2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Mobile Bot Rotation Using Sound Source Localization And Distant Speech Recognition

Swapnil Sontakke[1] Pradyoth Hegde[1] Prashant Bannulmath[2] and Deepak K T[2]

Abstract— **In the last few years, mobile robots such as floor cleaners, assistive robots, and home telepresence have become an essential part of our day-to-day activities. In human-robot interaction, speech is the preferred way of communication, especially in indoor environments. This paper proposes a speech module to rotate the mobile robot. It has two components, namely, a distant automatic speech recognizer and a sound source localizer. To build distant speech recognizer, far-field speech data is collected at 1, 3, and 5-meters distances. The model performs well even at a 5-meters distance with a Word Error Rate of 40.38% and a Character Error Rate of 28.85%. The direction of arrival of the speech signal is computed from the 4-mic circular array microphone. The speech module is integrated with the Robot Operating System and physically demonstrated on Turtlebot3 Waffle Pi. It is observed that the speech recognizer and sound source localizer work well in the reverberant indoor environment with a small single-board computer.**

I. INTRODUCTION

Human-Robot Interaction (HRI) in robotics and artificial intelligence domain stands as a pivotal and rapidly evolving area that holds the promise of reshaping the way humans and machines collaborate. It ventures into the domain of cognitive sciences, psychology, and user experience design. Mobile robots are designed majorly for navigating and to do specified tasks. Technically, the mobile robots include different sensors such LiDAR, Depth sensor, RGBD camera, and microphone. As conceptualising the intelligent mobile robot, it is expected to have the features similar to the humans such as seeing, listening, understanding, and responding.

For effective human-robot interaction, robots should possess the hearing and speaking capabilities [1]. When mobile robots are considered in indoor environment, speaker localization and tracking using speech are essential for various human-robot interaction applications [2] and navigation tasks [4][5]. Compared to speech based recognition and localization systems, many works on recognition and localization have been done using vision sensors. However, mobile robots using such sensors leverage only vision modality and are not able to detect the sound emitting sources [3]. In human-robot interaction this becomes the limitation as the mobile robots are limited to vision sensors and are not able to perceive the world as humans do using their multiple sensors.

[1]Department of Data Science and Intelligent Systems, Indian Institute of Information Technology, Dharwad, Karnataka, India sontakke.swapnil31@gmail.com, pradyothhegde@gmail.com

[2] Department of Electronics & Communication Engineering, Indian Institute of Information Technology Dharwad, Karnataka, India prashantb@iiitdwd.ac.in, deepak@iiitdwd.ac.in

In speech based mobile robot rotation, two tasks can be identified related to speech domain. They are automatic speech recognition (ASR) and sound source localization (SSL). ASR is used to recognize the speech and convert the speech signals into text. The sound source localization does the job of detecting the angle of the speaker.

In this work the speech is used as a sensory information in the mobile robot for effective human-robot interaction in the indoor environment. The inclusion of speech along with Direction of Arrival (DOA) of speech contains meaningful information with which robots can make intelligent decisions for numerous indoor robotic applications. The work of this paper are summarized as follows:

- **Distant Automatic Speech Recognizer:** An Automatic Speech Recognition model is trained on the curated distant Hindi speech data. The speech data is collected at 1, 3 and 5 meters distances in natural indoor environment.
- **Speech Module:** Speech module consists of distant ASR and angle received from SSL. The speech module understands the command and outputs the angle to rotate.
- **Rotating mobile robot:** The speech module is integrated into the Robot Operating System (ROS) installed on Raspberry Pi of Turtlebot3 Waffle Pi.

The remainder of this paper is organized as follows. In Section II, an overview of the related work to establish our approach in relation to integrating SSL and distant ASR is presented. Section III explains the integrated system structure and approach implemented in this work. In Section IV, performed experiments, obtained results and ablation study are discussed. Section VI concludes the paper with future work directions.

II. RELATED WORKS

Sound Source Localization plays a crucial role in enhancing Human-Robot Interaction by enabling robots to perceive and respond to acoustic cues in their environment. Earlier works in this domain predominantly utilized signal processing techniques to estimate the direction of sound arrival and localize the source [1]. Notable techniques include beamforming, Generalized Cross-Correlation with Phase Transform (GCC-PHAT), Multiple Signal Classification (MUSIC), among others. For instance, Sasaki et al. [6] engineered a 32-channel microphone array for the Nomad-XR4000 mobile robot, employing triangulation for localization and sound source separation.

Further, X. Li et al. [7] introduced a method based on the Direct-Path Relative Transfer Function (DP-RTF) to address

Sound Source Localization with a robot head, utilizing audio and audio-visual datasets for experimentation. Subsequently, work [8] extended this by incorporating DP-RTF estimation, multiple speaker localization, and variational tracking, particularly tailored for multiple speakers in reverberant environments. J. H. DiBiase et al. in [9] explored source localization methods presenting the SRP-PHAT algorithm and testing it with a 15-channel microphone array without real-time mobile robot integration. J. M. Valin et al. [10] proposed a system for simultaneous sound source separation using Geometric Source Separation and post-filtering methods within the context of mobile robot audition.

Y. Tamai et al. [11] employed a three-ring microphone array for sound localization and separation, employing Delay and Sum Beamforming (DSBF) and Frequency Band Selection (FBS). A multi-microphone speech recognition system targeting three simultaneous speakers was developed using Geometric Source Separation and post-filtering techniques [12]. In the realm of dynamic sound source tracking, Q. V. Nguyen et al. [13] applied a mixture Kalman Filter (KMF) for intermittent moving sound sources in a reverberant environment, although primarily demonstrated through simulation. Further, efforts have been made to improve Audio Speech Recognition (ASR) performance for effective human-robot interaction [14].

Expanding the horizon, C. Evers et al. [15] introduced a Simultaneous Localization and Mapping (SLAM) algorithm utilizing acoustic signals, mapping 3D positions of multiple sound sources, albeit within a simulated environment. Additionally, S. Michaud et al. [16] proposed cooperative sound mapping using microphone arrays on two robots, sharing a common reference map. The fusion of audio and visual data has emerged as a powerful approach in HRI. T. Zhang et al. [17] presented an audio-vision fusion method that integrates sound source direction into an RGB-D image to address dynamic obstacles in SLAM. Furthermore, Abdelrahman Younes et al. [18] proposed audio-vision fusion coupled with reinforcement learning for robot navigation, accommodating both static and dynamic sound sources.

S. Majumder et al. [19] introduced active audio-visual source separation as a reinforcement learning problem, demonstrating its efficacy in simulated 3D environments. Lastly, a multisensory simulation platform named SONICVERSE was presented in [20], designed to train agents with both visual and auditory capabilities, thereby advancing the frontier of perceptual intelligence in robots. These works collectively highlight the evolving landscape of sound source localization. Understanding the literature, the present work ventured onto building a speech module with distant automatic speech recognition and sound source localization techniques and integrate it to physical robot.

III. APPROACH

This section describes the proposed approach for the mobile robot localization using Distant ASR and SSL integration. The problem of localization without considering the map of the environment can be stated as follows.

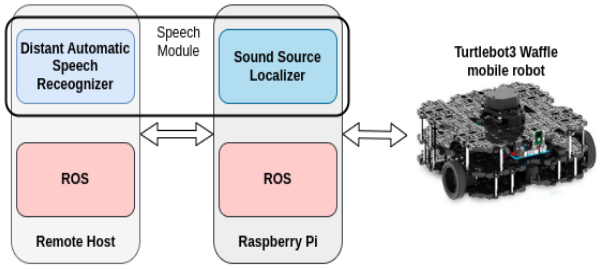

Fig. 1. Illustration of the integration of speech module with ROS and mobile robot.

Let, a vector $R = [C_i, C_j, 0]$ is the current orientation of the mobile robot in the given indoor environment. Here, front facing side of the robot always points the angle 0 when the robot is stationary. The 4-channel circular array microphone is installed on the mobile robot and is adjusted with respect to the front side of the mobile robot. This assures the 0 degree angle of robot and 0 degree angle of 4-channel circular array microphone points to the same direction. This helps robot get accurate sound source.

Let, a vector $H = [(H_x, H_y), \theta]$ is a sound source position i.e. human or sound emitting device with angle θ relative to the position of the microphone array.

Given speech as sensor data either by direct human or a recorded voice through speakers, and captured by 4-channel circular array microphone, the goal G is to localize the sound source and further rotate the mobile robot in the direction of sound source θ. This could be written as, $G = R_\theta[C_i, C_j, \theta]$

Further, the goal of speech module is to detect the DOA (θ) of a sound source H, record the audio signal $S(H)$ and receive the $\theta_{(DASR)}$ after processing and generate a relative angle $\theta_{(RA)}$. Finally, publish the angular velocity commands,

$$\omega(x) = \omega(y) = 0 \ rad/sec \tag{1}$$

$$\omega(z) = p \ rad/sec \tag{2}$$

Equation 1 sets the angular velocity of the robot with respect to x and y axes to 0. Equation 2 sets the angular velocity with respect to z-axis. Here, *p* is set manually. The remainder section discusses the methodology implemented to solve the above stated problem.

A. Overview

The proposed approach integrates Sound Source Localization (SSL) and distant Automatic Speech Recognition (ASR) technologies into the Turtlebot3 Waffle Pi mobile robot model. As illustrated in Figure 1, the system operates within the framework of the ROS. The hardware configuration of the Turtlebot3 Waffle Pi includes a circular array of microphone, serving as an audio sensor.

The microphone array captures speech at periodic intervals, yielding the DOA in degrees as an input parameter for ROS subscriber. Subsequently, the recorded audio data is transmitted to a remote server for transcription. Based on

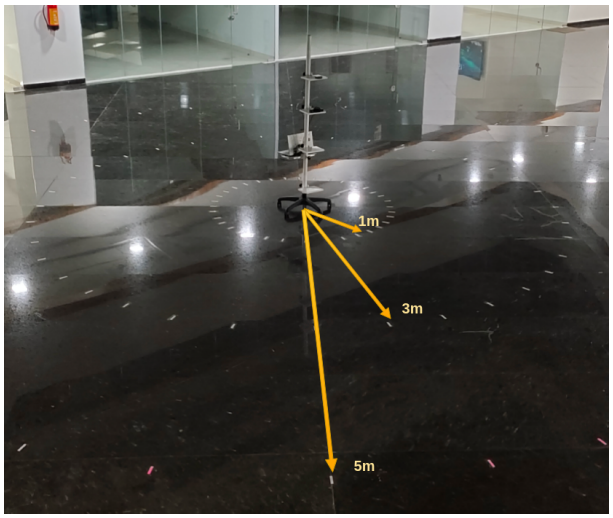

Fig. 2. Data Collection setup for the speech recognition system. At the center a recording sensors are placed. Three circles of 1, 3, 5 meters radii are pointed in the outer part.

predefined keyword values, the server generates a response in the form of an integer angle, which is then relayed back to the ROS system running on the Turtlebot3.

The ROS system combines the received response from the remote server with the input angle derived from the microphone array. The final evaluated value is then used by ROS to generate velocity commands for the Turtlebot3. These velocity commands are then used to rotate the Turtlebot3 in the appropriate direction.

B. Speech Data Collection

The training dataset for the speech module was collected in a natural reverberant environment, taking into account the requirements of the mobile robot to work within indoor settings. The dataset collection process was designed to cater to both SSL and distant ASR tasks. To capture the acoustic landscape, speech data was recorded at three distinct distances: 1, 3, and 5-meters, arranged in a circular configuration. Within each circular configuration, eight evenly distributed at 45-degree intervals discrete sound source positions were selected (i.e., 0, 45, 90, 135, 180, 225, 270, and 315 degrees). In total, this approach resulted in 24 distinct configurations. The speech data corpus consisted of contributions from 50 unique speakers, totaling 11 hours of recorded speech in Hindi language. Figure 2 shows the data collection setup and configurations.

C. Speech Module

The speech module has two key components that process speech and generate the required velocity commands for the mobile robot.

1) Sound Source Localizer: In robot audition, accurate estimation of a sound source is an essential task. In this work, a 4-channel circular array microphone is used to estimate the DOA. DOA estimation is performed using the Generalized Cross-Correlation with Phase Transform algorithm. GCC-PHAT effectively calculates time delays, allowing for the estimation of the sound source's angle in relation to the robot's orientation.

2) Distant Automatic Speech Recognizer: This work presents a distant speech recognizer that receives the speech from the microphone, transcribes it and extracts the keywords of interest essential for robot rotation. The ASR used in this work has two components: the acoustic model and the language model. The acoustic model is built using state-of-the-art wave2vec2.0 architecture which uses both self-supervised and supervised learning as shown in the Figure 3. In this architecture Convolutional Neural Network (CNN) takes the raw speech waveform (X_t) and encodes it to latent speech representations (Z_t). These representations are further fed to the transformer and a quantizer jointly, to generate the contextualized representations (C_t) and quantized representations (Q_t) which are discretized version of Z_t, respectively. Wave2vec2.0 solves the problem of speech recognition by training a contrastive task over masked latent speech representations and simultaneously learning the quantization of latents shared across selected languages [24]. The model is pre-trained via contrastive tasks where distractors are separated from true latent. After pre-training, the model is fine-tuned with distant speech data collected in house in Hindi language. The output of the fine tuned acoustic model is fed to the statistical based KenLM language model to decode the Hindi transcripts.

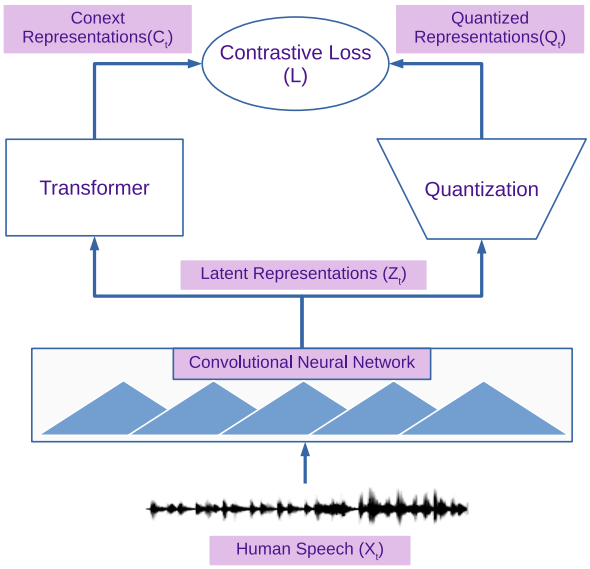

Fig. 3. Architecture of wav2vec2.0 model used for distant automatic speech recognition.

D. ROS Publisher and Subscriber Implementation

The ROS serves as a meta-operating system, offering a diverse array of services tailored for the development of robotic software. Within the ROS framework, nodes represent executable entities designed for facilitating seamless message exchange among various nodes. In ROS, there are two fundamental node types: the publisher and the subscriber.

Algorithm 1: publishInfo

Data: Sensor Data
Result: Computed Angle, .wav audio file
while *isRosRunning == True* **do**
  if *isMicAvailable == True* **then**
    if *isVoiceDetected == True* **then**
      $DOA\_angle \leftarrow angle\ from\ mic\ array$
      $rec\_data \leftarrow recorded\ audio\ in\ .wav$
      $cmd\_angle \leftarrow DASR(rec\_data)$
      $diff \leftarrow abs(DOA\_angle - cmd\_angle)$
      if $cmd\_angle == -1$ ***or*** $(diff \geq 0$ ***and*** $diff \leq 20)$ **then**
        publish(DOA_angle)
      else
        publish(cmd_angle)
      end
    else
      wait for voice detection
    end
  else
    wait for Mic availability
  end
end

Algorithm 2: subscribeInfo

Data: Computed Angle
Result: Velocity Commands
while *isRosRunning == True* **do**
  $A\_SPEED \leftarrow variable\ rad/sec$
  $PI \leftarrow 3.1416$
  /*(cmd_angle is received from publishInfo)*/
  $rt\_angle \leftarrow cmd\_angle$
  $linear.x \leftarrow linear.y \leftarrow linear.z \leftarrow 0$
  if $rt\_angle \geq 180$ **then**
    $rl\_angle \leftarrow ((360 - rt\_angle) * 2 * PI)/360$
    $angular.x \leftarrow angular.y \leftarrow 0$
    /*(Clockwise rotation)*/
    $angular.z \leftarrow -abs(A\_SPEED)$
  else
    $rl\_angle \leftarrow (rt\_angle * 2 * PI)/360$
    $angular.x \leftarrow angular.y = 0$
    /*(Anti-clockwise rotation)*/
    $angular.z \leftarrow abs(A\_SPEED)$
  end
  $start\_time \leftarrow current\ timestamp\ in\ seconds$
  $st\_angle \leftarrow 0$
  while $st\_angle \leq rl\_angle$ **do**
    publish($angular.z$)
    $ct\_time \leftarrow set\ current\ timestamp\ in\ sec$
    $st\_angle \leftarrow A\_SPEED * (ct\_time - st\_time)$
  end
end

In this work, a publisher node is tasked with the collection of sensor data, and a subscriber node actively listens to the published sensor data. This sensory information is utilized to make informed decisions, including the generation of velocity commands necessary for real-world mobile robot rotation. Algorithm 1 explains the working of the publisher node. The subscriber node is shown in Algorithm 2, that receives the output from publisher and generate the velocity commands for the robot.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

The hardware configuration of the hardware involved in the present is listed below:

1) Mobile Robot Configuration: Turtlebot3 Waffle Pi model consists Raspberry Pi 4 having 8GB RAM, 32 GB SD Card and running ROS 1 on top of the Raspbian OS.

2) Microphone: The Seed studio ReSpeaker USB 4-channel circular array microphone is used as an audio sensor to record speech. The channels are positioned in the angles 0, 90, 180 and 270 degrees. The microphone is mounted at a height of 0.125 meters from the robot's base. The array microphone is 7.5 cm in diameter.

3) Remote Server Configuration: Remote server is a remote host for Turtlebot3's Raspberry Pi 4. A laptop running Ubuntu 20.04 LTS operating system and ROS on top of it is used as a remote host. The server has an Intel i5 processor, 16 GB DDR4 RAM, and 256 GB Solid State Drive. Figure 4 shows the experimental setup for the robot localization in indoor environment.

B. Distant ASR

The objective was use the collected data to train and test a distant speech recognition model for robot rotation. For distant ASR training, recorded beamformed Hindi speech data of 45 speakers contributing a total of 10 hours is utilized. The validation dataset comprises 1 hour of speech data from 5 speakers. To build a distant automatic speech recognition system, a pre-trained model CLSRIL-23 [22] which is trained on 10,000 hours of 23 Indic languages is used. The model is fine-tuned with the collected distant Hindi speech data. An additional layer of fully connected layer is added on the top of wav2vec2.0 transformer model. The performance of the ASR is measured in terms of Word Error Rate (WER) and Character Error Rate (CER). The lowest WER on validation set of 40.38% is achieved using the language model and is shown in Figure 5. Figure 6 shows the decrease in training loss and validation loss.

The results of the fine-tuned model with and without language model (LM) are compared with the baseline model by Vakyansh [23] which is fine tuned with near field speech. Table I shows the comparison of base, distant ASR without language model and distant ASR with language model in terms of WER and CER.

C. Integration of Speech Module and ROS

In the context of real-world performance assessment, the integrated system's operational cycle starts with the speech

Fig. 4. Mounting of 4-channel circular array microphone on the Turtlebot3 Waffle Pi mobile robot.

Fig. 5. Distant ASR model performance evaluation graph.

TABLE I

PERFORMANCE COMPARISON OF DISTANT AUTOMATIC SPEECH RECOGNITION SYSTEM WITH AND WITHOUT ADDITION OF LANGUAGE MODEL.

Performance Criterion	Baseline Model	Distant ASR model without Language Model	Distant ASR with Language Model
WER (%) ↓	62.19	47.27	**40.38**
CER (%) ↓	35.48	**26.38**	28.85

detection, progressing through to its completion upon the successful rotation of the mobile robot. In the domain of human-robot interaction, metric for evaluation lies in the system's responsiveness. Specifically quantified by the time taken from the moment speech is received to the completion of the entire sequence. The testing setup is shown in the Figure 4. In this study, we test the system's overall performance by quantifying the time required for key stages: the processing of speech upon microphone reception, the generation of DOA estimations, and finally rotate the robot after generating the velocity commands for the mobile robot[1].

[1]The demo video is available at https://drive.google.com/file/d/158bNyiV69JpNxgotYE6C4TTLOlKvPK_z/view?usp=drive_link

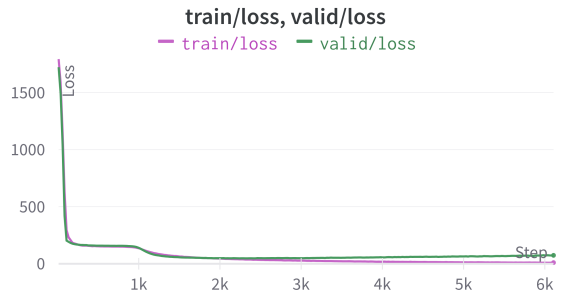

Fig. 6. Training and validation loss of the distant ASR.

To evaluate the performance of the integrated system Turnaround time and Response time are computed. Let, $t0 = 4\ secs$ is the time for which the speech module records the speech and generates DOA using SSL. $t1 = 0\ secs$ is the time at which the speech module sends the recorded speech to the ASR server. $t2 = t1 + k\ secs$ is the time at which the speech module receives the processed DOA from ASR server. $t3 = t2 + m\ secs$ is the time at which the mobile robot starts rotating. $t4 = t3 + n\ secs$ is the time at which the mobile robot completes the rotation.

$$t\_response = (t1 + t2 + t3)\ secs \tag{3}$$

$$t\_response_{avg} = (\sum_{i=1}^{n} t\_response(i))/n\ secs \tag{4}$$

$$t\_turnaround = (t1 + t2 + t3 + t4)\ secs \tag{5}$$

$$t\_turnaround_{avg} = (\sum_{i=1}^{n} t\_turnaround(i))/n\ secs \tag{6}$$

where, i is the number of trials performed.

Equations 4 and 6 give the average response time and average turnaround time, respectively. The following Table II shows the average $t\_turnaround$ time and average $t\_response$ time computed after 30 real-world successful trials and by varying the speed of the mobile robot. It can be seen from the table that Average Response Time is near 0.5 milliseconds; however the Turnaround Time is high. This is due to data transfer over the remote server for speech recognition.

TABLE II

REAL-TIME PERFORMANCE EVALUATION OF MOBILE ROBOT AND SPEECH MODULE INTEGRATION.

Sl. No.	Speed of the robot (rad/sec)	Average Response Time - ART (Sec)	Average Turnaround Time - ATT (Sec)
1	20	0.487	5.1377
2	30	0.5854	3.6422
3	40	0.6941	3.3692

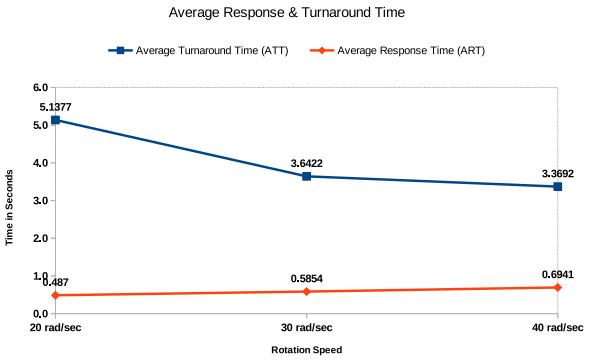

Fig. 7. Average Response Time and Average Turnaround Time of Turtlebot3

TABLE III

COMPARISON OF WER AND CER OF DISTANT ASR ON 4-CHANNEL BEAMFORMED SPEECH DATA AND SINGLE MICROPHONE SPEECH DATA.

Distance from the microphone	**4-channel beamformed output**		**Single microphone output**	
	WER ↓	CER ↓	WER ↓	CER ↓
1	28.70	16.95	36.30	31.21
3	40.26	27.47	47.95	37.81
5	53.86	43.52	63.92	60.02
Overall	40.38	28.85	48.91	43.03

D. Ablation Study

The proposed system consisted of beamformed signal from 4-channel circular array microphone. In this ablation study, only one microphone output is considered. The degradation of the performance could be seen in Table III.

In the real-world experiments of robot localization using SSL and distant ASR, sound reverberation in large and empty rooms needs to be improved to get the accurate DOA of speech. An area for improvement is the distance from which the speaker speaks with the robot. Up to 5 meters the distant ASR performs well. However, beyond 5 meters, due to the quality of the signal, the performance decreases substantially. Still, 5 meters can be considered a higher extent the speaker would command in an indoor setting.

V. CONCLUSIONS AND FUTURE WORK

This paper introduces a speech module designed to enhance the capabilities of mobile robots. The Turtlebot3 Waffle Pi mobile robot is mounted with 4-channel circular array microphone for the audio tasks. The speech module integrated to the ROS, consists of two components: a distant automatic speech recognizer and a sound source localizer. By making use of far-field speech data collected at varying distances till 5-meter range, the distant speech recognizer exhibits better performance with Word Error Rate of 40.38% and a Character Error Rate of 28.85%. With the help of multi-array microphone, the direction of arrival is calculated accurately even in the moderate reverberating environment. Complete setup works smoothly even on a resource efficient single-board computer. This work could be further extended by adding additional resources. Adding vision sensors such as camera and depth sensors will allow the mobile robot to navigate seamlessly.

ACKNOWLEDGMENT

We would like to express our gratitude towards the funding agencies Technology Innovation Hub on Autonomous Navigation and Data Acquisition Systems (TiHAN) Foundation at IIT Hyderabad, India and Ministry of Electronics and Information Technology (MeitY), Government of India.

REFERENCES

[1] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5610-5614, doi: 10.1109/ICASSP.2015.7179045.

[2] Xinyuan Qian, Zhengdong Wang, Jiadong Wang, Guohui Guan, Haizhou Li, "Audio-Visual Cross-Attention Network for Robotic Speaker Tracking", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.31, pp.550-562, 2023.

[3] P. P. Rao and A. R. Chowdhury, "Learning to Listen and Move: An Implementation of Audio-Aware Mobile Robot Navigation in Complex Indoor Environment," 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 2022, pp. 3699-3705, doi: 10.1109/ICRA46639.2022.9812193.

[4] G. Narang, K. Nakamura and K. Nakadai, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual SLAM," 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 2014, pp. 4021-4026, doi: 10.1109/SMC.2014.6974560.

[5] C. Gan, Y. Zhang, J. Wu, B. Gong and J. B. Tenenbaum, "Look, Listen, and Act: Towards Audio-Visual Embodied Navigation," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 9701-9707, doi: 10.1109/ICRA40945.2020.9197008.

[6] Y. Sasaki, Y. Tamai, S. Kagami and H. Mizoguchi, "2D sound source localization on a mobile robot with a concentric microphone array," 2005 IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, HI, USA, 2005, pp. 3528-3533 Vol. 4, doi: 10.1109/ICSMC.2005.1571694.

[7] X. Li, L. Girin, F. Badeig and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function", 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)., pp. 2819-2826, 2016.

[8] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments", IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 1, pp. 88-103, 2019.

[9] J. H. DiBiase, H. F. Silverman and M. S. Brandstein, "Robust localization in reverberant rooms" in Microphone arrays., Springer, pp. 157-180, 2001.

[10] J. -M. Valin, J. Rouat and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), Sendai, Japan, 2004, pp. 2123-2128 vol.3, doi: 10.1109/IROS.2004.1389723.

[11] Y. Tamai, Y. Sasaki, S. Kagami and H. Mizoguchi, "Three ring microphone array for 3D sound localization and separation for mobile robot audition," 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2005, pp. 4172-4177, doi: 10.1109/IROS.2005.1545095.

[12] J. -M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai and H. G. Okuno, "Robust Recognition of Simultaneous Speech by a Mobile Robot," in IEEE Transactions on Robotics, vol. 23, no. 4, pp. 742-752, Aug. 2007, doi: 10.1109/TRO.2007.900612.

[13] Q. V. Nguyen, F. Colas, E. Vincent and F. Charpillet, "Localizing an intermittent and moving sound source using a mobile robot," 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea (South), 2016, pp. 1986-1991, doi: 10.1109/IROS.2016.7759313.

[14] R. Gomez, K. Nakamura, T. Mizumoto and K. Nakadai, "Temporal smearing compensation in reverberant environment for speech-based human-robot interaction," 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 2015, pp. 3347-3353, doi: 10.1109/ICRA.2015.7139661.

[15] C. Evers and P. A. Naylor, "Acoustic slam", IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 26, no. 9, pp. 1484-1498, 2018.

[16] S. Michaud, S. Faucher, F. Grondin, J.-S. Lauzon, M. Labbé, D. Létourneau, et al., "3d localization of a sound source using mobile microphone arrays referenced by slam", 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10402-10407, 2020.

[17] T. Zhang, H. Zhang, X. Li, J. Chen, T. L. Lam and S. Vijayakumar, "AcousticFusion: Fusing Sound Source Localization to Visual SLAM in Dynamic Environments," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 6868-6875, doi: 10.1109/IROS51168.2021.9636585.

[18] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, Abhinav Valada, "Catch Me if You Hear Me: Audio-Visual Navigation in Complex Unmapped Environments With Moving Sounds", IEEE Robotics and Automation Letters, vol.8, no.2, pp.928-935, 2023.

[19] S. Majumder, Z. Al-Halah and K. Grauman, "Move2Hear: Active Audio-Visual Source Separation," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 275-285.

[20] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, Jiajun Wu, "Sonicverse: A Multisensory Simulation Platform for Embodied Household Agents that See and Hear", 2023 IEEE International Conference on Robotics and Automation (ICRA), pp.704-711, 2023.

[21] Caleb Rascon, Ivan Meza, "Localization of sound sources in robotics: A review, Robotics and Autonomous Systems", Volume 96, 2017, Pages 184-210, ISSN 0921-8890, https://doi.org/10.1016/j.robot.2017.07.011.

[22] A. Gupta, H. Singh Chadha, P. Shah, N. Chhimwal, A. Dhuriya, R. Gaur, V. Raghavan, "CLSRIL-23: Cross Lingual Speech Representations for Indic Languages", 2021 2107.07402 arXiv, https://doi.org/10.48550/arXiv.2107.07402.

[23] H. Singh Chadha, A. Gupta, P. Shah, N. Chhimwal, A. Dhuriya, R. Gaur, V. Raghavan, "Vakyansh: ASR Toolkit for Low Resource Indic languages", 2022 2203.16512, arXiv, https://doi.org/10.48550/arXiv.2203.16512.

[24] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli., "Wav2vec 2.0: a framework for self-supervised learning of speech representations" In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 1044, 12449–12460, 2020.