are, indeed, already equipped with or ready to be connected to a microphone, making possible a combined analysis of audio and video streams [20].

Two important applications in intelligent audio surveillance are abnormal event detection [3, 4] and sound source localization [13]. A comprehensive review of methods for audio surveillance has been recently published [2].

Existing methods for sound source localization can be

organized in two groups depending on the acquisition device that they employ. In the first group methods for binaural localization are present, which use two microphones as input devices. Such methods are based on the computation of bio-inspired cues: the interaural level difference (ILD) and the interaural time difference (ITD). The former measures the decay of the volume of the sound with distance. It also considers the absorption effect introduced by the head that stops the propagation of the sound waves [23]. The latter concerns the differences in the arrival time of the sound waves between the two microphones. The computation of ILD is generally approached by considering the energy ratio between the two audio streams at frequency higher than 1500 Hz [1]. The computation of ITD is instead based on the evaluation of the cross-correlation of the two audio streams at frequency lower than 1500 Hz [14]. The estimation of the auditory cues is done with a probabilistic approach in [23] and [15] or by comparing input values with theoretically generated curves in [7]. The latter approach shows less flexibility and lower robustness to noise with respect to the former ones. Other methods integrate audio and visual information to improve the reliability of localization [18, 8]. In [17, 19], auditory epipolar geometry, inspired by epipolar geometry in stereo vision, was introduced. An artificial human-like dummy head has been proposed for binaural hearing in telepresence operations [21].

In the second group, methods that employ array of more than two microphones are collected. Such methods aim at improving the reliability of binaural localization approaches and are designed to perform 3-dimensional localization. A localization system that uses 4 microphones arranged on the surface of a spherical robot head was suggested in [9]. The time-delay of arrival was computed on 6 pairs of microphones and the location estimations were combined so as to improve the reliability of the decision. In [22], the estimation of ITD is performed by cross-correlation of the audio signals for an array of 8 microphones in order to perform 3-dimensional sound source localization. A study about the optimization of the number of microphones in the array was

Figure 1: Architectural overview of the localization system. (a) A Kinect sensor (b) captures the audio signal. The localization method (c,d,e,g,h) is composed of the modules represented within the two pairs of straight lines, which perform a sound direction estimation at frame level. A (f) azimuth estimator integrates the frame-level decisions over a time window W_t and the results are presented to the user through a result visualization module.

presented in [6]. The authors exploit redundancy of spatial information to scale the system linearly with the number of microphones. Beamforming techniques and Kalman filter have been used to reduce the errors in tracking multiple moving speakers in noisy and echoic environments [16]. Such methods are more accurate and robust to noise than binaural localization methods, but require larger hardware resources for processing the input signals.

In this paper, we propose a real-time architecture for sound source localization based on a cheap microphone array, namely the audio acquisition card provided together with the Kinect sensor. Our contribution is in the design and implementation of a system for hardware-software integration of the estimates at frame-level provided by the binaural sound source localization method proposed in [15]. The proposed architecture is modular and, in principle, can be used together with any binaural localization method. The real-time implementation of the system and the use of a cheap audio sensor give the possibility of deploying large installations of the localization system with reduced costs. Furthermore, the proposed solution provides means for distributed analysis of sound sources and eventually to track them within the monitored areas. We evaluate the performance and the characteristics of the proposed architecture in different conditions by carrying out experiments on four data sets, namely the Surrey, Oldenburg, Aachen and MIVIA data sets (the final name will be released after anonymous submission and revision). In this work, we constructed and made public available for benchmark purpose the OUR localization data set¹.

The paper is organized as follows: in Section 2 we present the proposed hardware and software architecture, while in Section 3 we describe the data sets that we used

and discuss the results that we achieved in the experiments. Finally, we draw conclusions in Section 4.

2. System Architecture

In Figure 1, we depict the architectural overview of the proposed system. The presented framework integrates a cheap sensor device, namely the microphone array of the Kinect sensor, with a localization method based on probabilistic estimation of sound directions [15]. Together with the processing architecture, we propose an integration rule of the short-time estimations that improves the reliability of the overall system.

2.1. Localization method

The considered localization method is inspired by how the human auditory system processes sounds. The *cochlea* membrane in the inner ear vibrates according to the energy contained in frequency sub-bands of the input sound. In order to model this behavior, a Gammatone filterbank was adopted. As suggested in [5], the input signals to each microphone are decomposed into $N_c = 32$ auditory channels using a fourth-order gammatone filterbank, where the channel center frequencies are distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. Successively, binaural cues are estimated using a rectangular window of 20ms at a sampling frequency of $f_s = 44.1\text{kHz}$. The impulse response $g_{f_c}(t)$ of a Gammatone filter at the time t and central frequency f_c is:

$$g_{f_c}(t) = D t^{\eta-1} \cos(2\pi f_c t + \phi) e^{-2\pi t b(f_c) u(t)}, \quad (1)$$

where D is an arbitrary scaling constant, η the order of the filter (fourth in our case), ϕ is the phase, $u(t)$ is the unit step function (1 for $t > 0$ and 0 otherwise), and $b(f_c)$ is a

¹The data set is available at the url <http://OUR.->

function which determines the bandwidth for a given center frequency. It is formally defined as: $b(f_c) = a(24.7 + 0.108f_c)$, where a is a proportionality constant.

Given the response of the Gammatone filterbank for both microphones, the the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) are computed. ILD and ITD contain complementary information about the sound source position with respect the microphone array. Thus they are combined into a two-dimensional feature vector, called binaural vector, in order to represent the single frame within a specific Gammatone channel.

Finally, a Gaussian mixture model (GMM) is used to estimate the position of the sound source from the set of binaural feature vectors. In particular, considering that the binaural features corresponding to different channels tend to cluster in the feature space, the azimuth-dependent pdf is modeled by summing up the superimposed Gaussian components. During a preliminary training phase, a set of training sounds is used to learn the GMM model, which will be used during the operating phase for estimating the source direction. For more details about the localization method we refer the reader to [15].

2.2. Integration of azimuth estimation

The GMM provides an estimation of the direction of the sound source for small overlapped (50% of their length) frames of the input audio signal of duration 20ms. The combination of the estimation for the N channels of the Gammatone filterbank provides a reliable measure of short-time characteristics of the input signal. We integrate the short-time estimations on a larger time scale, using a sliding time window of length W_t that forward shifts on the audio signal by half of its size. This allows for more robust estimations to noise and outliers. The direction of the input sound is thus taken as the one of the audio frame with the highest estimated GMM likelihood within the time window W_t .

2.3. System implementation

The proposed architecture acquires the audio signal from the microphone array of the Microsoft Kinect sensor (Figure 1a). The software module for hardware interface and data acquisition (Figure 1b) is developed in C++ and communicates with a Matlab implementation of the method for the estimation of the sound source direction (Figure 1c,d,e,g,h) at frame-level through a bridge interface. The final estimated direction is obtained by integration of frame-level decisions (Figure 1f) and the result is made available by a visualization module (Figure 1i).

Figure 2: Setup of the kinect sensor for the realization of the MIVIA data set.

3. Experimental analysis

3.1. Data sets

We performed experiments on four publicly available data sets, namely the Surrey [10], Oldenburg [12], Aachen [11] and MIVIA data sets. The first three data sets contain sound sources at different angles and distances from the microphone array and sounds recorded in environments with different reverberation.

The **Surrey data set** was recorded by a Cortex Instruments Mk.2 Head and Torso Simulator (HATS). The loudspeakers were placed around the HATS on an arc with a 1.5 m radius between -90° and $+90^\circ$. The sounds were recorded at intervals of 5° and four different room configurations were considered. For each room a reverberation time RT_{60} was estimated according to BS EN ISO 3382 standard.

The **Oldenburg data set** [12] was recorded by a HATS in five semi-controlled environments.

The **Aachen data set** [11] was recorded in two rooms, with sound sources at a distance up to 10m from the microphones. The characteristics of the rooms for such data sets are reported in Table 1.

We recorded the **MIVIA data set** and made it available for research purposes. It was acquired with a double aim: (1) to evaluate the performance of localization systems with cheap and less accurate devices, instead of very expensive devices which are more difficult to be used in real applications; (2) to evaluate the performance of the proposed approach by varying the distance between the microphones. We considered two different environments, namely a living room $6 \times 4 \times 2,7$ m (M.LR) and a laboratory of our university $8 \times 8 \times 7$ m (M.L). The sounds were recorded by using the microphone array of a Kinect sensor. The hardware set was expanded so as to build an array of four microphones, where the maximum distance between them is 138 cm, as shown in Figure 2. Note that the distance between the two central microphones were fixed to 23 cm, as in the original setup of the Kinect sensor. The recording was made by considering the following acquisition angles: -90° , -60° , -45° , -20° , 0° , 20° , 45° , 60° , 90° . As for M.LR, a man repeats the same word aloud for 10 seconds at a distance of 2 m from the microphones. In the second scenario, a Bose loud-

Surrey data set

S.AR	the room size was $17.04 \times 14.53 \times 6.5$ m (l x w x h) and the anechoic conditions were simulated by truncating the first reflected waves
S.A	the room is a small sized ($5.72 \times 6.64 \times 2.31$) office with seats for 8 persons. The reverberation time is $RT_{60} = 320$ ms
S.B	medium sized classroom ($4.65 \times 9.6 \times 2.68$) with $RT_{60} = 470$ ms
S.C	large sized room ($23.5 \times 18.8 \times 2.31$) used as cinema hall for more than 400 people with $RT_{60} = 680$ ms
S.D	a medium sized room ($8.72 \times 8.02 \times 4.25$) for presentations with high ceiling and $RT_{60} = 890$ ms

Oldenburg data set

O.AR	Anechoic chamber; the HATS was mounted in front of the speaker, with a distance varying from 0.8m to 3m. The azimuth angle of the source ranged in the interval $-90^\circ, +90^\circ$, with a reverberation time $RT_{60} < 50$ ms
O.O1	Office room (3.20×4.55 m) The HATS was mounted on a desk close to the center of the room, while the speaker was moved in the front hemisphere $-90^\circ, +90^\circ$ at a distance of 1 and m with an elevation angle of 0° . The reverberation time $RT_{60} = 300$ ms
O.O2	Small office room (3.30×6.00 m) by using the same setup used for O.O1. The door and the window were left opened.
O.OC	Busy cafeteria at lunch time. The HATS was moved in different positions, with an average reverberation time of 1250 ms and an average SNR of 75.6 dB.
O.CY	Courtyard crossed by pedestrians and bicycles. The setup is similar to the one in the cafeteria, with an average reverberation time of 900 ms and an average SNR of 86.1 dB

Aachen data set

A.MR	Medium sized room $8.00 \times 5.00 \times 3.10$ m suitable for a meeting. The distance between the loudspeaker and the receiver ranged from 1.45 m up to 2.80 m, with an average reverberation time of 0.23 s
A.LR	Lecture room whose size is $10.80\text{m} \times 10.90\text{m} \times 3.15\text{m}$. The loudspeaker was placed in different positions, with a distance ranging from 4 m up to 10.20 m. The average reverberation time is 0.78 s.

Table 1: Description of the rooms where the sounds contained in the data sets were recorded.

speaker was used to reproduce part of a song (10 seconds for each recording) at two distances from the array center: 1 m and 3 m.

3.2. Performance evaluation

We organized the experiments in four groups, so as to evaluate the performance of the system when the sound source is (1) at different angles and (2) different distances from the microphones, (3) different SNR values and (4) different configurations of the binaural microphone set. Each group of tests involves specific rooms from the four data sets, which we report in the following:

- **TEST1:** S.AR, S.A, S.B, S.C, S.D from the Surrey data set and the rooms O.AR and O.O1 from the Oldenburg data set.
- **TEST2:** A.MR and A.LR from the Aachen data set.
- **TEST3:** O.O2, O.C and O.CY from the Oldenburg data set.
- **TEST4:** M.LR and M.L from the MIVIA data set.

We measured the performance of the real-time sound localization system by computing the accuracy (Acc) and the mean absolute error (MAE) of the angle estimation, as follows:

$$Acc[\%] = 100 - \frac{N_e * 100}{N}, MAE = \frac{\sum_{i=1}^N |x_i - \hat{x}_i|}{N}$$

where N is the number of considered sounds in a particular test and N_e is the number of wrongly estimated angles. For the computation of the MAE, x_i is the estimated direction while $\hat{x}_i$ is the nominal one.

3.3. Results discussion

In the analysis of the performance of the real-time localization method we considered several tolerance values for the localization error. Specifically, we considered different margins for maximum acceptable error on the estimation of the sound source position. A sound source is correctly localized if the estimated angle falls into a tolerance interval around the nominal value of the sound direction. The value of the tolerance degree influences both accuracy and MAE. Indeed, only wrongly estimated directions are considered for the computation of the MAE.

For the experiments in the groups “TEST1” and “TEST2” we performed a quantitative analysis of results by computing the accuracy and MAE. In Table 2 and Table 3 we report the results that we achieved for the different environments in “TEST 1” and “TEST 2”, respectively. We observed generally high performance and precise estimations. The results decrease for highly noisy environments.

In the experiment “TEST3”, we evaluated the performance of the system in environments with several values of signal-to-noise ratio. For the rooms in the Oldenburg data set, precise measures on the nominal sound source direction are not provided, but rather only an indication of the direction of the sound. In Table 5, we report the angle estimated

Test 1				
		Tolerance		
Room		5°	10°	15°
S.AR	Acc	100%	100%	100%
	MAE	0°	0°	0°
S.A	Acc	78.37%	100%	100%
	MAE	1.08°	0°	0°
S.B	Acc	81.08%	97.29%	100%
	MAE	1.08°	0°	0°
S.C	Acc	86.48%	94.59%	94.59%
	MAE	2.83°	2.43°	2.43°
S.D	Acc	51.35%	70.27%	70.27%
	MAE	14°	13°	13°
O.AR	Acc	97.29%	100%	100%
	MAE	0°	0°	0°
O.O1	Acc	48.64%	70.27%	75.67%
	MAE	11.41°	10.33°	9.79°

Table 2: Performance results for the group of experiments “TEST 1” with different values of the error tolerance interval.

Test 2				
		Tolerance		
Room		5°	10°	15°
A.MR	Acc	100%	100%	100%
	MAE	0°	0°	0°
A.LR	Acc	50%	100%	100%
	MAE	3.53°	0°	0°

Table 3: Performance results for the group of experiments “TEST 2” with different values of the error tolerance interval.

by the real-time localization system together with the indication of the direction of the sound. For each room, various positions of the sound source are considered.

For the experiments “TEST4”, we evaluated the performance of the method when the physical configuration of the microphone set changes. In particular we compared the estimated direction when the pair of microphones from the Kinect sensor are placed at a distance of 23cm (CONF1) and 138cm (CONF2). We observed a decrease of the performance of the localization system for sound sources at directions farther than 60° with respect to the frontal direction. This is mainly due to the directionality of the microphones of the Kinect sensor. When considering the configuration CONF2, we observed that the reliability of the direction estimation decreases. The model proposed in [15] is constructed by taking inspiration from the functions of the human auditory system, which are not effectively applicable for large binaural inter-distances.

Furthermore, we studied the average required processing time with respect to different sizes of the time windows W_t . We report in Table 6 the processing time required by each block of the architecture and the total processing time of the

Test 3			
Room	Configuration	Expected direction	Estimated angle
O.O2	1C	Frontal	−5°
O.O2	1D	Frontal	5°
O.O2	1B	Right	35°
O.O2	2A	Right	15°
O.O2	2B	Left	−10°
O.C	1A	Frontal	0°
O.C	2D	Frontal	5°
O.C	2E	Right	60°
O.C	1D	Right	90°
O.C	1B	Left	−30°
O.C	1C	Left	−90°
O.CY	1A	Right	35°
O.CY	1B	Right	10°
O.CY	1C	Left	−20°
O.CY	1D	Left	−50°
O.CY	1E	Left	−60°

Table 4: Qualitative analysis of the results for the group of experiments “TEST 3”.

Test 4			
	Expected Angle	CONF1	CONF2
Living Room (M.LR)	−90°	−58°	−45°
	−60°	−45°	−33°
	−45°	−40°	−37°
	−20°	−26°	−21°
	0°	0°	0°
	20°	23°	19°
	45°	30°	37°
	60°	55°	40°
	90°	56°	55°
Laboratory (M.L)	−90°	−62°	−62°
	−60°	−55°	−52°
	−45°	−40°	−38°
	−20°	−23°	−21°
	0°	0°	0°
	20°	20°	19°
	45°	44°	41°
	60°	57°	55°
	90°	56°	62°

Table 5: Qualitative analysis of the results for the group of experiments “TEST 4”.

system. Block #1 refers to the acquisition module, blocks #2 and #3 are the features computation and frame-level angle estimation, while block #4 refers to the integration module. We used audio chunks of 4 seconds to be analyzed and observed that in order to obtain real-time response on a machine with an Intel i5 CPU, a reasonable choice for the size of W_t is 0.8 seconds.

The real-time response and the non-expensiveness of the audio sensor are key factors for the deployment of the proposed system in real scenarios. The sound directions estimated by several sensors can be combined together in order to obtain more precise localization and to eventually track

Time analysis					
W_t	#1	#2	#3	#4	Total
0.8	0.51	0.5	2.63	0.002	3.642
1.1	0.51	0.5	2.92	0.002	3.932
1.4	0.51	0.5	3.24	0.002	4.252
1.7	0.51	0.5	3.56	0.002	4.572
2.0	0.51	0.5	3.97	0.002	4.982

Table 6: Contribution to the processing time of the modules of the architecture for an input sound of 4 seconds. #1 is the acquisition module, #2 and #3 are the features computation and frame-level angle estimation, while #4 refers to the integration module. Time values are in seconds.

the movements of the sound sources within the test environment.

4. Conclusions

We proposed a real-time architecture for sound source localization that combines existing localization techniques with a cheap audio sensor, namely the Microsoft Kinect sensor. The architecture that we presented provides real-time responses. Moreover, it is flexible and modular, which makes it usable with any localization method that requires up to four microphones. We carried out an extensive experimental analysis on four data sets and evaluated different conditions that a localization method has to face with in real scenarios. We realized and made publicly available one of the four data sets used for the experimental analysis.

References

- [1] J. Blauert. *Spatial hearing : the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997.
- [2] M. Crocco, M. Cristani, A. Trucco, and V. Murino. Audio surveillance: A systematic review. *ACM Comput. Surv.*, 48(4):52:1–52:46, Feb. 2016.
- [3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *Intelligent Transportation Systems, IEEE Trans. on*, PP(99):1–10, 2015.
- [4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22–28, 2015.
- [5] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103–138, 1990.
- [6] A. Handzel. Planar spherical diffraction-arrays: Linear sound localization algorithms. In *Sensor Array and Multichannel Processing, 2006. Fourth IEEE Workshop on*, pages 655–658, July 2006.
- [7] A. Handzel and P. Krishnaprasad. Biomimetic sound-source localization. *IEEE Sensors J.*, 2(6):607–616, Dec 2002.
- [8] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda. Sound localization for humanoid robots - building audio-motor maps based on the hrtf. In *IROS, 2006 IEEE/RSJ Intern. Conf. on*, pages 1170–1176, Oct 2006.
- [9] J. Huang, K. Kume, A. Saji, M. Nishihashi, T. Watanabe, and W. L. Martens. Robotic spatial sound localization and its 3-d sound human interface. pages 0191–, Washington, DC, USA, 2002. IEEE Computer Society.
- [10] C. Hummersone, R. Mason, and T. Brookes. Dynamic precedence effect modeling for source separation in reverberant environments. *Audio, Speech, and Language Processing, IEEE Trans. on*, 18(7):1867–1871, Sept 2010.
- [11] M. Jeub, M. Schafer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Digital Signal Processing, 2009 16th Inter. Conf. on*, pages 1–5, July 2009.
- [12] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP J. Adv. Signal Process*, 2009:6:1–6:10, Jan. 2009.
- [13] F. Keyrouz, K. Diepold, and S. Keyrouz. High performance 3d sound localization for surveillance applications. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 563–566, Sept 2007.
- [14] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Trans. on*, 24(4):320–327, Aug 1976.
- [15] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):1–13, Jan 2011.
- [16] M. Murase, S. Yamamoto, J. marc Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. G. Okuno. Multiple moving speaker tracking by microphone array on mobile robot. 1:143–145, 2005.
- [17] K. Nakadai, H. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *IROS, 2002. IEEE/RSJ Intern. Conf. on*, volume 2, pages 1320–1325 vol.2, 2002.
- [18] K. Nakadai, H. G. Okuno, H. Kitano, H. G. Okuno, and H. Kitano. Real-time sound source localization and separation for robot audition. In *in Proceedings IEEE ICSLIP, 2002*, pages 193–196, 2002.
- [19] H. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano. Human-robot interaction through real-time auditory and visual multiple-talker tracking. In *IROS, 2001. Proceedings. 2001 IEEE/RSJ Intern. Conf. on*, volume 3, pages 1402–1409 vol.3, 2001.
- [20] S. T. Shivappa, B. D. Rao, and M. M. Trivedi. Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation. *IEEE J. Sel. Topics Signal Process.*, 4(5):882–894, Oct 2010.
- [21] I. Toshima, S. Aoki, and T. Hirahara. An acoustical telepresence robot: Telehead ii. In *IROS, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ Intern. Conf. on*, volume 3, pages 2105–2110 vol.3, Sept 2004.
- [22] J. Valin, F. Michaud, J. Rouat, and D. Letourneau. Robust sound source localization using a microphone array on a

mobile robot. In *IROS, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ Intern. Conf. on*, volume 2, pages 1228–1233 vol.2, Oct 2003.

- [23] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner. A probabilistic model for binaural sound localization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, 36(5):982–994, Oct 2006.