Evaluation of a MUSIC-based Real-time Sound Localization of Multiple Sound Sources in Real Noisy Environments

Carlos T. Ishi, Olivier Chatot, Hiroshi Ishiguro, and Norihiro Hagita, *Member, IEEE*

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. The effects of several parameters related to the MUSIC algorithm on sound source localization and real-time performances were evaluated, for recordings in different environments. Real-time processing could be achieved by reducing the frame size to 4 ms, without degrading the sound localization performance. A method was also proposed for determination of the number of sources, which is an important parameter that influences the performance of the MUSIC algorithm. The proposed method achieved localization accuracies and insertion rates comparable with the case where the ideal number of sources is given.

I. 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.

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.

In the present paper, we discuss about these two issues, by analyzing the effects of several parameters related to the MUSIC algorithm, on the sound localization accuracy and real-time performance.

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 the present work, 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. We evaluated the effects of the environment change on the MUSIC algorithm and proposed a method to improve sound localization robustness.

This paper is organized as follows. In Section II, descriptions about the hardware and data collection are given. In Section III, the proposed method is explained, and in Section IV, analyses and evaluation results are presented. Section V concludes the paper.

II. HARDWARE AND DATA COLLECTION

A. The microphone array

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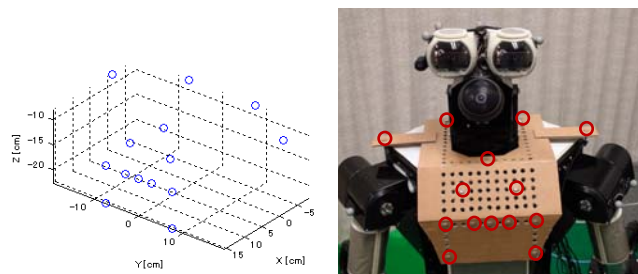

Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

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.

B. 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.

C. 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 is currently being executed [10]. 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 II.*D*.

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.

D. 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.

III. Proposed Method

A. The sound localization algorithm

Fig. 2 shows the block diagram of the implemented sound localization algorithm. The algorithm structure is basically the same of 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 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, the broadband MUSIC response by averaging the (narrowband) responses over a frequency range, and finally a peak picking in the MUSIC response to get the desired direction of arrival (DOA) of the sound sources.

In the proposed approach, some of the parameters are analyzed in order to obtain a real-time processing, and keeping the DOA estimation performance. We also analyze and propose a method for determining the number of sources,

which is an important parameter necessary for obtaining a good performance by the MUSIC algorithm. Each analyzed parameter is described in detail in the following sections.
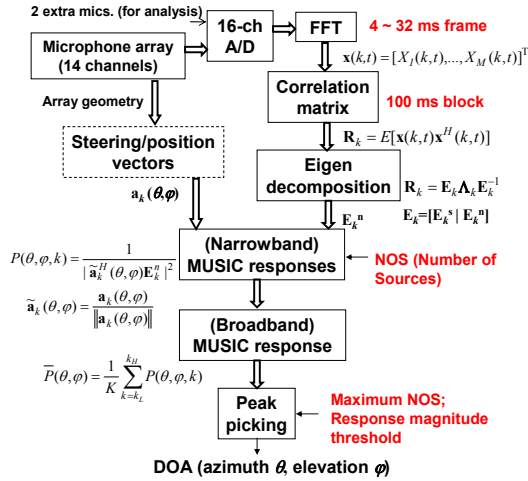

Fig. 2. The MUSIC-based sound localization algorithm, and related parameters.

B. The 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. 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.

C. Definition of 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 $512 \sim 1024$ points

is commonly used (corresponding to $32 \sim 64$ ms frame length at 16 kHz), we proposed the use of smaller FFT sizes ($64 \sim 128$). This will reduce the computation not only of the FFT stage, but also the subsequent correlation matrix, eigenvalue decomposition, and MUSIC response computations. Evaluation of the effects of the FFT size reduction is reported in Section IV.

In the next step of the MUSIC algorithm, a correlation matrix is averaged for the frames within a time block. A time block length of 1 second interval has been set in [1]. However, such a long block length could result in a low resolution in the DOA estimation, if the sound source is moving. In the present work, we decided to use a smaller time block length of 100 ms.

D. Estimation of the number of sources (NOS) for the MUSIC algorithm

For each time block, the number of sources (NOS) present in the input signals has to be attributed to the MUSIC algorithm, to decide how many eigenvectors have to be included in the computation of the MUSIC response.

Classical methods for estimating the number of sources use the eigenvalues obtained from the correlation matrix of the array signals. In theory, strong eigenvalues would correspond to directional sources, while weak eigenvalues would correspond to non-directional noise sources.

However, in practice, it is very difficult to distinguish between strong and weak eigenvalues, due to reflections and possibly due to the geometric imperfections of the array implementation. Fig. 3 shows the eigenvalue profiles (averaged over the frequency bins) for the recordings in different environments (OFC, UCW1 and UCW6), arranged by PNOS (ideal number of sources obtained using the power of the source signals).
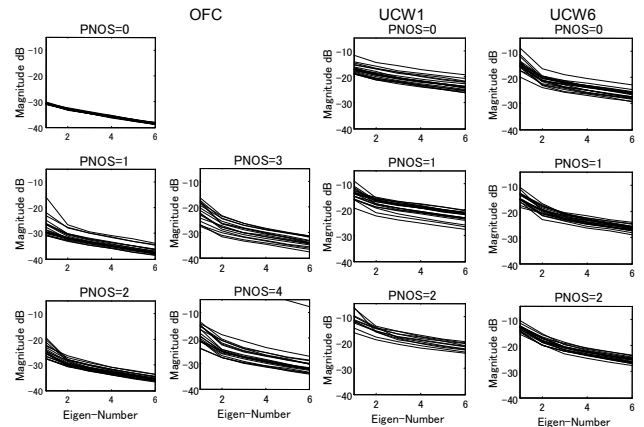

Fig. 3. Eigenvalue profiles arranged by the ideal number of sources (PNOS), in three different environments: OFC1, UCW1 (loudspeakers are far from the robot), and UCW6 (loudspeakers are close to the robot).

We can observe in Fig. 3 that 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. We can also

observe that there are many overlaps between different PNOS.

Fig. 3 also shows that the environment noise has a strong impact on the shapes of the eigenvalues. Both magnitude and slope of the profiles are affected. We can notice this by comparing the eigenvalue profiles of PNOS=0 for the recordings OFC1 and UCW1 in Fig. 3. The ones from UCW1 clearly have a higher magnitude than the ones from OFC1, due to the background music. Also, due to the varying nature of the background music, the profiles from UCW1 for PNOS=0 are not as densely packed as the ones from OFC1.

Furthermore, we notice that the proximity of the environmental music sources also affects the shapes of the profiles. The profiles from UCW6 for PNOS=0 are very similar in shape to the profiles from UCW1 for PNOS=1. This indicates that when the music source is near the robot, it is possible to treat it as an additional directional source. Otherwise, when the robot is far away from the loudspeakers, the environmental music becomes more non-directional. In the cases the music source is directional, it would be useful to localize it as well and to use it to improve a subsequent sound source separation process.

Considering the difficulties in estimating NOS from the eigenvalues of the spatial correlation matrix, we proposed a method of using 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 MUSIC response computation is limited to a fixed small number, the combination of frequency bins to compute the broadband MUSIC response may produce more peaks than the fixed number.

To avoid over-estimation of the number of sources, we set a threshold for the magnitude of the broadband MUSIC response, to determine if a response peak can be considered as a source.

E. Finding the directions of arrival (DOA) from the broadband MUSIC responses

Once the MUSIC response has been computed for each time block, it is possible to find the DOA by finding the local maxima of the response that have the highest magnitudes.

To find local maxima in the 2D (azimuth vs. elevation) MUSIC response, the following procedure was adopted. The algorithm starts by finding the local maxima that has the highest magnitude, recording its direction as one of the detected DOA. Then, a 2D Gaussian (azimuth vs. elevation) is subtracted from the response. This Gaussian, centered in the direction of the last detected DOA, has standard deviations that fit the usual shape of the response for one source and is scaled to match the magnitude of the response at its maximum. Subtracting this Gaussian emulates the removal of the source responsible for the strongest local maximum from the response. This is repeated until the number of DOA found is equal to NOS.

IV. ANALYSES AND EXPERIMENTAL RESULTS

A. 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**". And the third value measures the real-time performance, as the ratio between the actual processing time and the actual recording time. We will call this quantity "**real-time rate**".

The actual processing time was measured, by running all trials with an Intel Xeon processor running at 3 GHz.

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 II.D) 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.

B. The effects of the number of FFT points (frame length) on DOA estimation and real-time performance

Fig. 4 shows the DOA accuracies, the DOA insertion rates, and the real-time rates, as a function of different values of NFFT (number of FFT points) (NFFT = 64, 128, 256 and 512), for speech and music sources in OFC and UCW.
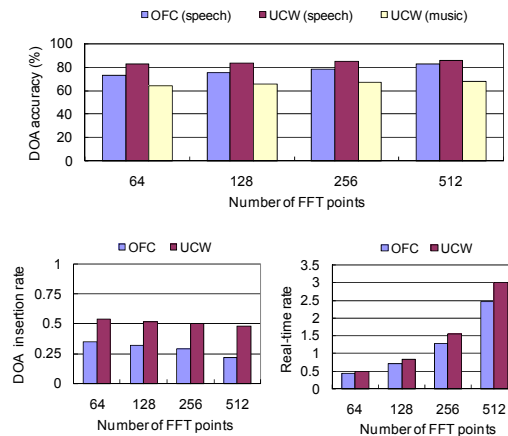

Fig. 4. DOA estimation and real-time performances as a function of number of FFT points, for OFC and UCW recordings. For all trials, frequency range = 1 – 6 kHz, and the ideal number of sources (PNOS) are provided.

From Fig. 4, we can observe that the DOA accuracies and the DOA insertion rates are almost the same for the different NFFT values. However, a considerably large reduction in

processing time can be observed for smaller NFFT values. We can observe that real-time processing can be achieved (i.e., real-time rate smaller than 1), for NFFT = 64 and 128. It was also confirmed that real-time processing can be achieved for NFFT=64, in a 2GHz processor. Since the DOA performance does not degrade much and computation time can be largely reduced by using smaller NFFT values, we decided to use a NFFT of 64 points (or equivalently a frame size of 4 ms) in all subsequent analysis.

C. The effects of the frequency range on DOA estimation

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).

Fig. 5 shows the effects of the frequency range on DOA estimation, for NFFT = 64, and using the ideal number of sources (PNOS). The results in Fig. 5 show that higher frequency boundaries provide better performances (higher DOA accuracy, and lower DOA insertion rates), since fricative sounds which contains higher frequency components can also be correctly detected.
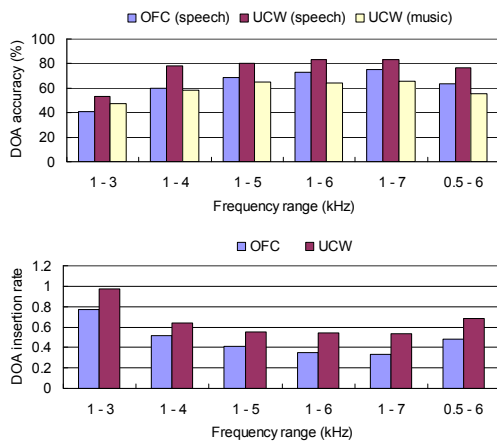

Fig. 5. DOA estimation performances as a function of the frequency range of operation, for OFC and UCW recordings. For all trials, NFFT = 64, and the ideal number of sources (PNOS) are provided.

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. This fact is reflected in the results of "1 – 6 kHz" and "0.5 – 6 kHz" in Fig. 5, where DOA accuracy is lower and DOA insertion rate is higher in "0.5 – 6 kHz", where frequency components lower than 1 kHz are included in the MUSIC response computation.

Although 1 – 7 kHz provided the best performance, we decided to use 1 – 6 kHz for the subsequent analyses, to avoid possible effects of spatial aliasing.

D. The effects of number of sources on DOA estimation

In this section, we analyze the three parameters involved in the estimation of the number of sources: the fixed NOS for the MUSIC response computation (fixed NOS), the maximum NOS detectable from the broadband MUSIC response (max NOS), and the threshold for the magnitude of the broadband MUSIC response.

First, we set the max NOS to 4 or 5, which is considered to be enough for human/robot interaction purposes. The fixed NOS can have a value smaller than or equal to the max NOS. Then, the threshold for magnitude of the broadband MUSIC response is set to avoid insertion errors, since we would be over-estimating NOS by setting a large max NOS value. However, a too large threshold would also cause deletions of the actual sources.

Fig. 6 shows the DOA estimation performance for several combinations of fixed NOS, max NOS, and MUSIC response magnitude thresholds.
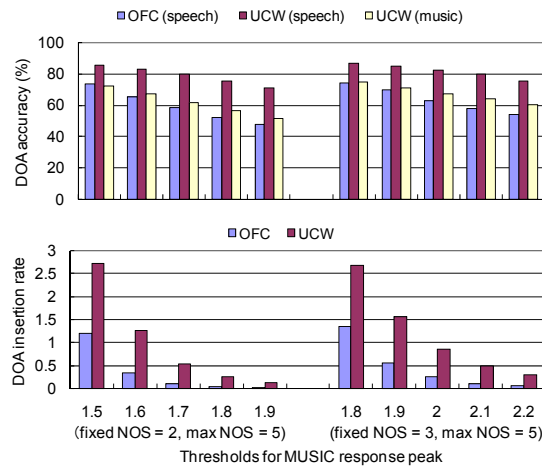

Fig. 6. DOA estimation performances as a function of the different values for the threshold for MUSIC response, for fixed NOS = 2 or 3, and max NOS =5. For all trials, NFFT = 64, frequency range = 1 – 6 kHz.

Fig. 6 shows the DOA estimation performance for different threshold values, for fixed NOS = 2 and max NOS = 5, a threshold of 1.7 is found to have a good balance between DOA accuracy and DOA insertion rate. For fixed NOS = 3 and max NOS = 5, a threshold of 2.0 ~ 2.1 is found to have a good balance between DOA accuracy and insertion rate.

E. Analysis of DOA estimation in different trials

Fig. 7 shows the DOA estimation performance for individual trials, fixing NFFT = 64, frequency range = 1 – 6 kHz, fixed NOS = 2, max NOS = 5, and MUSIC response threshold = 1.7.

Sources S2 and S4 in OFC2, S2 in UCW9, and S1 in UCW12 showed lower DOA accuracy, probably because these sources come from the back side of the robot, so that both power and spatial resolutions are lower than the sources coming from the front side.

Regarding the ambient music sources, the DOA accuracies

were low in UCW1 to UCW4 and in UCW9, since the robot was relatively far from the ceiling loudspeakers. DOA accuracies were relatively high for UCW5 to UCW8, where the robot was closer to one of the ceiling loudspeakers, while DOA accuracies were almost 100 % in UCW10 to UCW13, when the robot was right under one of the loudspeakers.
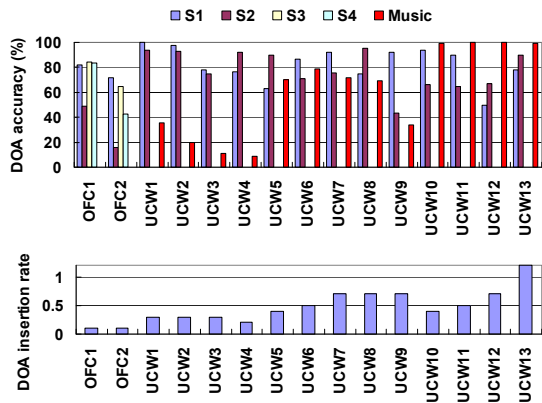

Fig. 7. DOA estimation performances for each source and each trial in OFC and UCW. S1 to S4 are speech sources, while M is the background music source. For all trials, NFFT = 64, frequency range = 1 – 6 kHz, fixed NOS = 2, max NOS = 5, and threshold for MUSIC = 1.7.

The larger DOA insertion rate in UCW13 was due to the misdetection of a source coming from 180 degrees azimuth and low elevation angles, as shown in the DOA estimations in Fig. 8. This seems to be a sidelobe effect of the source at 0 degrees azimuth or of the music source, rather than a reflection. Further analyses on the sidelobe effects are left for future work. The right panels in Fig. 8 show relatively successful results of DOA estimations for trial UCW8, where one of the sources is moving in front of the robot, and a directional music source is present in the first half of the trial.
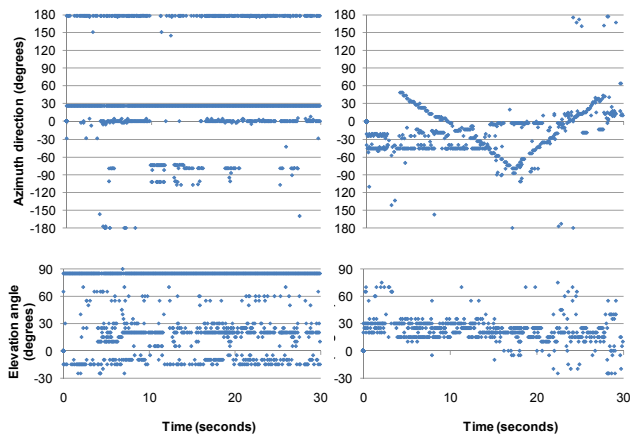

Fig. 8. DOA estimations of trial UCW13 (left) and UCW8 (right). In UCW13, music source is detected at about 90 degrees elevation 30 degrees azimuth. S1 source is detected at 0 degrees azimuth and moving from 15 to 25 degrees elevation at 6 to 7 seconds. S2 source is detected at about -80 to 85 degrees azimuth and 25 degrees elevation. In UCW8, source S2 is moving in front of the robot around 50 to -80 degrees azimuth, while source S1 is around -20 to 0 degrees azimuth. The music source was present in the first half at about -45 degrees azimuth, and the volume casually reduced in the second half. For these trials, NFFT = 64, frequency range = 1 – 6 kHz, fixed NOS = 2, max NOS = 5, and threshold for MUSIC = 1.7.

V. CONCLUSIONS AND FUTURE WORKS

A 3D-space sound localization of multiple sound sources based on the MUSIC algorithm was implemented and evaluated in our humanoid robot embedded in real noisy environments.

Evaluation results first indicated that reducing the FFT size to 64 (or equivalently reducing the frame size to 4 ms) was effective to allow real-time processing without a big degradation in the estimation of the directions of arrival (DOA) of sound sources. The evaluation of the proposed method of determination of the number of sources was also effective to keep a reasonable estimation performance, with a low insertion rate.

The evaluation results in the present work are based on the raw DOA estimation results, so that a post-processing for example by grouping and interpolating the detection results will probably increase the accuracy numbers. Post-processing of the detected DOA is scope of our next work.

For future works, we are planning the implementation and evaluation of sound source separation algorithms using the localization results from the present work.

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 the 2006 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 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 the 2007 IEEE/RSJ, Intl. Conf. on Intelligent Robots and Systems*, 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 the 2006 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. 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 the 2006 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 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 the 2006 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 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 the 2006 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, 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," *2007 IEEE Intl. Conf. on Robotics and Automation*, Roma, Italy, 2007, pp. 4733–4738.

[10] 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.