MUSIC空間スペクトログラムを用いた複数音源の発話区間検出の検討

Investigation of utterance detection of multiple sound sources based on the MUSIC spatial spectrogram

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
梁棟（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
石黒浩（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Liang DONG, Hiroshi ISHIGURO, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, liang@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - With the goal of improving human-robot speech communication, the localization of multiple sound sources in the 3D-space based on the MUSIC algorithm was implemented and evaluated in a humanoid robot embedded in real noisy environments. A method for tracking sound intervals of multiple sound sources was then proposed based on the sound directivity inferred from the MUSIC spectrogram. The proposed method achieved good sound interval detection accuracies and low insertion rates compared with previous sound localization results.

1 Introduction

In human-robot speech communication, the microphones on the robot are usually far (more than 1 m) from the human users, so that the signal-to-noise ratio becomes lower than for example in telephone speech, where the microphone is centimeters from the user's mouth. Due to this fact, interference signals, such as voices of other subjects close to the robot, and the background environment noise, would degrade the performance of the robot's speech recognition. Therefore, sound source localization and posterior separation become particularly important in robotics applications.

One of the difficulties that degrade speech recognition performances in the robot's real environment is the lack of accuracy in utterance detection. In the present work we propose the use of sound localization (or more specifically, sound directivity) for improving utterance detection.

There are many works about sound source localization [1]-[9]. The sound localization method adopted in the present work is the MUSIC (MUltiple SIgnal Classification) algorithm, which is a well-known high-resolution method for source localization [1]-[3]. However, there are two issues regarding the MUSIC algorithm, which constrain its application for sound localization in practice. One is the heavy computational cost, while the other is the need of previous knowledge about the actual number of sources present in the input signal.

Regarding evaluation, although there are many works related to sound localization, most of them only evaluate simulation data or laboratory data in very controlled conditions. Also, only a few works evaluate sound localization in the 3D space, i.e., considering both azimuth and elevation directions [8]-[9]. Looking at the user's face while the subject is speaking is also an important behavior for improving human-robot dialogue interaction, and for that, a sound localization in 3D space becomes useful.

Taking the facts stated above into account, in our previous work [10], we constructed a MUSIC-based 3D-space sound localization (i.e., estimation of both azimuth and elevation directions) in the communication robot of our laboratory, "Robovie", and evaluated it in real noisy environments. However, only the raw data (without sound interval segmentation) was evaluated. Also, there still are considerable insertion error rates, for getting high detection rates.

In the present work, we propose and evaluate a sound interval detection method, by tracking sound sources directly on the MUSIC spectrogram and delta-spectrogram.

This paper is organized as follows. In Section 2, descriptions about the hardware and data collection are given. In Section 3, the proposed method is explained, and in Section 4, analyses and evaluation results are presented. Section 5 concludes the paper.

2 Hardware and data description

2.1 The microphone array

A 14-element microphone array was constructed in order to fit the chest geometry of Robovie, as shown in Fig. 1.

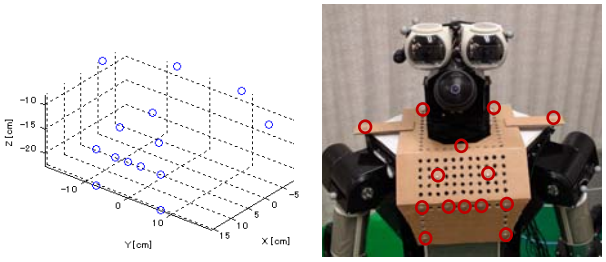

Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

The chest was chosen, instead of the head, due to geometric limitations of Robovie's head. Several 3D array architectures were tested using simulations of the MUSIC algorithm. The array geometries were designed in such a way to cover all three-dimensional coordinate axes, giving emphasis to resolution in azimuth direction, and sounds coming from the front. The array configuration shown in Fig. 1 was chosen since it produced fewer side-lobes and had a fairly good response over different frequency bins.

A 16-channel A/D converter TD-BD-16ADUSB from Tokyo Electron Device Limited was used to capture the signals from the array microphones. Sony ECM-C10 omni-directional electret condenser microphones were used as sensors. Audio signals were captured at 16 kHz and 16 bits.

2.2 Recording setup

The microphone array was set on the robot's chest structure, as shown in Fig. 1. The robot was turned on to account for the noise produced by its internal hardware. The sources (subjects) were positioned around the robot in different configurations and were instructed to speak to the robot in a natural way. Each subject had an additional microphone to capture their utterance. The signals from these additional microphones, which we will call "**source signals**" throughout the paper, will be used only for analysis and evaluation. Nonetheless, the source signals are not required by the proposed method in its final implementation.

2.3 Data collection and environmental conditions

Recoding data using the microphone array was collected in two different environments. One is an office environment (OFC), where the main noise sources are the room's air conditioner and the robot's internal hardware noises. The second environment is a hallway of an outdoor shopping mall (called Universal City Walk Osaka – UCW), where a field trial experiment has been executed [11]. The main noise source in UCW was a loud pop/rock background music coming from the loudspeakers on the hallway ceiling. The ceiling height is about 3.5 meters. Recordings were done with the robot faced to different directions, in several places.

In OFC, four sources (male subjects) are present. At first, each source speaks to the robot for about 10 seconds, as the others remain silent. In the last 15 seconds of the recording, all four sources speak at the same time. For this recording, two of the subjects wore microphones connected to the two remaining channels of the 16-channel A/D device, while the other two subjects wore microphones connected to a different audio capture device (M-audio USB audio). A clap at the beginning of the recording was used to manually synchronize the signals of these two speakers to the array signals. It is worth to mention that a strict synchronization between the source signals was not necessary, because only power information of the source signals will be used, as will be explained in Section 2.4.

In UCW, there are two speech sources (male subjects) present in all recordings. In most of the trials, the sources take turns to speak for about 10 seconds each and then proceed to talk at the same time. In two of the trials (UCW7 and UCW8), one source is moving and the other is static, both speaking at the same time most of time. In five trials (UCW1-4, UCW9), the robot is far from the ceiling loudspeakers, while in four trials, the robot is close (a few meters) to a loudspeaker (UCW5-8), and in another four trials, the robot is right under a loudspeaker (UCW10-13). All trials have different configurations for the robot facing direction and/or source locations.

2.4 Computation of the reference number of sources from the power of the source signals (PNOS)

The number of sources (**NOS**) is an important parameter required by the MUSIC algorithm, which influences on the performance of DOA estimation. For analysis and evaluation of the NOS in the DOA estimation performance, reference NOS were computed from the power of the source signals. These power-based NOS values will be referred as **PNOS**.

Prior to compute the power of each source, a cross-channel spectral binary masking was conducted among the source signals in order to reduce the inter-channel leakage interferences, and get more reliable reference signals. In addition, the signal of the microphone in the center position in the array was used to remove the ambient music noise from all the source signals. Finally, the signal was also manually attenuated in the intervals where interference leakage persisted after the above processing. This resulted in much clearer source signals.

The average power of the signal was computed for each 100 ms, which corresponds to the block interval used in the MUSIC algorithm. A threshold was manually adjusted to discriminate the blocks with sound activity for each source signal. For each block, PNOS is then given by the summation of the source signals with activity.

In the UCW recordings, an additional source (due to the background music) was added to PNOS.

3 Proposed method

3.1 The broadband MUSIC spectrum

Fig. 2 shows the block diagram of the algorithm for computing the broadband MUSIC spectrum. The algorithm structure is similar to a classical approach of the MUSIC algorithm: getting the Fourier transform (FFT) for computation of the multi-channel spectrum, computing the cross-spectrum correlation matrix, making the eigenvalue decomposition of the averaged correlation matrix over a time block, computing the (narrowband) MUSIC responses for each frequency bin using the eigenvectors corresponding to the noise subspace and the steering/position vectors prepared beforehand for the desired search space, and finally computing the broadband MUSIC response by averaging the narrowband responses over a frequency range.

The broadband MUSIC responses are referred as MUSIC spectrum, while the sequence of the MUSIC spectrums along the time is referred as MUSIC spectrogram.

In our previous work, some of the parameters related to the MUSIC response computation were analyzed in order to obtain real-time processing, while keeping the DOA (direction of arrival) estimation performance. These parameters related to the MUSIC computation are described in detail in the following sub-sections 3.1.1 to 3.1.4.

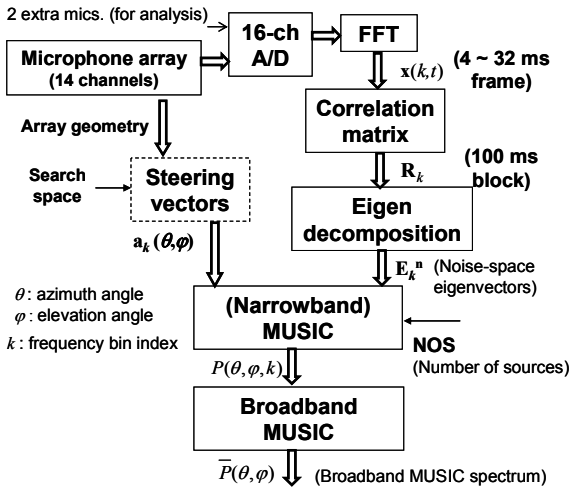

Fig. 2. The MUSIC-based sound localization algorithm, and related parameters.

3.1.1 Search space for DOA (directions of arrival)

The MUSIC algorithm was implemented to obtain not only the azimuth but also the elevation angle of the direction of arrival (DOA) of each source signal. Since the goal of this development is to enhance the human/robot interaction, we considered that it was not necessary to estimate the distance between the robot and the source(s) and that the DOA was the important piece of information. Nonetheless, the MUSIC algorithm can easily be extended to estimate also the distance between the array and the source, by adding the corresponding steering/position vectors. However, this would considerably increase the processing time.

A spherical mesh with a step of 5 degrees was constructed for defining the directions to be searched by the MUSIC algorithm. The mesh was constructed by setting elevations in intervals of 5 degrees, and setting different number of azimuth points for each elevation. The number of azimuths is maximum for 0 degrees elevation (having 5 degrees azimuth intervals), and gradually reduces for higher elevations, in such a way that the arc between two points is kept as close as possible to the arc corresponding to 5 degrees azimuth in 0 degrees elevation. This reduces the number of directions to be scanned by the MUSIC algorithm, reducing computation time. The directions with elevation angles lower than -30 degrees were also removed to speed up the computation, resulting in a total of 1216 directions.

The origin of the coordinate frame is set to the intersection point of the rotational axis of the degrees of freedom of the Robovie's head. This way, the output from the DOA estimation algorithm can be directly used to servo the head.

3.1.2 Frame length and block length

The frame length, which is related to the number of FFT points to be computed in the first stage, is an important parameter that can drastically reduce the computational costs of the MUSIC algorithm. Although FFT of 512 ~ 1024 points is commonly used (corresponding to 32 ~ 64 ms frame length at 16 kHz), we have proposed the use of smaller FFT sizes (64 ~ 128). This allows reducing the computation not only of the FFT stage, but also of the subsequent correlation matrix, eigenvalue decomposition, and MUSIC response computations. We have found that reducing the frame size to 4 ms (or equivalently reducing the FFT size to 64) was effective to allow real-time processing without a big degradation in the estimation of the directions of arrival (DOA) of sound sources.

In the next step of the MUSIC algorithm, a correlation matrix is averaged for the frames within a time block. The block length has to be long enough for getting good estimation of the averaged correlation matrix. On the other hand, it also should be short enough for getting good temporal resolution (considering that a sound source can move) and low latency. In the present work, we decided to use a time block length of 100 ms.

3.1.3 Frequency range of operation

Although speech contains information over a broad frequency band (vowels in 100 – 4000 Hz and fricative consonants in frequencies above 4000 Hz), the frequency range of operation for DOA estimation has to be limited, given the geometric limitations of the array (shown in Fig. 1).

The smallest distance between a pair of microphones is 3 cm, so that on theory the highest frequency of

operation to avoid spatial aliasing would be about 5.6 kHz (according to Rayleigh's Law).

Regarding the lowest frequency boundary, although speech contain important information in frequency bands lower than 1 kHz, the array geometry limitations do not allow good spatial resolution in these low frequency bands. In the present work we use the frequency range of 1 – 6 kHz, for avoiding the issues above.

3.1.4 Number of sources (NOS)

The number of sources is an important parameter necessary for getting a good estimate of the MUSIC spectrum. In theory, there is some relationship between the number of sources and the shapes of the eigenvalue profiles. However, a threshold between strong and weak eigenvalues is difficult to be determined. The environment noise has also a strong impact on the shapes of the eigenvalues, so that both magnitude and slope of the profiles are affected.

Considering the difficulties in estimating NOS from the eigenvalues of the spatial correlation matrix, we have proposed the use of a fixed number of sources for the (narrowband) MUSIC response computation ("fixed NOS"), and establishing a maximum number of sources detectable from the broadband MUSIC response ("max NOS"). Here, we allow the maximum number of sources detectable being larger than the fixed number of sources for the MUSIC response computation. This idea is based on the assumptions that at an instant time, the predominance of different broadband sound sources varies depending on the frequency bins. Therefore, even if the NOS used to the narrowband MUSIC computation is limited to a fixed small number, the combination of frequency bins to compute the broadband MUSIC spectrum may produce more peaks than the fixed number.

Fig. 3 shows examples of MUSIC spectrograms for OFC1, where 4 sources are speaking in front of the robot, UCW8, where one of the sources is speaking in front of the robot, while the other speaker is moving in front of the robot while speaking, and a directional music source is present in the first half of the trial, and UCW13, where the robot is right under a loudspeaker. A suitable plotting of the MUSIC spectrogram is difficult because there are several dimensions: azimuth angle, elevation angle, MUSIC power and time. In the MUSIC spectrogram of Fig. 3, we show the azimuth angle in the y-axis, time in x-axis, different colors for different elevations, and different tonalities according to the MUSIC power. (The colors can be viewed in the electronic version.) Note that in the middle panel of Fig. 3, the music source in the first half appears in green, because the loudspeakers are in higher elevations compared to humans, while there are strong lines in pink/red in the bottom panel of Fig. 3, because there is a loudspeaker over the robot.

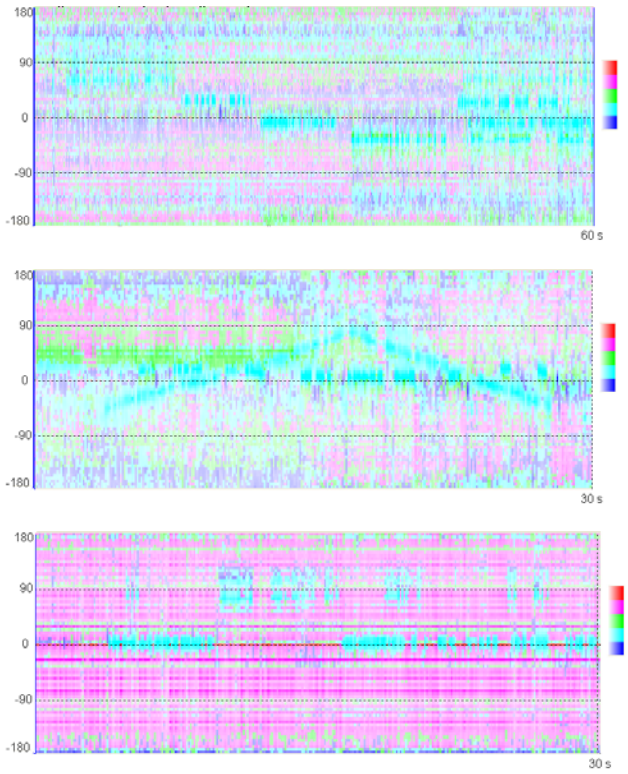

Fig. 3. Examples of MUSIC spectrograms for OFC1 (where 4 sources are present in front of the robot), for UCW8 (where one of the sources is moving in front of the robot), and for UCW13 (where the robot is right under a loudspeaker). The elevation angles are displayed by different colors. (Please refer to the electronic document to see the colors.)

3.2 Sound interval detection from the MUSIC spectrogram

In a classical approach for determining the direction of arrival (DOA) of the sound sources, peak picking is realized on the MUSIC spatial spectrum. In the present work, we proposed a method for detecting sound source intervals, based on a MUSIC spectrogram and delta-spectrogram.

Fig. 4 shows a block diagram of the proposed method for sound interval detection.

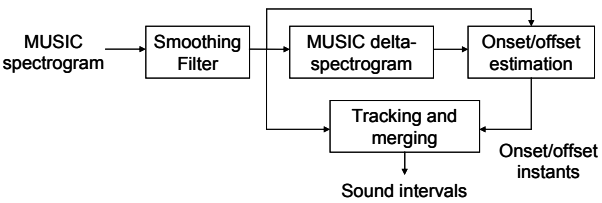

Fig. 4. The proposed sound interval detection based on MUSIC spectrogram and delta-spectrogram.

First, a moving-average smoothing filter is passed through each direction (azimuth vs. elevation) of the raw MUSIC spectrogram, by a Hamming window of 5 taps, for reducing temporal distortions.

Then, a delta-spectrogram is computed by taking, for each direction (azimuth vs. elevation), the minimum difference between the MUSIC power values of the

current block and the neighbor directions in the previous block.

Thresholds are set for estimation of onset and offset instants. First, thresholds are imposed to the positive and negative peaks of the MUSIC delta-spectrogram, to get raw estimates of onset and offset instants. To avoid over-estimation of the number of sources, we set a threshold for the magnitude of the MUSIC power, as proposed in our previous work. Further, the maximum number of onsets per block is constrained, assuming that the probability of multiple sources starting at the same instant is low. This last constraint is in particular important to reduce insertion errors due to sidelobe effects.

Also, to avoid misdetection of the offset instant, we forced offset if the MUSIC power becomes lower than the MUSIC power of the block previous to the onset instant (onset MUSIC power) plus a bias factor alpha. The following summarizes the thresholds involved in the onset/offset detection.

- Onset: [MUSIC delta power > 1.0 dB] and [MUSIC power > 1.8 dB] and [Max. number of onsets per block <= 2]
- Offset: [MUSIC delta power < -1.2 dB] or [MUSIC power < onset MUSIC power + alpha]

Finally, the path with maximum MUSIC power is tracked in the MUSIC spectrogram from the onset to the offset instant. Segments separated by short pauses smaller than 4 blocks (400 ms), and with continuity in the direction are merged.

4 Analysis and experimental results

4.1 The evaluation setup

To measure the performance of the DOA estimation, we used three scalar values. The first represents the percentage of ideal DOA that were detected successfully by the algorithm. We will call this quantity "**DOA accuracy**". The second represents the number of additional sources (insertions) that were detected, on average, per time block. We will call this quantity "**DOA insertion rate**".

To get the ideal DOA of the sources, we used information about the sound source activity (obtained from the power of the source signals – Section 2.4) and raw estimates of the DOA obtained by using the ideal number of sources (PNOS). Piecewise straight lines were fit to the contours of the raw DOA estimates in the intervals where each source is active. Video data were also used to check the instants where a source is moving.

4.2 Analysis of DOA estimation in different trials

Fig. 5 shows the DOA estimation performances (accuracies and insertion rates) for individual trials in office (OFC) and shopping mall (UCW) environments, for several parameter conditions related to the sound interval detection logic: with/without smoothing filter, with/without the threshold for forcing offset, with/without maximum number of onsets, and before/after tracking. For the computation of the MUSIC spectrogram, the following parameters were used: NFFT = 64, frequency range = 1 – 6 kHz, and fixed NOS = 2. Although larger NFFT provide slightly better performances, we used 64, because real-time can be achieved even when running in a 2GHz Centrino CPU.

The average DOA accuracies for speech sources are shown in the left part of the top panel in Fig. 5, and the DOA insertion rates are shown in the middle panel of Fig. 5, for each experimental condition and for each trial in OFC and UCW.

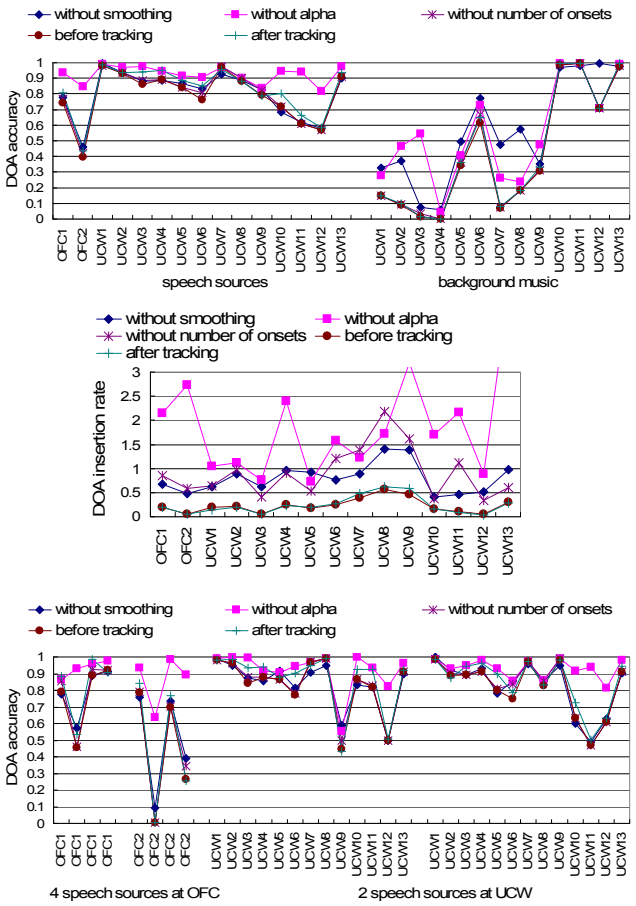

Fig. 5. DOA estimation performances (accuracies and insertion rates) for each source and each trial in OFC and UCW. For all trials, NFFT = 64, frequency range = 1 – 6 kHz, fixed NOS = 2, MUSIC delta-spectrogram Onset threshold = 1.0 dB, Offset threshold = -1.2 dB, and MUSIC power threshold = 2.0 dB.

We can first observe that "without alpha" (forcing an offset according to the onset MUSIC power plus a bias factor alpha) gives the best DOA accuracies. However, it also gives the worst insertion rates. The DOA accuracies of "without smoothing", "without number of onsets" and "before tracking" are very similar. However, a clear reduction in the insertion rate can be observed for

"before tracking", showing the effectiveness of both smoothing procedure and constraining the number of onsets per block. Finally, comparing "before tracking" and "after tracking", a slight improvement is observed in DOA accuracy for UCW3-6 and UCW10-11, while only a very small increasing in DOA insertion rate is observed in UCW8-9.

Regarding the ambient music sources, it can be observed in the right side of the top panel in Fig. 5 that the DOA accuracies were low in UCW1-4 and UCW7-9, since the robot was relatively far from the ceiling loudspeakers. DOA accuracies were almost 100 % in UCW10-13, when the robot was right under one of the loudspeakers, where the background music can be clearly considered as a directional source, while DOA accuracies show intermediate detection rates for UCW5-6, where the robot was relatively closer to one of the ceiling loudspeakers.

Regarding performances for individual sources, it can be observed in the bottom panel of Fig. 5 that the second and fourth sources in OFC2, the first source in UCW9, and the second source in UCW12 show lower DOA accuracy. An explanation is that these sources come from the back side of the robot, so that both power and directivity are lower than the sources coming from the front side.

5 Conclusion and Future works

Sound interval detection of multiple sound sources using a 3D-space sound directivity based on the MUSIC spectrogram was implemented and evaluated in our humanoid robot embedded in real noisy environments.

Evaluation of the proposed method showed lower insertion rates could be achieved. However, the detection accuracies also degraded in some of the trials where the loudspeaker is right over the robot. We are currently investigating the reasons of this degradation.

Finally, although the goal of the present work is to detect speech intervals, the technology here presented is to detect sound intervals. However, in robot applications, other modalities, such as vision, can be used to determine if the detected sound is speech or not. This will be scope of our future work. We are also planning the implementation and evaluation of sound source separation algorithms using the localization and sound interval detection results from the present work.

Acknowledgement

This work was supported in part by the Ministry of Internal Affairs and Communication, and by the Ministry of Education, Culture, Sports, Science and Technology.

References

1) F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application on automatic speech recognition," in *Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1013–1016.
2) K. Nakadai, H. Nakajima, M. Murase, H.G. Okuno, Y. Hasegawa and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 852–859.
3) S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics," in *Proc. of IROS 2007*, San Diego, CA, USA, 2007, pp. 2009–2014.
4) M. Heckmann, T. Rodermann, F. Joublin, C. Goerick, B. Schölling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. of IROS 2006,* Beijing, China, 2006, pp.368–373.
5) T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, B. Schölling, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.860–865.
6) J. C. Murray, S. Wermter, H. R. Erwin, "Bioinspired auditory sound localization for improving the signal to noise ratio of socially interactive robots," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 1206–1211.
7) Y. Sasaki, S. Kagami, H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 380–385.
8) J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *IEEE ICASSP 2006*, Toulouse, France, pp. IV 841–844.
9) B. Rudzyn, W. Kadous, C. Sammut, "Real time robot audition system incorporating both 3D sound source localization and voice characterization," Procs.of *ICRA 2007*, Roma, Italy, 2007, pp. 4733–4738.
10) C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.
11) T. Kanda, D. F. Glas, M. Shiomi, H. Ishiguro, and N. Hagita, "Who will be the customer?: A social robot that anticipates people's behavior from their trajectories," *Tenth International Conference on Ubiquitous Computing* (*UbiComp 2008*), 2008.