Design of a Microphone Array for Rollin' Justin

Marco Sewtz Tim Bodenmüller Rudolph Triebel

I. Introduction

For a humanoid robot operating in populated environments it is a key competence to interact with humans naturally and intuitively. Therefore, research in human-robot interaction explores the interpretation of visual, auditive and even tactile sensing modalities. One important aspect here is to recognize robustly the intention of the human interacting with the robot. To achieve this, robot audition is a suitable modality, as it allows for detecting and tracking speakers from arbitrary positions around the robot and also from distant places. In this paper we present the design of a microphone array for the head of our humanoid robot Rollin' Justin that allows us to localize and track sound sources within a certain distance to the robot. Until now, high-level interfacing to the robot was only possible using a tablet or by verbal communication via a headset using speech recognition. However, neither method allows to localize the operator. With the microphone array in the head we can do sound source localization and then re-position the head sensors towards the speaker, enabling advanced interaction possibilities.

In recent years, the research in the field of service robotics focuses on human-centered interfaces. This includes easy-to-use as well as easy-to-understand systems like robots with audio input [1]. Several systems have been developed using microphone arrays to extract speech [2], [3]. More complex systems use multiple techniques for processing including sound source localization, feature extraction and speech-to-text engines [4].

In the following we present the design of a microphone array for Rollin' Justin for sound source localization and speech recognition. First we describe the considered household scenario and present an overview of the proposed processing system. We than discuss our design considerations regarding the microphone array as well as possible sound source localization and speech processing. We conclude with a description of system integration.

II. Design

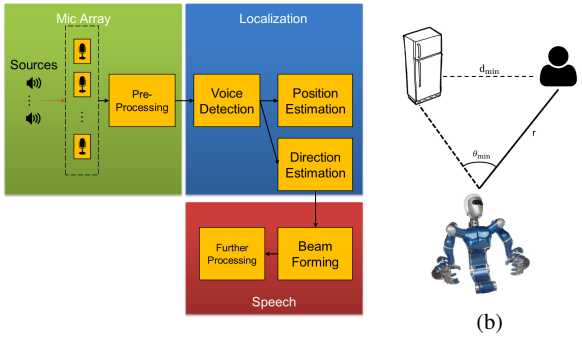
For the design of our microphone array we consider the following scenario: our robot is located in a typical indoor environment, for instance an apartment (Figure 1). This implies that we have multiple sources of noise, e.g. a fridge, and reverberations. In addition, we suppose that there is only one person at a time speaking to the robot, called the operator. The expected distance r between the robot and the operator is between $1\,\mathrm{m}$ and $4\,\mathrm{m}$. We assume that, from the robot's point of view the operator and any other sound source have a minimum tangential distance of at least $d_{\min} = 1\,\mathrm{m}$. This results in a minimum angular distance $\theta_{\min} = \arctan(d_{\min}/r_{\max}) \approx \pm 14^\circ$ separating the speech from any other sound source. Figure 2b illustrates this scenario.

A. System Overview

We plan our system as illustrated in Figure 2a. The sound from the sound sources is received and sampled by a microphone array. In a preprocessing step we remove ambient noise and the robot's ego-noise. Then, we localize the sound source by calculating the direction and distance of the source. This is used by subsequent beam forming to improve the signal for speech recognition.

All authors are with: Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany. Email: marco.sewtz@dlr.de, tim.bodenmueller@dlr.de, rudolph.triebel@dlr.de

Figure 1: Lab environment imitating a typical small apartment



(a) (b)

Figure 2: (left) System overview and (right) illustration of assumed audio scenario.

B. Microphone Array

For the design of the microphone array, we consider the expected sound signals, the processing requirements, as well as geometrical limitations. In general, we follow the approach of a broadband microphone sub-array [5]. Hence, we divide the frequency spectrum into three sub-bands ($< 1\,\mathrm{kHz}$, 1-2 kHz, $> 2\,\mathrm{kHz}$), each handled by a specific sub-array. For lower frequencies, we need large distances between microphones to capture longer delays. For higher frequencies, we have to use a smaller spacing. However, the range of feasible distances between the microphones is small due to limited space within the robot's face mask. Also, they have to be placed on a single plane because of limited space and appearance constraints. In total we use eight microphones that are arranged as shown in Figure 4b and grouped into sub-arrays as shown in Figure 4a. The outer microphones have a distance of $146\,\mathrm{mm}$. To improve the speech quality [6], we sample with at least $16\,\mathrm{kHz}$ instead of the usual $8\,\mathrm{kHz}$. The signals are bandpass filtered, amplified and summed up to the complete signal. We optimized the microphone positions by a free-air simulation. The resulting combined directivity pattern is shown in Figure 3. The main lobe of the array is focused around our defined 14° corridor at -3 dB, and sources at larger angles are suppressed. In the lower frequency range, the lobe is somewhat less focused due to a limited maximum inter-signal delay.

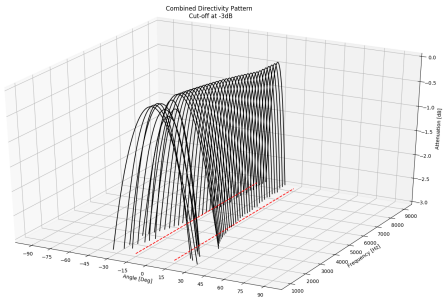
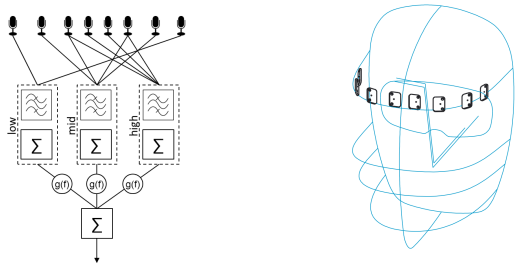


Figure 3: Directivity pattern for the combined array approach. The figure shows the pattern truncated at -3 dB. The dashed red line illustrates a $\pm 14^{\circ}$ corridor.



(a) Sub-Array concept

(b) CAD drawing of the proposed microphone positions

Figure 4: (left) Illustration of the sub-array approach, (right) CAD drawing of the microphones on the robot's face.

C. Sound Source Localization

For localization, we only use signal snippets that contain speech, which we identify by means of the Long-Term Speech Divergence [7]. For an accurate and robust localization from these snippets we use a modified version of the MUSIC algorithm [8].

The main principle of the estimation process relies on the delay between the received signals. In case of a linear array, the delay is proportional to the spacing between the microphones. For our 2D array design, the delay Δt_i between microphone i and a reference point assuming a signal from direction θ is given by the projection onto the direction vector of the arriving sound wave, i.e.

$$\Delta t_i = (\mathbf{p}_i^{\top} \mathbf{e}_\theta)/c_0 \ ,$$

where $\mathbf{p}_i$ is the position of the i-th microphone with respect to the reference point, $\mathbf{e}_\theta$ the direction unit vector of the sound wave and $c_0 \approx 343\,\mathrm{m/s}$ the speed of sound. Figure 5 illustrates this relation.

D. Further Processing

With the direction of the sound source, we steer our system towards the source using a delay-and-sum beamformer in combination with our sub-array approach. We expect a high increase in signal-to-noise ratio as well as noise suppression from this technique. The retrieved speech signal will be used to extract commands from the operator. We will use already available offline speech processing engines such as CMU Sphinx [9], which has already been used with our headset. We also plan to adapt voice assistant paradigms as in the NAOMI Project [10].

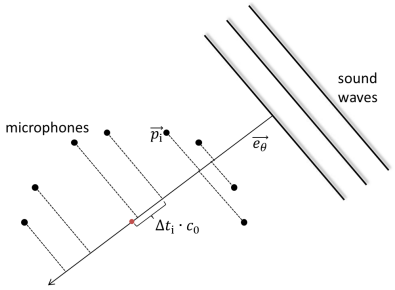


Figure 5: Illustration of the delay calculation for an arriving sound wave. The intersections of the dashed lines show the projection onto the direction vector $\vec{e_\theta}$. The red dot represents the reference point.

III. Integration

For the array we chose SPH0645LM4H-8 MEMS microphones with I^2S support[1]. They will be sampled simultaneously by the native I^2S ports of an Nvidia Jetson TX2 board mounted onto an Auvidea J140 carrier. The whole system will be integrated physically into the head of the robot, and the software will be part of our "Links and Nodes" framework.

IV. Conclusion

We presented the design of a microphone array for sound source localization and speech processing on the humanoid robot Rollin' Justin. We introduced our overall system architecture, gave an overview of the subsequent sound processing, and described design considerations, implementation details and preliminary evaluations. In a next step, we will finalize the implementation on the robot and execute several experiments in realistic scenarios.

V. Acknowledgment

We want to thank Werner Friedl for his contributions to the mechanical design and implementation.

References

[1] A. Di Nuovo, F. Broz, N. Wang, T. Belpaeme, A. Cangelosi, R. Jones, R. Esposito, F. Cavallo, and P. Dario, "The multi-modal interface of Robot-Era multi-robot services tailored for the elderly," *Intelligent Service Robotics*, 2018.

[2] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *IROS*, 2004.

[3] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *IROS*, 2016.

[4] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system "HARK" — open source software for listening to three simultaneous speakers," *Advanced Robotics*, 2010.

[5] I. A. McCowan, "Robust speech recognition using microphone arrays," Ph.D. dissertation, Queensland University of Technology, 2001.

[6] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, 2014.

[7] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, 2004.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, 1986.

[9] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The CMU Sphinx-4 speech recognition system," in *ICASSP*, 2003.

[10] The Naomi Community and Project Naomi. [Online]. Available: https://projectnaomi.com/

[1] https://www.knowles.com/