AMIGO (A SOCIAL ROBOT):DEVELOPMENT OF A ROBOT HEARING SYSTEM

A.M.N.C. Attanayake
Department of Electrical and
Telecommunication Engineering
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
nimashaca@gmail.com

W.G.R.U. Hansamali
Department of Electrical and
Telecommunication Engineering
South Eastern Universityof Sri Lanka
Oluvil, Sri Lanka
ruviniuthpala2@gmail.com

R. Hirshan
Department of Electrical and
Telecommunication Engineering
South Eastern Universityof Sri Lanka
Oluvil, Sri Lanka
ruviniuthpala2@gmail.com

M.A.L.A. Haleem
Department of Electrical and
Telecommunication Engineering
South Eastern Universityof Sri Lanka
Oluvil, Sri Lanka
mala haleem@seu.ac.lk _

M.N.A. Hinas
Department of Computer Science
and Engineering
South Eastern Universityof Sri Lanka
Oluvil, Sri Lanka
ajmalhinas@seu.ac.lk

***Abstract*—currently, humanoid robotic applications are increasing. Humanoid robots with the capability of listening to the human voice are useful in a number of applications. In many studies, hearing systems are not designed to work in real time. However, when a humanoid robot is used for a specific purpose, a triggering method is required. If not, it will be difficult to put it into practice. When using a humanoid robot in an environment with frequent human voices, it is important to identify which sound the robot should be responding to. This paper proposesa hearing system for a humanoid robot that has the ability to turn towards the speaker who called its name "Amigo". Thesystem uses the Angle of Arrival technique (also called Direction of arrival) and speech recognition. The Angle of Arrival (AOA)is performed by using the Time Difference of Arrival (TDOA) technique. Google Speech Recognition Application Programming Interface (API) is used for speech recognition. Cross-correlation of two acoustic signals is used to measure the TDOA. In order to test the system, a robotic head was developed using 3D printed components, a Raspberry Pi computer, and a stepper motor. The Raspberry Pi computer is used for audio signal processing and motor control. Two I2S microphones were used as audio devices. Experimental results show 82.54% of accuracy in the indoor environment and 85.94% accuracy in the outdoor environment.**

***Index Terms*—Angle of arrival, Cross-correlation, Speech recognition, Time Difference of Arrival**

I. INTRODUCTION

Humanoid robots have attracted a significant amount of interest in recent years. The most suitable way to relieve hu- mans from tedious routine tasks is to develop humanoid robots with real-time sensing features. But the design of that type of humanoid robot is not an easy task. Often a separate study of different body parts takes place. When using humanoid robots for applications in education, reception, and patient caring, they should be able to listen to the human. It improves human-robot interaction to a better level. Hearing is defined as a process or function of perceiving sound while listening is defined as paying attention to sound [1]. Developing an auditory system for robotic applications with the capability of listening to the human voice is quite challenging. The main objective of this research was to design and develop a robotic hearing system consisting of listening capability with low cost and high computational efficiency. The developed robot head is capable of estimating the direction of the sound source. Speech recognition was added to the system and it gave the robot more vitality. When a person calls the robot saying "Hello Amigo", the sentence is recognized by our robot using Google Speech Recognition API. Then the system is triggered and the angle from which the acoustic signal was received is measured. Since this robot is only triggered for its name, it can be used for specific purposes in an environment with interfering conversations. In addition, it was designed to answer some questions to make it attractive to the user and to illustrate the usage of this robot. The developed system was tested in indoor and outdoor environment. It works with high precision for real-time use. This paper discusses the hearing part of the social robot "Amigo" and a picture of the robot is shown in fig. 1. All the scripts were written in the Python programming language.

The remainder of the paper is organized as follows. A summary of related published work is described in section II. Section III introduces the methodology of the system. Section IV presents results with experimental procedure and the resultsare analyzed in section V. Section VI concludes the paper.

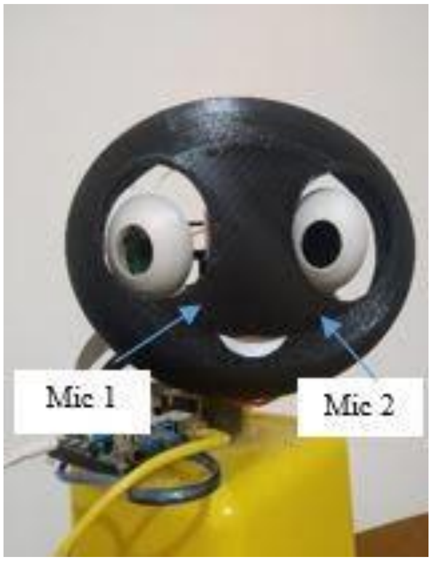

Fig. 1. Overview of the robot.

II. RELATED STUDIES

Several different approaches such as Time difference of arrival (TDOA), Time of arrival (TOA) and received signal strength (RSS) were used to measure the features of acoustic signals in artificial hearing systems [2]. In several previous studies, sound source tracking was achieved by TDOA and TOA. In [3], the location of a single sound source was found by using two separate microphone pairs with USB sound card and the system could follow the instrumental music in an iterative manner. Furthermore, it is possible to use microphone array which consists of three microphones to determine the direction towards the acoustic source [4].However, rather than using a lot of microphones in an array, squared microphone array which consists of four microphones was shown to be more appropriate for applications related to humanoid robots [5], [6]. An exciting method found in literature was the monaural localization approach which uses machine learning techniques, a single microphone and a set of small partial enclosures for the microphone that serves as "artificial pinna" [7]. HEARBO is an advanced robot that has ability to localize a sound source, separate sound, and recognize of the human speech [8]. Robot HEARBO first captures all the sounds and recognizes them. Then determines the location they're coming from. Finally, it gives attention to each source one by one. Ava is a social robot which is able to turn towards the speaker in noisy environments [9]. This robot is able to differentiate speech from non-speech segments. Ava works in 10 dB noisy environments with a precision of +/-5°. A comparison of different types of limitations in these studies is listed in Table 1.

III. METHODOLOGY

A. Assumptions

When implementing and testing the hearing system, following assumptions were made.

- Sound waves detected by two microphones are parallelto each other.
- Change in velocity of sound in air due to variations inpressure, temperature and humidity is negligible.
- The effect of reverberation is negligible.

B. Background

In signal processing, cross-correlation is used to measure the similarity between two signals. The point of maximum similarity is identified by the peak point of cross correlation signal. The cross-correlation function is denoted in (1), where x and y are two real continuous functions of time. The limits of the integral must be changed according to the length of the signal

$$\emptyset_{xy} = \int_{-\infty}^{+\infty} x(\tau - t)y(\tau)d\tau$$

The basic task in AOA estimation is determining the angle of a received acoustic signal. When the microphones are spatially separated, the sound signals arrive at them at differenttimes. The AOA of the signal can be determined from mea- sured time delays, if the microphone array geometry is known.The time taken to get maximum similarity is considered as TDOA. For this robot's hearing system, two microphoneswere used. According to Fig.2, the angle θ must be foundto determine the direction of the sound source. According to the parallel wave assumption, signal 1 takes Δt time morethan signal 2, to reach the microphone. The cross-correlation method was used to find this Δt time. Finally, the angle θ was found by using (2).

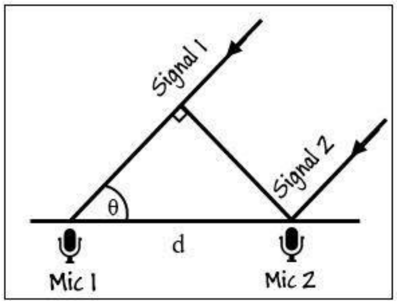

Fig. 2. Diagram showing difference in path lengths of arrival

$$\theta = \arccos(\Delta t * V/d), \qquad (2)$$

where,
θ = Angle of arrival
d = Distance between two microphones
Δt = Time difference of arrival
V = Velocity of Sound in air.

Speech recognition involves translation of speech froma verbal format to a text format. There are many speech recognition applications, but modern speech recognition uses deep neural network algorithms. However, several speech

TABLE I. COMPARISON OF LIMITATIONS IN RELATED STUDIES

Study	Computing device	Features	Limitations
HEARBO robot [8]	N/A	Sound localization, separation, and recognition.	N/A
Ava (A Social Robot): Design and Performance of a Robotic Hearing Apparatus [9]	PC	Sound localization, and recognition.	Computational power is high.
Robust Speech Direction Detection for Low Cost Robotics Applications [3]	Raspberry Pi	Assumed that the music used would be of a much higher frequency than 300 Hz. The robot move towards the instrumental music or singing voice.	Cannot distinguish between two sounds which is in same or most similar space in the frequency spectrum.
Acoustic Source Localization and and Navigation Robot [4]	Arduino	Estimate the direction of sound sources	Working for any sound signals that greater than or lower than the threshold value.
Acoustic Source Localization [10]	PC	Localize an acoustic source within a frequency band from 100Hz to 4 KHz	Working for all voice signals between 100Hz to 4 KHz.

recognition services are available for online usage through an API. The audio of a speech recording is broken down into individual words and properties like amplitude and frequency are taken into account to recognize those words separately. Finally, those sounds are converted into text format. Simply, this is how Speech recognition systems work. In this project, we have used Google speech recognition API with Python SpeechRecognition package to recognize the sentence “Hello Amigo” [11]. As the computing device, a Raspberry Pi 4model B (4GB RAM) computer was used. Raspberry Pi isa low cost, small form factor, single board computer. It does not have an audio input. When using the Raspberry Pi for real time audio signal processing, it is necessary to use USB soundcard or I2S via IO pins. Two I2s microphones (INMP441) werechosen for recording sound [12]. A simple 3D printed face was connected to the stepper motor and the motor smoothly rotates towards the angle θ that is measured by the system.For that, firstly, the system measures the TDOA using cross- correlation and then calculates the angle. Then the robot turns the neck to the calculated angle smoothly. Considering the normal behavior of a human neck, the robot neck was designedto be able to rotate in a range of 120 degrees maximum [13]. Further, the robot head includes features to answer questions from the human user. In order to achieve this, pre-defined questions were included in a Python script in text formatand answers to the corresponding questions were saved inaudio format. If the user asked one question from this listof questions, the system plays the answer to that question.

C. Algorithm

There were two threads used in this system; to always keep twenty second stereo audio recording as a queue andto simultaneously wait for acoustic signal. When any acoustic signal with high amplitude is detected, a set of timing data is obtained and the detected acoustic signal is checked to see whether it is the trigger word or not. If it is the trigger word, the recorded audio portion is retrieved according to timing data. Then the cross correlation between the stereo componentsis taken and TDOA is estimated. After finding the angle by this process, the stepper motor is rotated accordingly. Motoris reset to the initial position, when either the word “Bye” is recognized or conversation does not continue within a minute. The flow diagram of the entire hearing system is shown in fig. 3.

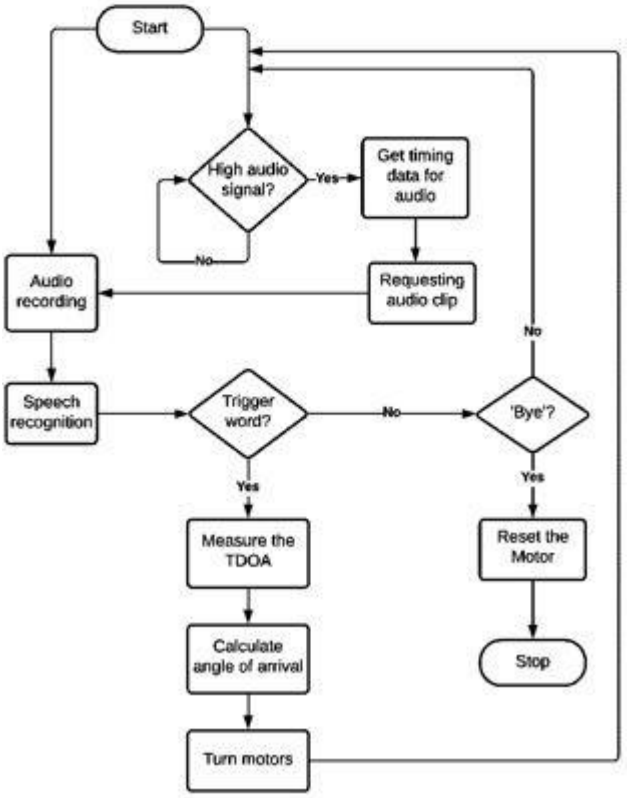

Fig. 3. Flow Diagram of the System

IV. RESULTS

We tested the AOA estimation performance of the robot for seven different orientations of the speaker (0°, -20°, +20°,-40°, +40°, -60°, +60°). The distance between two microphones and the distance from microphones to sound source were changed for each of the above angles. The microphone pair separation was set at 3cm, 6cm, 12cm, 18cm, and 24cm, whereas the distances from microphone pair to sound source was set to

0.5m, 1m, 1.5m, 2m, and 2.5m. The testing procedure was to fix the distance between microphones and to change the position of the speaker with respect to the robot as denoted by black dots in Fig. 6. The success rate was evaluated by repeating the same trigger word 20 times for each position of the speaker. The above procedure was repeated for different microphone pair separations. A trial was declared a success if the AOA estimated by the system were within +/- 5°of the actual direction. Success rate was evaluated as a percentage by counting the number of successes out of 20 trials. Fig. 4 and Fig. 5 show the "heatmap" of success rate as percentages. The mean of success rates over all positional setting wastaken as the overall accuracy. With this series of experiments, the overall accuracy of AOA estimation has been evaluatedto be 82.54% in indoor environment and 85.94% in outdoor environment.

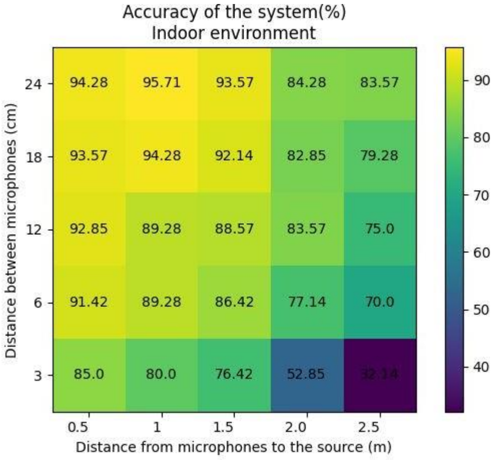

Fig. 4. Accuracy in the indoor environment

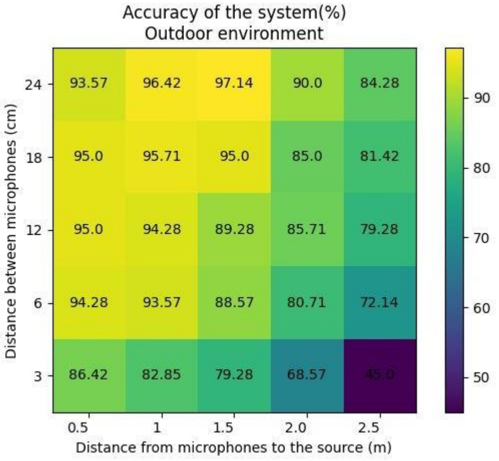

Fig. 5. Accuracy in the outdoor environment

V. Discussion

According to the results, we can deduce that system accuracy is increased when distance between two microphones is increased and distance between microphones to the sound source is decreased. When the two microphones are too close, the time difference is much less and the accuracy will be less. But, due to our assumption of parallel sound waves, it is not advisable to keep them too far away. The distance between pair of microphones and sound source should always be keptat a sufficiently larger value compared to the distance betweenthe two microphones. When selecting the optimal distance between two microphones, accuracy and the type of application need to be considered. When the microphones are fixed in the robot head, distance between two microphones must be matched with dimensions of the robot head. The reason for the decrease in accuracy in the indoor environment than the outdoor environment is that the system was subjected to reverberation. Other than that, when the trigger word was long, the accuracy of cross correlation result was better than that of a shorter trigger word. This was the reason to use the trigger word as "Hello Amigo" instead of only using "Amigo". Since Google speech recognition was used for speech recognition, good network connection was always expected for proper functioning of the system. The robot's response time also depends on the speed of the internet connection and it was 2 seconds at 10Mbps for the trigger word ”Hello Amigo”. If the user didn't speak clearly, the robot may misinterpret the words.

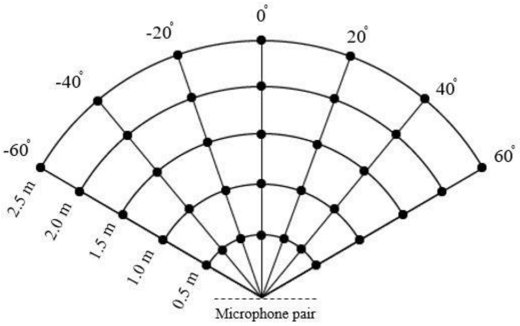

Fig. 6. The positions where the sound source was placed when the microphone pair were at the fixed position (This process was repeated foreach microphone pair separations)

VI. Conclusion

In signal processing, the direction of arrival (DOA) denotes the direction in which waves are coming to the sensors, and there are several effective use cases of finding the DOA of acoustic signals, but the problems arise with the cost and the computational power. In many recent studies, there have been instances where computational power, practicality, and cost of applications have not been taken into account. This paper presents a computationally efficient and low-cost method to find the DOA and investigates the effect of applying the method in a humanoid robotic head to improve the hearing ability to an interactive level. The current system still has some limitations. The system is susceptible to reverberations in an indoor environment and the operation of the neck is limited to the horizontal plane. Reverberation filters could be used to improve the accuracy in the indoor environment. The work andresults presented in this paper are a part of a social robot whichis called “Amigo”. Future work of this study will consist ofan intelligence system, natural-looking eyes, and more natural neck behavior.

REFERENCES

[1] S. Lindberg, “What’s the difference between hearing and listening?” Available at https://www.healthline.com/health/hearing-vs-listening (2021/02/04).

[2] X. Li, Z. D. Denga, L. T. Rauchenstein, and T. J. Carlson, “Contributed review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements,” *Review of ScientificInstruments*, p. 13, 2016.

[3] S. Ramnath and G. Schuller, “Robust speech direction detection for low cost robotics applications,” Kos island, Greece, 2017.

[4] A. R. J. T, V. R. G, and S. K, “Acoustic source localization and navigation robot,” *in International Journal of Engineering and AdvancedTechnology (IJEAT)*, 2019.

[5] A. K. Pandey and R. Gelin, “A mass-produced sociable humanoid robot:Pepper: The first machine of its kind,” *IEEE Robotics and Automation Magazine*, p. 10, 2018.

[6] X. Lv and M. Z. X. Liu, “A sound source tracking system based on robothearing and vision,” *International Conference on Computer Science andSoftware Engineering*, 2008.

[7] A. Saxena and A. Y. Ng, “Learning sound location from a single mi- crophone,” *IEEE International Conference on Robotics and Automation*,2009.

A. Lim, “Hearbo robot has superhearing,” *IEEE Spectrum*, 2012.

[8] E. Saffari, A. Meghdari, B. Vazirnezhad, and M. Alemi, “Ava (a social robot): Design and performance,” *7th International Conference on SocialRobotics, Paris, France*, 2015.

[9] S. Paulose, E. Sebastian, and D. B. Paul, “Acoustic source localization,” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2013.

[10] “The ultimate guide to speech recognition with python,” Avail- able at https://realpython.com/python-speech-recognition/#putting-it-all- together-a-guess-the-word-game, (2020/12/29).

[11] “I2s mems microphone for raspberry pi (inmp441),” Available at https://makersportal.com/shop/i2s-mems-microphone-for-raspberry-pi- inmp441, (2020/12/15).

[12] S. LaValle, “The physiology of human vision,” *Virtual Reality, Cam-bridge University Press*, p. 426, 2019.