Robotic Orientation towards Speaker for Human-Robot Interaction

Caleb Rascón, Héctor Avilés, and Luis A. Pineda*

Instituto de Investigaciones en Matemáticas Aplicacas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)
{caleb,haviles}@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

Abstract. The orientation of conversational robots to face their interlocutors is essential for natural and efficient Human-Robot Interaction (HRI). In this paper, progress towards this objective is presented: a service robot able to detect the direction of a user, and orient itself towards him/her, in a complex auditive environment, using only voice and a 3-microphone system. This functionality is integrated within Spoken HRI using dialogue models and a cognitive architecture. The paper further discusses applications where robotic orientation benefits HRI, such as a tour-guide robot capable to guide a poster session and a "Marco Polo" game where a robot aims to follow a user purely by sound.

Keywords: Cognitive architecture, Human-robot interaction, Direction-of-arrival estimation, Robotic orientation.

1 Introduction

The orientation of a service robot is essential for a natural Human-Robot Interaction (HRI). From the user experience point of view, the human will feel the interaction as more 'natural' if the robot is facing him/her during the conversation. However, as such ability requires the direction of the user to be known, it can benefit other parts of HRI.

For instance, once the direction of the user is known, voice recognition can be improved using directional noise cancellation [7]. In addition, it is well known that face detection and recognition provide rich information relevant to HRI: the identity of the user, the direction the user is looking at, his mood, etc. [4,18] However, such analysis is carried out by visual means, and the user cannot always be expected to be in the line of sight. Knowing the direction of the user by sound alone, and facing the user accordingly, tackles this issue straight on. Moreover, it is imperative to know the direction of the user in a situation such as the one proposed in the test *Follow me*, taking place in the popular *Robocup@Home* competition[1], where the robot is required to track and follow the user.

* The authors would like to acknowledge the support from the grants CONACYT 81965 and PAPIIT-UNAM IN-104408 and IN-115710.

[1] www.robocup2010.org

A. Kuri-Morales and G.R. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 10–19, 2010.

Knowing the direction of the user in regard to the robot is also a good first step towards localizing him/her in the environment. In a 3-dimensional polar coordinate system, the horizontal angle (i.e. the direction of the user) is one of three values that define a location (the other two being: vertical angle and distance from origin). And the location of the user is, in turn, another important variable in HRI. During a human-robot conversation, the phrase "robot, come here" may be emitted by the human. In this situation, even with the phrase recognized correctly, the robot may know that it needs to move, but, because the term 'here' lacks context, it will not know **where** to move.

Unfortunately, locating the user in the environment is a difficult task to carry out by sound analysis alone. It is of interest having several types of modalities (vision, sound, laser, etc.) working in conjunction for that objective [14,19,13], where the direction of the user plays an important role [12].

In addition, many types of noise may hinder the estimation of the sound source Direction-of-Arrival (DOA), such as reverberation [2]. A sophisticated recording system may be able to overcome them, such as the one proposed in [17] that used a 24-microphone 1-D array for precision. However, a high amount of microphones may be impractical to carry by many of the currently-in-use service robots [20,11], such as our in-house robot, Golem, herein described.

Golem is a service robot built with a primary focus on HRI. It is integrated by a cognitive architecture focused on HRI, explained in detail in [14], which can take advantage of different types of information interpreted from the world, including the direction of the user. Because Golem is a conversational robot, it is of interest to implement a robotic-orientation module that is triggered by sound. This implies that the module needs to be sufficiently light on the hardware side for the robot to carry, but robust enough in the software side to handle different types of noise and disturbances. Moreover, for tracking purposes, the module should be able to estimate the direction of the user in a -179°– 180° range, and fast enough to do so with small utterances from the user.

The paper is organized as follows: Section 2 further discusses how a direction-aware service robot may benefit HRI, as well as give a brief review of current algorithms that aim to estimate the direction of a speaker; Section 3 describes the hardware setup and the proposed algorithm; the different trials on an actual service robot and their results are given in Section 4; and conclusions are discussed in Section 5.

2 Background

2.1 On Benefits to Human-Robot Interaction

A cognitive architecture for a service robot, designed to focus primarily on HRI, is presented in [14], and is implemented in our service robot Golem. This cognitive architecture has a top level, called the 'Representation & Inference' phase, where the set of expected multimodal situations are defined and ordered in what is called a Dialogue Model (DM), which guides the HRI process and assumes an abstract-but-meaningful interpretation of the world. This interpretation is

achieved by first obtaining an uninterpreted image [14] (i.e. the Recognition phase), and then providing a meaning to that representation (i.e. the Interpretation phase), guided by the contents of the episodic memory of the architecture and the current expectations of the DM. With this architecture, complex objectives can be fulfilled, such as a tour-guide robot capable to guide a poster session [1] and a spanish-spoken "guess the card" game [8].

An example of a DM is shown in Figure 1, which describes the popular Marco Polo game[2]. The HRI process flows through the DM, with each node describing a situation. Each node is joined to other nodes by arrows that are tagged with an "Expectation:Action" label. If the Expectation with which an arrow is tagged is met, the corresponding Action is carried out and the system moves to the node the arrow is pointing at.

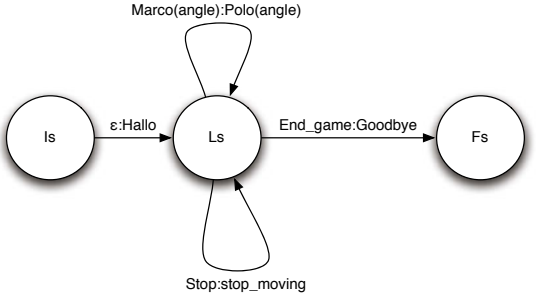


Fig. 1. Dialogue Model for the Marco Polo game

An Expectation is met when the abstract interpretation of the world matches the expected occurrence in the world, such as the user greeting the robot at the start of a tour or the robot bumping into something when moving. If an expectation is met accordingly, the DM furthers along the HRI process; if something happens in the world, but no expectations were in place for it in the current situation in the DM, the system executes a recovery dialogue model to continue the HRI.

In the example of Figure 1, the expectation *Marco(angle)* is met when the user specifically says the word "Marco", detected to be coming from the direction *angle*, and it triggers the action *Polo(angle)*, indicating to the robot to stop, say the word "Polo", turn to the direction *angle*, and continue moving. The expectation *Stop* is met when the user expresses the wish for the robot to stop, and it triggers the action *stop_moving*, indicating to the robot to stop. Finally, the expectation *End_game* is met when the user expresses the wish for the game to end, and it triggers the action *Goodbye*, indicating to the robot to stop and express a farewell.

[2] A game between two parties, where one is blind and tries to follow the other party by sound alone. The followed party must yell "Marco" to signify where he/she is, and the following blind party must respond "Polo" and move to that location.

The module proposed in this paper will reside inside the Recognition phase, where the sound of the outside world will be tagged with the characteristic of "direction". This characteristic can be used by the Interpretation phase to enrich the meaning of the information received by other modalities, such as in the example presented in [1], where the tour-guide robot could approach and greet the user at the start of the tour, triggered by sound alone (instead of manual means); the robot could also face the user when being talked to during the tour, enhancing the 'naturality' of the conversation. The "direction" characteristic may also be used directly by the DM, as in the case of the Marco Polo example, where such information is essential for carrying out HRI. To do any of these, however, a sound source direction estimator needs to be implemented first.

2.2 On Source Direction Estimation

Estimating a Sound Source Direction of Arrival (DOA) is a well written-about topic in Signal Processing. It has been proven useful in applications ranging from fault monitoring in aircrafts [17], to intricate robotic pets [6], to close-to-life insect emulation [5]. In addition, the principles employed in DOA estimation have been applied in the design of hearing aids [7].

Having two audio sensors (i.e. microphones), the Inter-aural Time Difference (ITD) is the delay of a sound from one microphone to the other. It is one of the main sources of information for DOA estimation, as it provides a clear relation to the direction of the source (it was applied in [10] with limited results). The Inter-aural Intensity Difference (IID) is the difference in magnitude between both microphones and can also be used for DOA estimation, although a training stage is usually necessary for it to be useful, as it was observed in [11].

In [2], the concept of Inter-aural Coherence (IC) is introduced, which is the highest correlation measure of the cross-correlation vector between the two signals. If a high IC is present, the signals are deemed coherent and, thus, an analysis using ITD and/or IID can proceed. This methodology was implemented in [6], and it was observed that it didn't improve DOA estimation when dealing with complex signals (e.g. more than one source, reverberation present, etc.).

The Multiple Signal Classification algorithm (MUSIC) [16] is able to detect the Direction of Arrival (DOA) of as many sources as one less the available microphones (e.g. 1 source for 2 microphones). It does this by projecting the received signals in a DOA subspace, based on their eigenvectors, similar to Principal Component Analysis. It was applied in [9] with good results, although it has been observed that its performance decreases considerably in the presence of reverberation [21] (pp. 169).

Classic beamforming can be applied for precise DOA estimation [17], but requires a large quantity of microphones which is impractical for smaller robots.

The DOA of a source has been able to be estimated using one microphone by implementing an 'artificial ear' [15], but the sound was required to be known *a priori* and any modification to the ear (even its location in relation to the microphone) required re-training.

More specifically in the area for service robots and HRI, DOA estimation can be considered to be in its initial stages. In [20], the sources are separated from each other, in order to enhance speech recognition, and as a preamble for DOA estimation. However, it required an array of 8 microphones positioned in a cube-like manner to work, which may be impractical for service robots.

Other attempts, such as [12], have approached the DOA estimation problem using a two-microphone array. The reasoning behind using only two microphones ranges from that of practicality (it is lightweight) to that of biological similarity [21], where the robot is meant to be the most human-like possible [3]. However, doing so comes with three main problems.

The feature most used for DOA estimation when using a 2-mic array is the ITD, as there is a direct relation between the two, described in equation (1).

$$\theta = arcsin\left(\frac{ITD \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{1}$$

where θ is the DOA angle; ITD is the Inter-aural Time Difference in number of samples; V_{sound} is the speed of sound ($\sim$ 343 m/s); F_{sample} is the sampling frequency (usually 44.1 kHz); and d is the distance between microphones.

In Figure 2, the DOA is plotted against the ITD, and it can be seen that in the -50°–50° range, the relation between both seem almost linear. However, in the outer ranges, the relation becomes exponential. This causes major errors when estimating angles that are located in the sides of the robot [12].

As it can also be seen in Figure 2, a 2-mic array only estimates DOAs in the -90°– 90° range. This can be surmounted by implementing 'artificial ears' that can detect if the sound source is coming from the front or back of the robot, but it has been proven impractical [15]. This can also be tackled by a two-phase strategy: a first pair of signals are used to estimate an initial DOA, the robot then rotates briefly, and then another pair of signals are acquired to estimate a second DOA. Comparing both DOAs provides an angle in the -179°– 180° range,

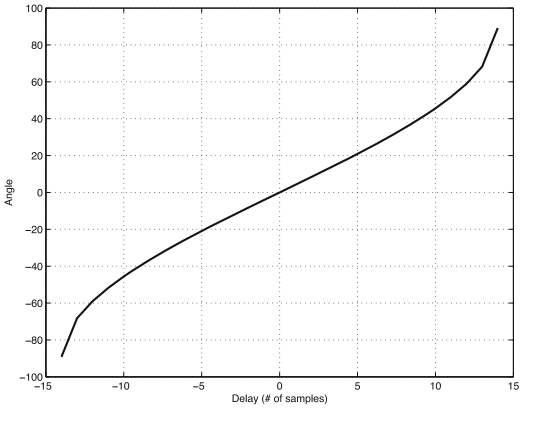


Fig. 2. DOA (or *Angle*) versus ITD (or *Delay*) in number of samples

but has its own set of issues: it requires considerably more time than when using one DOA estimate, the required rotation may hinder navigational requirements, and the user may be moving as well, rendering the DOA comparison mute.

Finally, the estimation of the ITD is usually based on the calculation of the cross-correlation vector between the two signals. This process can be very sensitive to reverberations and other noise sources when using only 2 microphones [21] (pp. 213-215), which may result in significant errors in the DOA estimation.

3 A 3-Microphone System

To avoid the problems described in the last section that arise when estimating a DOA using only two microphones, and to maintain a relatively light hardware setup, a three-microphone system, as it is shown in Figure 3, is proposed.

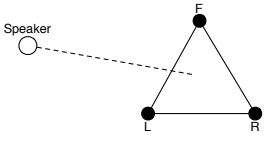


Fig. 3. The proposed 3-microphone system

The distances RL, LF, and FR are equal, thus a preliminary angle can be calculated using equation (2).

$$\theta_m = arcsin\left(\frac{ITD_{min} \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{2}$$

where ITD_{min} is the ITD with the lowest absolute value of the three ITDs. Equation (3) is then used to calculate the final DOA value (θ).

$$\theta = f(\theta_m, ITD_{RL}, ITD_{LF}, ITD_{FR}) \tag{3}$$

where f is a function that first checks if the three ITDs (ITD_{RL}, ITD_{LF}, and ITD_{FR}) point to the same angle sector. It is a type of redundancy check against situations with reverberation and/or with more than one sound source. If the ITDS are redundant, f shifts θ_m to the pointed angle sector to obtain θ.

With this setup, θ is always estimated using a θ_m inside the -30°– 30° range (well within the close-to-linear -50°– 50° range), resulting in a close-to-linear ITD-DOA relation all through the -179°– 180° range.

In addition, because of the redundancy check and the close-to-linear relation, the maximum error of this system can be known beforehand using equation (4).

$$|error^{\circ}_{max}| = arcsin\left(\frac{ITD_{>30^\circ} \cdot V_{sound}}{F_{sample} \cdot d}\right) - arcsin\left(\frac{ITD_{<30^\circ} \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{4}$$

where $ITD_{>30^\circ}$ and $ITD_{<30^\circ}$ are the ITDs (measured in number of samples) that provide the closest ceil and floor measurements, respectively, to 30°. In the case of Golem, sampling at 44.1 kHz and with the microphones spaced at 18 cm, a maximum error of ±2.8747° can be expected. In those same circumstances, when using a 2-Mic array, the maximum expected error is ±15.0548°, which occurs when the sound source is close to either side of the robot.

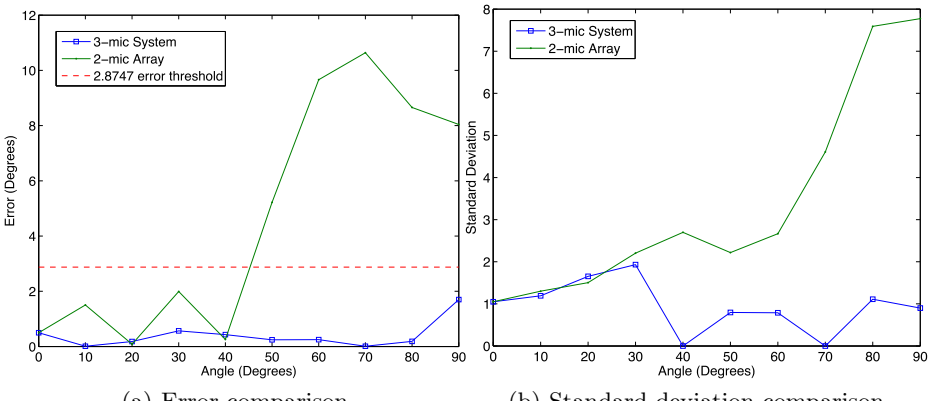
4 Trials and Results

Two trials were implemented: one offline, where the angle is estimated in a relatively quiet environment, and another online, where the angle was used as part of a simple navigation system implemented in Golem.

The offline trial was carried out to compare the proposed 3-mic system and a 2-mic array in a situation where both systems can potentially provide good results. Because a 2-mic array is not able to detect angles outside the -90°– 90° range, 10 equally spaced angle values within the 0°– 90° range were used as a testbed for both systems (the -90°– 90° range is symmetrical).

At each angle value, a sound source was located at 20 cm from the system, and 10 sets of 4410 samples were taken. With each sample set, an angle was estimated; the average of the 10 estimated angles was considered as the final estimated angle. The final estimation was then subtracted from the known angle to calculate the error. The standard deviation of the 10 angles was also calculated. The results of the trial are shown in Figure 4.

As it can be seen in Figure 4a, the 3-mic system is able to provide good angle estimates throughout the angle range, while the 2-mic array error increases substantially with angles greater than 40°. In addition, it can be seen in Figure 4b that the 2-mic array, when estimating large angles, estimations deviate considerably more from one moment to another than the 3-mic system. This results show that the 3-mic system not only provides better angle estimations than the



(a) Error comparison (b) Standard deviation comparison

Fig. 4. Results of offline trial

two 2-mic array, but also that the estimations are more consistent throughout the angle range. Adding on the fact that this consistency is provided in all of the -179°– 180° range, this results show that 3-mic system is superior in all respects to the 2-mic array, while being comparably lightweight.

Because the 3-mic system overly outperformed the 2-mic array, it was meant appropriate to install over Golem, which triggered the online trial. The testing area was a large room, with higher reverberation than that of the offline trial, and in presence of higher ambient noise (fans, keyboard typing, low-to-medium chatter between desks). A simple navigation system was implemented, such that when an angle is provided, the system would turn towards that angle. The sound source was a user who was instructed to utter the phrase "Golem" from 9 different angular positions 2 meters away from the robot. All estimations were carried out starting from the 0° position and only the 3-mic system was tested[3]. The results are shown in Figure 5.

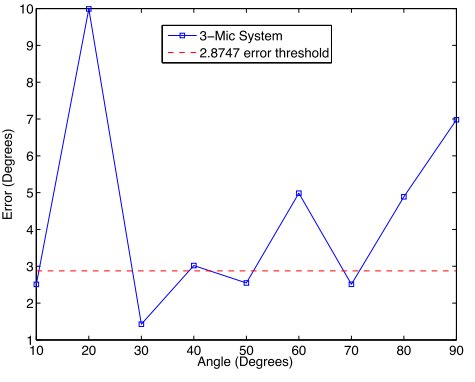


Fig. 5. Results of online trial

As it can be seen, the estimations are considerably worsened in the online trial. However, in all cases, except for the 20° and 90° position, the estimation is no more than 3° above the offline error threshold, which for real-life use is more than adequate. In fact, the user that was testing the robot regarded its orientation as adequate in all the angular positions, except at 20°. The error spikes in the 20° and 90° positions are currently being investigated.

It is important to note that this was a real-life test. The sonic complexity of the room was high, the user placement can be expected to be inconsistent, the navigation system could have produced some positional errors, and no reverb filtering was carried out. When considering all of this, the 3-mic system has shown it is an adequate solution to the user DOA estimation problem.

[3] The 2-mic array was deemed inappropriate for this scenario. It consistently misidentified the sound source, making the robot behave erratically.

5 Conclusion and Future Work

Human-Robot Interaction benefits from a rich perception of the world. Having the robot orient itself towards the user during a conversation enhances HRI from the point of view of both the user and the robot: the 'naturality' of the conversation is improved, and the acquisition of more information from the user (face recognition, voice context, etc.) is simplified. To do this, however, the direction of the user is required. Because a conversation is carried out via voice, it is appropriate that the direction of the user be estimated by sound analysis.

It was shown that a 3-microphone system provides a more reliable Direction-of-Arrival estimation than the more-widely-used 2-mic array, while being light enough to be carried by a service robot. It also provided a robust estimation in the presence of reverberation and other types of noise. However, the 3-mic system could only go so far, as the implemented navigation system in Golem (our service robot) contributed some orientation errors that need to be amended.

A good first step was presented towards the integration of a robotic-orientation module in various HRI applications, such as a tour-guide robot and a "Marco Polo" game. This integration is considered as future work in this project.

It is important to note that it is intended for the 3-mic system to be the medium with which the Automatic Speech Recognizer (ASR) gets audio data. Currently, when using the onboard 3-mic system, the performance of the ASR decreases significantly, compared to when using a headset. The reason for this is that, even though the direction estimator is now more robust against reverberation, the ASR is not. To this effect, the next phase of this work is to tackle the reverb problem, aided by the 3-mic system, to enhance voice recognition.

References

1. Avilés, H., Alvarado-González, A.M., Venegas, E., Rascón, C., Meza, I., Pineda, L.A.: Development of a tour–guide robot using dialog models and a cognitive architecture. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 512–521. Springer, Heidelberg (2010)
2. Faller, C., Merimaa, J.: Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. The Journal of the Acoustical Society of America 116(5), 3075–3089 (2004)
3. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42(3-4), 143–166 (2003)
4. Hjelmås, E., Low, B.K.: Face detection: A survey. Computer Vision and Image Understanding 83(3), 236–274 (2001)
5. Horchler, A.D., Reeve, R.E., Webb, B., Quinn, R.D.: Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization. In: Sound Localization, 11th International Conference on Advanced Robotics (ICAR 2003), pp. 1749–1756 (2003)
6. Liu, R., Wang, Y.: Azimuthal source localization using interaural coherence in a robotic dog: modeling and application. Robotica First View, 1–8 (2010)

7. Lockwood, M.E., Jones, D.L., Bilger, R.C., Lansing, C.R., O'Brien Jr., W.D., Wheeler, B.C., Feng, A.S.: Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. Journal of the Acoustical Society of America 115(1), 379–391 (2004)
8. Meza, I., Salinas, L., Venegas, E., Castellanos, H., Chavarria, A., Pineda, L.A.: Specification and evaluation of a spanish conversational system using dialogue models. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 346–355. Springer, Heidelberg (2010)
9. Mohan, S., Lockwood, M.E., Kramer, M.L., Jones, D.L.: Localization of multiple acoustic sources with small arrays using a coherence test. Journal of the Acoustical Society of America 123(4), 2136–2147 (2008)
10. Murray, J.C., Erwin, H., Wermter, S.: Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In: AI Workshop on NeuroBotics (2004)
11. Murray, J.C., Erwin, H.R., Wermter, S.: Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. Neural Networks 22(2), 173–189 (2009)
12. Nakadai, K., Okuno, H.G., Kitano, H.: Real-time sound source localization and separation for robot audition. In: Proceedings IEEE International Conference on Spoken Language Processing, pp. 193–196 (2002)
13. Pineau, J., Montemerlo, M., Pollack, M., Roy, N., Thrun, S.: Towards robotic assistants in nursing homes: Challenges and results. Robotics and Autonomous Systems 42(3-4), 271–281 (2003)
14. Pineda, L.A., Meza, I., Salinas, L.: Dialogue model specification and interpretation for intelligent multimodal HCI. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 20–29. Springer, Heidelberg (2010)
15. Saxena, A., Ng, A.Y.: Learning sound location from a single microphone. In: ICRA 2009: Proceedings of the 2009 IEEE International Conference on Robotics and Automation, pp. 4310–4315. IEEE Press, Piscataway (2009)
16. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation 34(3), 276–280 (1986)
17. Smith, M., Kim, K., Thompson, D.: Noise source identification using microphone arrays. Proceedings of the Institute of Acoustics 29(5) (January 2007)
18. Stiefelhagen, R., Ekenel, H.K., Fugen, C., Gieselmann, P., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., Waibel, A.: Enabling multimodal human-robot interaction for the karlsruhe humanoid robot. IEEE Transactions on Robotics 23(5), 840–851 (2007)
19. Tanawongsuwan, R., Stoytchev, A., Stoytchev, E., Essa, I.: Robust tracking of people by a mobile robotic agent. Tech. rep. (1999)
20. Valin, J., Rouat, J., Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, pp. 2123–2128 (2004)
21. Wang, D., Brown, G.J.(eds.): Computational auditory scene analysis: Principles, Algorithms, and Applications. IEEE Press/Wiley-Interscience (2006), `http://www.casabook.org/`