A Robust Real-Time Sound Source Localization System for Olivia Robot

Shengkui Zhao*, Eng Siong Chng*, Nguyen Trung Hieu† and Haizhou Li†
* Nanyang Technological University, Nanyang Avenue, 639798 Singapore
† Institute for Infocomm Research, 1 Fusionopolis Way, 138632 Singapore
E-mail: skzhao@ntu.edu.sg, aseschng@ntu.edu.sg Tel: +65-6790 6200
E-mail: thnguyen@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg Tel: +65 6408 2773

Abstract—In this paper, we present a robust real-time sound source localization system implemented on a social robot platform developed in Institute for Infocomm Research, Singapore. The audio localization system provides the robot with auditory senses and enables the robot to direct its face to a speaker outside its frontal vision system. As the localization system exploits time difference of arrival (TDOA), the placement of the 8 microphone system array is crucial. This paper discusses the configuration and implementation of our system for the Olivia robot platform for accurate 3D localization under high babble noise condition.

I. INTRODUCTION

When deploying mobile robot to interact with human beings, only the visual sense of the robot may not be enough, for example, in the cases of poor lighting or if the interested subject is located outside visual field of the visual sensors. Hence, the auditory sense of a robot is also important as it enables the robot to react to sound events and directs the robot to turn to the sound source. To this end, there have been increasing research interests in the field of sound source localization for mobile robot. For example, the robot ROBITA [5] uses 2 microphones to follow a conversation between two persons based on the head related transfer functions (HRTF) approach. The humanoid robot SIG [6], [7] uses two pairs of microphones to locate the sound source in 3D space based on both the time difference of arrival (TDOA) and intensity level difference (ILD). For good robustness in noisy condition, an array of 8 microphones was mounted on an ActiveMedia Pioneer 2 robot platform for 3D localization based only on the TDOA in [4]. The HARK robot audition system in [17] provides the sound source localization with an array of 8 microphones based on the TDOA and steered beamformer. In [8], the placement of 4 microphones on an IRP-2004 robot platform was studied for 3D localization in anechoic chamber. In addition, the Humanoid HRP-2 reported in [2] uses the fusing of the audio and video information, where audio information is used in the detection of speech events and video is used for human tracking.

However, it is difficult to apply the above mentioned methods on a robot with human-like head working in a general environment. It is known that using a pair of microphones on a robot is a difficult challenge to match the hearing capabilities of humans. The current sound source localization methods based on a pair of microphones are based on the known robot design and working environment. The methods based on more microphones either didn’t address the placement of the microphones due to the indirect path problem or didn’t provide a general configuration of the microphone array suitable for human-like robot head. In this paper we present a robust real-time sound source localization system with 8 microphones for the Olivia robot, a social robot for intelligent human-robot interactions. The indirect path problem is addressed by carefully configuring the microphone array. Our experiments show that the system is robust to noisy condition and capable of accurately localizing the azimuth angle ($0^\circ \sim 360^\circ$) for the sound source coming from all around the robot and the elevation angle ($-45^\circ \sim 45^\circ$) for the sound sources in front of the robot within a view of $60^\circ$.

The rest of the paper is organized as follows. In Section II, we first give a short description of the TDOA estimation and the direction estimation. Then, in Section III, we present our new configuration of microphone array on Olivia robot platform and discuss the selection of the microphone pairs for sound source coming from all around the robot. In Section IV, we conduct experiments in real-world environment and report the performance of sound source localization system implemented on Olivia robot platform. In Section V, we give a summary of the paper.

II. THE TDOA AND DIRECTION ESTIMATIONS

A. The TDOA Estimation

The very popular method for the sound source localization is based on the TDOA estimation. The time difference measurement between a pair of microphones is usually realized by selecting the peak index of the cross-correlation between the signals perceived by the two microphones as defined below:

$$\hat{\tau}_{ml} = \arg \max_{\tau_{ml}} R_{x_mx_l}(\tau_{ml}), \ m, l = 1, 2, ..., M$$

(1)

where $M$ denotes the number of microphones, $\hat{\tau}_{ml}$ is the TDOA estimation, and $R_{x_mx_l}(\tau_{ml})$ is the cross-correlation function of the microphone signals $x_m(n)$ and $x_l(n)$.

To improve the estimation accuracy in noisy and reverberant environment, the traditional used method is the Generalized Cross-Correlation (GCC) method [10-12], which is defined as

$$R_{x_mx_l}(\tau_{ml}) = \sum_{k=0}^{N-1} G(k)X_m(k)X_l^*(k)e^{j2\pi k\tau_{ml}/N}$$

(2)
Fig. 1. The TDOA estimation based on far field model.

Fig. 2. Illustration for the computation of azimuth angle $\theta_1$ and elevation angle $\theta_2$.

where $X_m(k)$ and $X_l(k)$ is the short time discrete Fourier transforms of $x_m(n)$ and $x_l(n)$, respectively. $X_m(k)X_l^*(k)$ is the short time cross-spectrum of $x_m(n)$ and $x_l(n)$. $G(k)$ denotes the general frequency weighting function. The selection of $G(k)$ is based on criteria of optimizing certain performance and has been investigated by many researchers [13-15]. Among these approaches, the phase transform (PHAT) using the following weighting function

$$G(k) = \frac{1}{|X_m(k)||X_l(k)|}$$

is known to perform in general better than the other approaches in the presence of reverberations. It is because only the phase of the cross-power spectrum is used and the amplitude is discarded. This means that the phase with the most number of occurrences naturally produces the highest peak in the cross-correlation function.

B. Direction Estimation of Sound Source

After the TDOA estimation is performed, the next step is to obtain the position or direction of the sound source. In general, the direction of sound source in 3D space is represented by two angles, i.e., the azimuth angle and elevation angle. With the far-field assumption, the source radiates a plane wave having the waveform propagates through the non-dispersive medium air as shown in Fig. 1, a linear equation can be constructed

$$\vec{d}_{ml} \cdot \vec{u} = \|\vec{x}_{ml}\| \|\vec{u}\| \cos \theta = d_{ml}$$

where $\vec{x}_{ml}$ is the vector from Mic $m$ to Mic $l$, $\vec{u}$ is a unit vector pointing in the direction of the sound source, and $d_{ml}$ is the distance delay between Mic $m$ and Mic $l$ from the sound source given by $d_{ml} = c \cdot F \cdot s \cdot \tau_{ml}$ with the speed of sound $c$ and the sampling rate $F$. By selecting the Mic $l$ as the reference microphone and the rest microphones with direct paths for $m$, from (4) we have a system of linear equations. Let $\vec{u} = (x_u, y_u, z_u)$. By solving for the unity vector $\vec{u}$, the azimuth angle $\theta_1$ along $x$ axis and elevation angle $\theta_2$ along $z$ axis as shown in Fig. 2 can then be computed as follows:

$$\theta_1 = \cos^{-1} \frac{x_u}{\sqrt{x_u^2 + y_u^2}}, \quad \theta_2 = \cos^{-1} \frac{\sqrt{x_u^2 + y_u^2}}{\sqrt{x_u^2 + y_u^2 + z_u^2}}.$$  \hspace{1cm} (5)

Noted that although the computation is made under the far-field assumption, for the near field cases, the method is still applicable since the distance deviation corresponding to the angular error is often smaller than human face.

III. CONFIGURATION OF MICROPHONE ARRAY ON THE OLIVIA ROBOT

The Olivia robot platform is developed in the Robotics Lab, I²R, Singapore. The height of the robot is $1.65m$. The head of the robot has a 3D motion varied $0^{\circ} \sim 180^{\circ}$ for azimuth angle and $-20^{\circ} \sim 20^{\circ}$ for elevation angle. The Olivia robot is designed to be a receptionist robot, therefore, the robot should be able to localize the visitors based on voices when being spoken to. The other functionalities the robot will perform include tracking visitors by vision technology, conversing with visitors by speech recognition technology, as well as gesture learning. To achieve the purpose of 3D sound source localization, the array of 8 microphones are mounted on the robot head as shown in Fig. 3. The relative coordinates of the microphones are listed in Table I. The signal acquisition is performed using two M-Audio Delta sound cards installed on the robot platform. All the processes are running on the robot platform.

Due to the shape of the robot head and direct path nature of the TDOA estimation, a practical issue is to configure the microphone array on the robot head such that more
microphones will have direct paths to the sound source. To address this issue, the horizontal direction is uniformly divided into 6 regions as shown in Fig. 4 from the top view. Each region has a range of 60°. It is seen that based on this configuration we can select at least 4 microphones for each region under the far-field assumption.

Our program localizes a sound source as follows. First, the region of sound source is determined based on the signs of the TDOA estimations of several microphone pairs. Then, a subset of microphones and a reference microphone are selected corresponding to the region. The microphones within or close to the region are selected for the region, and the central microphone is selected as a reference microphone. The TDOA estimations between the reference microphone and the other microphones from the subset are obtained. Finally, the elevation angle and/or the azimuth angle are computed from the linear equations constructed based on the TDOA estimations and the coordinates of the microphones. As the microphones in regions 4, 5 and 6 are almost on the same plane, only the azimuth angle is computed. This is not a major limitation as our design favors the robot to have a better focus for its front facing. For the case when the speaker’s speech is from the back of the robot, then the system need only to detect its horizontal direction and direct the head to turn.

Table 2 shows the region selection in details. The TDOA value $\tau_{27}$ between Mic 2 and Mic 7 is firstly computed. If sign of $\tau_{27}$ is non-negative, the sound source is considered coming from the front of the robot. Otherwise, the sound source is assumed to be from the back of the robot. If the sound source is from the front, the TDOA estimations between Mic 2 and Mic 4, and Mic 2 and Mic 5 will be computed. If the sign of $\tau_{24}$ is non-negative and the sign of $\tau_{25}$ is negative, the sound source is determined in the region 1. Therefore, the microphone numbers 1, 2, 3, 5, 6 and the reference microphone number of 5 are selected for the 3D localization. The similar steps are also taken for the other cases. The details of the microphone selection are shown in Table II.

The experiments showed that the accuracy of the region selection is over 98% for a very noisy environment with SNR = 5dB under babble noise condition. In general, even if the direct path is blocked, the sign of the TDOA estimation is unlikely to be corrupted.

IV. EXPERIMENTAL RESULTS

The system is evaluated in a normal size room of 6m × 3.5m × 2.5m. The room reverberation is around 0.15s. The testing were carried out as follows. We used a person with height of 1.73m speaking to the robot with same sentence for different positions in the same test. For Test A, the distance between the speaker to the robot is 2m. The positions of the speaker varied 0° ~ 360° for azimuth angle at each 30 degrees. The total length of recorded data in all positions is 288 seconds with 57.8% of speech. For Test B, the distance is changed to 1m and the positions varied by 45 degrees each time. The total length of recorded data in all positions is 96 seconds with 36.5% of speech. Both tests were repeated under 4 types of noise. The ambient noise are from air-cons and computer fans in the room. The babble noise was simulated by playing NOISEX-92 babble noise signal. The human noise includes the sounds of typing keyboard, clicking table, opening doors and human whispering. All the data were processed with a sampling rate of 48kHz on a frame by frame base. Beside the 8 channels for localization, a close talk microphone was also recorded for the ground truth. We used a frame length of 160ms without overlapping. The reason of choosing the long frame length is for fast processing and high accuracy of GCC PHAT.

The performance of the system were measured with frame base by the precision and recall only for the speech frames.
and defined as

\[
\text{Precision} = \frac{N_F}{N_T + N_F} \times 100\%,
\]

\[
\text{Recall} = \frac{N_F}{N_T + N_M} \times 100\%
\]

where \(N_T\) is the number of speech frames with which the system correctly detected the direction of sound source, \(N_F\) is the number of speech frames with which the system has false alarm, and \(N_M\) is the number of speech frames where the system missed. In practice, a high precision is of much greater importance as compared to recall. Therefore, our Voice Activity Detector (VAD) system based on periodicity modified from [18] detects only voiced frames. As the TDOA estimation is more robust on the voiced frames, we only evaluated the sound source localization for voiced frames. The average precision and recall from the experiments are shown in Table III. It is seen that system achieves a very high precision even in high babble noise and human noise conditions. Although the recall is low (around 30%-50%), this recall rate is sufficient for the robot’s localization purpose. Note that for human conversation voiced sounds often occur even in very short sentences. In the experiments, for the data collected in all positions the number of voiced frames is 71\% of speech frames in Test A and 82\% in Test B.

From Table III, the mean error of angle of the localization system is from 0.15° to 4.28°. For such angular errors, the distance deviation of the sound source localization is often smaller than human face. These results show that the system can perform well even in the environment with high babble noise or human noise.

Our testing shows that the system is capable of accurately estimating the elevation angle if the sound source is within the front horizontal region 2 from 60° to 120° and the elevation range from −45° to 45°. Outside this range, the accuracy of the elevation estimation decreases. This is because the directions of sound source will deviate from the direct path due to the spherical robot head. As mentioned early, this is not a major limitation as our design favors the robot to have a better focus for its front facing. In addition, the Olivia robot only need to move from up to down within the range of 20° to −20°. Although we only show the testing results for the distance of 1m and 2m. In fact, the system can function satisfactorily for distances varying from 0.5m to 5m.

V. CONCLUSIONS

In this paper, we have presented a robust real-time sound source localization system on a real social robot. An array of 8 microphones mounted on the robot head are configured such that the system is able to localize a sound source in 3D space based on the TDOA method. Experiments conducted in real environment show that the system achieves very high precision and also good recall for human voice in environments with high babble noise or human noise.

REFERENCES