2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

# Robust MUSIC-Based Sound Source Localization in Reverberant and Echoic Environments

Marco Sewtz[1] Tim Bodenmüller[1] Rudolph Triebel[1,2]

***Abstract*— Intuitive human robot interfaces like speech or gesture recognition are essential for gaining acceptance for robots in daily life. However, such interaction requires that the robot detects the human's intention to interact, tracks his position and keeps its sensor systems in an optimal configuration. Audio is a suitable modality for such task as it allows for detecting a speaker in arbitrary positions around the robot. In this paper, we present a novel approach for localization of sound sources by analyzing the frequency spectrum of the received signal and applying a motion model to the estimation process. We use an improved version of the Generalized Singular Value Decomposition (GSVD) based MUltiple SIgnal Classification (MUSIC) algorithm as a direction of arrival (DoA) estimator. Further, we introduce a motion model to enable robust localization in reverberant and echoic environments.**

**We evaluate the system under real conditions in an experimental setup. Our experiments show that our approach outperforms current state-of-the-art algorithm and demonstrate the robustness against the previously mentioned disruptive factors.**

## I. Introduction

The ability of mobile robots to interact with people in an intuitive and maybe anthropomorphic manner is a key to the acceptance of robots in human-dominated environments. Human-robot-interaction (HRI) can be visual (e.g. gestures), tactile (e.g. guiding) as well as auditive (e.g. instructing). However, all modalities require that the robot recognizes the intention of a human to interact. Visual systems can only recognize intention in the sensor's field of view, which is usually limited. Tactile systems require that the human is nearby. Robot audition, however, allows for detecting and tracking a speaker from arbitrary positions around the robot and also from distant places. Figure 1 illustrates a typical situation. The human on the sofa wants to interact with the robot, but the latter is currently performing another task, thus, positioning its visual sensor in the opposite direction. Moreover, audio also allows for gaining information about the environment or to separate between different speakers. The information about the speaker's position can also be used to enhance the audio input signal, e.g. to improve speech processing as well as getting more information about the position of humans in the scenario.

In this work we present a novel approach for localization of speakers by use of a microphone array. First we detect speech phases in the audio stream using a voice activity detector. During the detection we calculate a score for analyzing the frequency spectrum. We introduce a frequency selection based on this score to enhance the MUSIC estimation and reduce the processing time. For estimation we use the established SEVD-MUSIC [1] algorithm. Further on we propose a motion model to check the calculated Direction of Arrival (DoA) of the received signal. We can show that this enhances robustness against reverberation and echo. Therefore, we present result of realistic experiments that verify our claims.

Fig. 1: Illustration of the interaction recognition problem: The robot is turned away from the operator. While the vision system might not recognize him, the audio input will do so.

## II. Related Work

In recent years, research has been done to imitate the binaural audio localization of animals and humans [2]–[5]. Using both the interaural phase difference (IPD) and the interaural intensity difference (IID). These techniques take into account the head-related transfer function [6], [7] as well as the reverberant properties of the environment to achieve accurate results. Incorporation of a particle filter approach to be used on binaural measurements improves the estimation of sound sources as well [8]. Nonetheless these systems need a demanding hardware setup and calibration.

Other approaches use an array of microphones to overcome the hardware requirements and to estimate the direction of arrival (DoA) of a signal [9], [10]. It is possible to calculate the most probable DoA by estimating the time delay between the signals received by each microphone.

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Oberpfaffenhofen, Germany.
[2]Dep. of Computer Science, Technical Univ. of Munich, Germany
marco.sewtz@dlr.de tim.bodenmueller@dlr.de
rudolph.triebel@dlr.de

rmc.dlr.de/rm/de/staff/marco.sewtz/SSL

Combining these methods with delay and sum beam forming (DSBF) as well as random sample consensus (RANSAC), more than one sound source can be localized [11]. However, these approaches have problems with low signal-to-noise-ratios (SNR) input signals, changing acoustic conditions and varying speakers. Different approaches using neural networks have been studied to tackle these problems. Nevertheless, they need training dedicated to the specific speaker or require very large amounts of data for generalizing [12]–[16]. Furthermore, Sasaki et al. present an approach incorporating a hypothesis tracking system which exploits the physical constraints of a dynamic moving object [17].

More recently, subspace approaches like Multiple Signal Classification (MUSIC) [18] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [19] have received more interest. They overcome the resolution limit constrained by the sampling rate and are more robust to signal noise but they are computational costly [20]–[23].

There have been several extensions for MUSIC, e.g. using singular value decomposition [24] to reduce the computational complexity while enhancing robustness against noise. Incremental versions are introduced to reach real-time performance while enhancing robustness against noise [25], [26]. Enhancements to further reduce the computational costs in the representation space is done in [27], [28].

However, even recent sound source localization systems face problems when detecting humans in indoor scenarios under non-optimal acoustic conditions.

First, the estimation of speech is challenging. The receiving sound event consists of several words, each composed of vowels and consonants with different frequencies and durations. It is therefore hard to implement a filter a-priori. Active filter system which adapts to the current information in real-time as proposed by Hoshiba et al. [29] tackle this problem. However, human speech consists of frequencies distributed on a wide spectrum. Using only a bandpass which narrows the calculations to small portions of the complete spectrum neglects additional information encoded in the signal or may even led to falsely estimations when the filter adapts to a noise source.

Secondly, indoor scenes often face the problem of having a high reverberation time and shadow sources created by echo. The first phenomena is the superposition of several reflections of the same signal which results in a "fading-out" effect and lower the SNR. The latter one is the reflection of the full signal at a surface and the system perceives an additional source at the location of the reflecting obstacle.

In this work we propose a novel framework based on the generalized singular value decomposition approach to reduce the complexity for estimating the DoA for localizing speakers. In addition we focus on raising the robustness in reverberant and echoic environments by exploiting intermediate steps of a noise evaluation process and validation based on a motion model. We aim to enable proven state-of-the-art methods for indoor scenarios under real-time constraints.

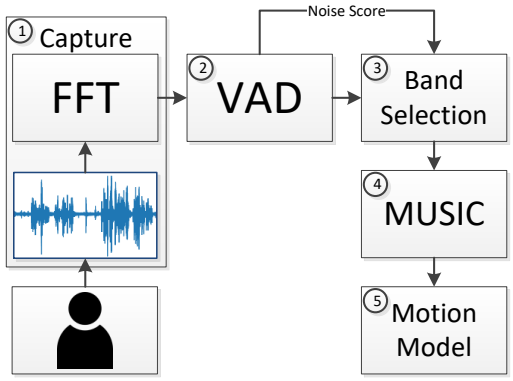


Fig. 2: System overview: ① Voice capture and transform into frequency domain. ② Classification of input as speech or noise phase. ③ Selection of appropriate frequencies. ④ DoA estiamtion with MUSIC. ⑤ Verification of DoA by a motion model.

## III. Sound Source Localization

In order to tackle the challenges of indoor environments, as discussed in the previous section, we propose a sound source localization framework called **Motion Model Enhanced MUSIC** (MME-MUSIC). Our system is based on the SEVD-MUSIC [1] approach. We enhance the estimation process by active selection of significant frequencies during Voice Activity Detection (VAD), as well as post-filtering the estimates by application of a motion model. A flow chart of our processing pipeline is given in Figure 2.

### A. Voice Activity Detection and Frequency Selection

We split the incoming audio recordings into smaller and overlapping frames and transform them into the frequency domain using the fast fourier transform (FFT). Afterwards, the frames are classified into the categories "speech" or "noise". We implement the Longterm Speech Divergence (LTSD) approach of Ramírez *et al.* [30], which assumes that the spectrum of noise differs significantly from frames containing speech. Yet, short time sound events like clapping or door closing are suppressed. For classification, the divergence of each frequency bin compared to a noise spectrum is computed, which we denote the **noise score** $\nu(k)$

$$\nu(k) = \frac{\overline{\mathrm{LTSE}}_\tau(k)^2}{\boldsymbol{X}_\Sigma(k)^2} \quad , \tag{1}$$

where $\overline{\mathrm{LTSE}}_\tau(k)$ is the average maximal amplitude of frequency band $k$ in a frame neighborhood $\tau$, and $\boldsymbol{X}_\Sigma$ a reference noise spectrum. The complete derivation can be found in [30]. Intuitively, a higher noise score means that the frequency bin differs more from the noise reference. If a frame is classified as "speech", then the noise score is used to analyze the frequency spectrum.

As mentioned above, considering the complete signal spectrum is not practical. However, a simple bandpass filter approach as in Hoshiba *et al.* [29] omits a lot of useful

information in the case of human speech. Therefore, we use the noise score $\nu$ to extract the $m$ bands with the highest score. This removes frequencies from the computation that do not contribute to the source signal. We show the selected bins from each algorithm in Figure 4. These bins are then fed into the SEVD-MUSIC estimator.

### B. SEVD-MUSIC

First, we derive the details to estimate the Direction of Arrival (DoA) for acoustic signals. We model our sound source as a point that emits a sinusoidal wave with center frequency $f_k$ and corresponding time-dependent amplitude $\lambda_k(t)$, where $k$ is the index of one out of $K$ frequency bands. Using the complex frequency notation we have

$$s(t) = \lambda_k(t) e^{j2\pi f_k t} = \lambda_k(t) e^{j\omega_k t} \ . \tag{2}$$

We consider a sensor array that consists of $N$ microphones, thus we obtain the system equation

$$\begin{bmatrix} 1 \\ e^{-jw_k\Delta_1} \\ \vdots \\ e^{-jw_k\Delta_{N-1}} \end{bmatrix} s(t) =: \boldsymbol{a}_k s(t) \ , \tag{3}$$

where $\Delta_n$ is the relative propagation delay with respect to the $n$th reference microphone. For a one-dimensional linear microphone array and under the assumption of planar waves, the delay is calculated as

$$\Delta_n = \frac{d_n \sin(\theta)}{c_0} \ , \tag{4}$$

where $d_n$ is the sensor's distance to the reference, $\theta$ the direction of arrival and $c_0$ the speed of sound, i.e. approximately $334\,m/s$ at room temperature. The vector $\boldsymbol{a}_k \in \mathbb{C}^{\mathrm{N}}$ in Equation (3) is denoted the **steering vector** for the frequency $f_k$. To obtain the complete **signal vector** we extend the system equation to

$$\boldsymbol{x}(t) = \boldsymbol{a}_k s(t) + \boldsymbol{n}(t) \ , \tag{5}$$

where $\boldsymbol{n}(t)$ is additional uncorrelated system noise.

When a new signal is received, we split it into smaller frames of fixed length and transform them into the frequency domain. Then, we compute the correlation matrix $\boldsymbol{R} \in \mathbb{C}^{N\times N}$ using

$$\boldsymbol{R} = \overline{\boldsymbol{X}(k)\boldsymbol{X}^{\mathrm{H}}(k)} \ , \tag{6}$$

where $\boldsymbol{X}(k) \in \mathbb{C}^{N\times F}$ contains the transformed Fourier coefficients of band $k$ for all $F$ frames and $N$ microphones. Here, $\boldsymbol{X}^{\mathrm{H}}$ is the Hermitian of $\boldsymbol{X}$. Using Singular Value Decomposition (SVD) on $\boldsymbol{R}$ to separate the contained subspaces, we get

$$\mathrm{SVD}\,(\boldsymbol{R}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{7}$$

$$\begin{aligned} \boldsymbol{U} &= [\boldsymbol{u}_0 \quad \boldsymbol{u}_1 \cdots \boldsymbol{u}_{N-1}] \\ &= [\boldsymbol{U}_{\mathrm{S}} \quad \boldsymbol{U}_{\Sigma}] \ , \end{aligned} \tag{8}$$

where $\boldsymbol{U}_{\mathrm{S}}$ is the **signal space** and $\boldsymbol{U}_{\Sigma}$ the **noise space**. As the system noise is uncorrelated it is present in all subspaces. The previously defined Steering Vector $\boldsymbol{a}_k$ is a property of a receiving signal and therefore defined in the signal space. This implies

$$\boldsymbol{a}_k \in \boldsymbol{U}_{\mathrm{S}} \ , \tag{9}$$

$$\Rightarrow \boldsymbol{a}_k \perp \boldsymbol{U}_{\Sigma} \ . \tag{10}$$

Hence, the inner product (denoted as $\langle\cdot,\cdot\rangle$) of the steering vector and the noise space is zero.

Natural sound events, especially the human speech, are composed of several frequencies. To take this into account we consider the complete frequency spectrum and combine it into a single representation. A common approach for that is the broadband pseudospectrum, which is defined over all frequency bands $K$ as

$$P(\theta) = \sum_{k=1}^{K} \frac{1}{\langle \boldsymbol{a}_k(\theta), \boldsymbol{U}_{\Sigma}\rangle^2} \ . \tag{11}$$

The DoA is found as the maximum of the estimator's response, i.e.

$$\tilde{\theta} = \operatorname{argmax} \ P(\theta) \ . \tag{12}$$

### C. Motion Model

We check the plausibility of the received angle by evaluating it with a motion model. To do this, we assume for time span $t_m$ that the source moves with mean angular velocity $\bar{\omega}$, i.e.

$$\bar{\omega}(t_m) = \overline{\left(\frac{\Delta\theta}{\Delta t}\right)} \approx \frac{1}{M} \sum_{n\in\mathcal{N}(t_m)} \frac{\tilde{\theta}_n - \tilde{\theta}_{n-1}}{t_n - t_{n-1}} \ , \tag{13}$$

where $\mathcal{N}(t_m)$ is the index set of all $M$ angular measurements $\tilde{\theta}_n$ within the time span $t_m$. A subsequent measurement $\tilde{\theta}_{m+1}$ is considered as valid, if

$$\left|\tilde{\theta}_{m+1} - \bar{\omega}(t_m)\right| < \theta_{tol}, \tag{14}$$

with the constant motion tolerance $\theta_{tol}$.

When receiving a new DoA from the previous steps we gather all estimations within the time span $t_m$. If at least two valid points are found we use our motion model to verify the new one. Otherwise we use all DoAs for the motion vector, at least three estimations are necessary. The first estimations are used to calculate $\bar{\omega}(t_m)$ and the last one to verify the model. If the motion can be explained by our model we mark all DoAs as valid estimations.

This motion model allows for filter out echo, because measurements that stem from echoes have a direction that is not consistent with the source, and they are timed shortly after the arrival of the original signal.

## IV. Experiments

### A. Evaluation Data Set

To evaluate the performance of our system in different and challenging conditions, we recorded static and moving speakers in an office building. We selected six representative rooms of different type and measured the reverberation time

TABLE I: Evaluation Data Set: Measured reverberation time $T_{60}$ and room size for six different room types.

| Room | $T_{60}$ [s] | Area [$m^2$] |
|---|---|---|
| Lab (large) | 1.158 | 291.3 |
| Lab (small) | 1.646 | 101.8 |
| Entrance Hall | 3.149 | 211.9 |
| Common Room | 1.971 | 80.28 |
| Lecture Hall | 1.077 | 141.97 |
| Office | 0.345 | 24.1 |


Fig. 3: The printed circuit board (PCB) with the 4 microphones (red circles). The microphones are spaced 1.5cm, 6cm and 9cm from the reference microphone on the right.

$T_{60}$ for each. Table I lists the measured $T_{60}$ time as well as the room sizes.

The data was recorded with a sensor array consisting of four microphones placed on a printed circuit board (PCB). A picture is shown in Figure 3. The microphones are arranged non-equally spaced over a distance of 9 cm, the positions of the microphones are marked by circles. The recording was done with a sampling rate of 16 kHz.

We created an evaluation data set with the aim to analyze different conditions where echo, reverberation and other effects degrade the localization performance. Hence, we placed the microphone array at different positions, to reflect a variety of scenarios for a robotic systems, and distances ranging from 3 m to 15 m. We took into account positions next to structures like walls or furniture, as well as placing the system in the center of the room. For example for the office room we placed the array into a corner next to two reflecting surfaces, centered in the room next to a desktop including screens and next to an open door.

### *B. Experiment Procedure*

The recorded data sets were fed to the different sound source localization algorithms. For our experiments we compared our method (MME-MUSIC) with the well established Generalized Singular Value Decomposition based MUSIC (GSVD-MUSIC) [24], and the recently published MUSIC with Active Frequency Range Filtering (AFRF-MUSIC) [29]. We do not consider any cross-correlation-based algorithms as they use a different approach than the previous mentioned subspace-based algorithms. In addition, most methods need a significant larger amount of sensor input for enabling the same theoretical accuracy [31]. The

TABLE II: Parameter constraints for experimental evaluation. If a parameter is applicable is indicated by a ✓.

| Parameter | Value | GSVD | AFRF | MME |
|---|---|---|---|---|
| $\omega_L$ [Hz] | 1000 | ✓ | ✓ | ✓ |
| $\omega_H$ [Hz] | 8000 | ✓ | ✓ | ✓ |
| $n_{\mathrm{FFT}}$ | 1024 | ✓ | ✓ | ✓ |
| $n_{\mathrm{Step}}$ | 64 | ✓ | ✓ | ✓ |
| $n_{\mathrm{Total}}$ | $4 \cdot n_{\mathrm{FFT}}$ | ✓ | ✓ | ✓ |
| $n_{\mathrm{Bins}}$ | 100 | | ✓ | ✓ |
| $t_{\mathrm{motion}}$ [s] | 0.5 | | | ✓ |
| $\theta_{\mathrm{tol}}$ [Deg] | 4.5 | | | ✓ |

used parameter set is given in Table II. We constrained all methods to a frequency band between 1 kHz and 8 kHz to remove low frequent system noise and focus on human speech. For each estimation a total frame of length $n_{\mathrm{Total}}$ was sliced into smaller frames of $n_{\mathrm{FFT}}$ points which are shifted by $n_{\mathrm{Step}}$. For the number of bins $n_{\mathrm{Bins}}$ the improved MUSIC methods shall process we took 100 as it showed to be a good trade-off for accuracy, processing time and estimation miss-matches. For our motion model we used a motion time $t_{\mathrm{motion}}$ of 0.5 s and a tolerated motion deviations $\theta_{\mathrm{tol}}$ of $4.5°$. Both have been determined empirically for static and dynamic sources.

We examine only true speech phase of each recording. Miss-classification of the VAD are not considered. For all positions we define a tolerated corridor of $\pm 2.5°$ around the ground truth to classify an estimation as successful or miss. Ground truth was obtained by measuring the angles of placed markers and positioning the speakers on them. $2.5°$ correspond to a deviation of approx. 20cm at a distance of 5m. This is a sufficient accuracy to recognize a speaker within a group of people standing next to each other.

In addition we evaluated the performance of our frequency band selection based on the noise score. Our goal was to reduce the computational cost which are introduced by each estimation step of the MUSIC response for each frequency band. Furthermore we wanted to use the wide spectrum of the human voice to be represented in our selection. In Figure 4 we show the selection of each algorithm for a received frame, from left to right GSVD, AFRF and MME. As previously described the selection is limited to the range from 1 kHz to 8 kHz for all methods.

As GSVD does not use a filtering technique to reduce the amount of frequency bands, it uses every received bin and feeds it to the estimator. AFRF focuses on the bin with the highest fourier coefficient corresponding to the primary frequency contributing the signal. The bandpass tremendously reduces the amount of calculations in subsequent steps, however as seen in the figure it is only using a small and limited portion of the signal. In contrast, the selection of MME is as wide as in the GSVD approach, but the amount of used bins is the same as in AFRF. The figure illustrates how the selection is gathering bins around the main frequencies in the signal while omitting frequencies which do not contribute.

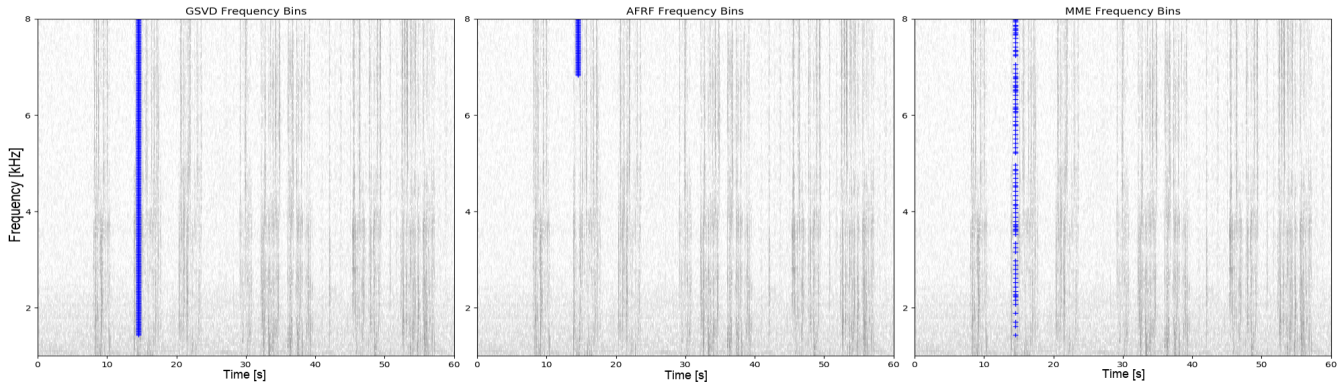


Fig. 4: Selected frequency bins for each algorithm. Each blue mark represents a selected frequency bin based on the algorithm's selection strategy. Left-hand side shows the GSVD approach, center the bandpass of AFRF and right-hand side the selected frequencies based on the noise score for MME.

TABLE III: Experimental results. The first columns present the total number of estimated DoA for each room, the last ones the rate of successful estimations.

| | $n_{\text{Total}}$ | | | success rate | | |
|---|---|---|---|---|---|---|
| Room | GSVD | AFRF | MME | GSVD | AFRF | MME |
| Lecture Hall | 263 | 263 | 229 | 0.91 | 0.79 | **0.95** |
| Common Room | 77 | 77 | 69 | 0.82 | 0.78 | **0.91** |
| Entrance | 78 | 78 | 39 | 0.72 | 0.46 | **0.95** |
| Office | 98 | 98 | 57 | 0.55 | 0.46 | **0.74** |
| Lab (large) | 73 | 73 | 49 | 0.78 | 0.64 | **0.82** |
| Lab (small) | 52 | 52 | 24 | 0.58 | 0.48 | **0.88** |

### C. Experimental Results

The comparison of GSVD-MUSIC, AFRF-MUSIC and MME-MUSIC for all rooms is summarized in Table III. It shows that our MME-MUSIC approach outperforms GSVD-MUSIC and AFRF-MUSIC in all experiment.

We want to discuss the results of the experimental evaluation exemplary on the lecture and entrance hall. The first one represents an environment with average acoustics, the latter one illustrates the worst case scenario with huge reverberation time $\mathrm{T}_{60}$ and numerous reflecting surfaces. The DoA estimation over time of each algorithm is displayed in Figure 5 and 6. A corresponding image of the environment is shown on the left-hand side of each. In Figure 7 we show exemplary one estimation result with corresponding ground truth and tolerated corridor. For better readability of the figures we skipped them for the rest of the evaluation.

In the lecture hall the GSVD-MUSIC algorithm has a good performance with overall 91% successful estimations. However it has some outliers which are created by echo of dominate sounds which can be seen at $t = 10\text{s}$, $t = 15\text{s}$ and $t = 22\text{s}$.

The AFRF-MUSIC algorithm is actively filtering for the main frequency in the current frame. This makes it faster than the standard GSVD approach and robust against other sources of noise, nevertheless it fails if the main frequency in the frame is not part of the source. Again at the endings of words, when the echo is dominant, AFRF-MUSIC solely focus on the frequencies which are created by the shadow source and neglects the frequencies of the original source. This yields to only 79% successful estimations.

MME-MUSIC introduces a motion model which checks if the estimated DoA is coherent to previous estimations. By that the algorithm removes outliers which were created from echo during the speech phases. The model not only considers static sources but also dynamic ones as the moving speaker at $t = 55\text{s}$. However the total amount of estimations is less compared to the other approaches. Despite that the rate of successful estimations is 95%.

The entrance hall is a more challenging environment for the algorithms as it has a high reverberation time and consists of a lot of reflecting surfaces. This is seen in the success rate of GSVD-MUSIC and AFRF-MUSIC with 72% and 46% respectively. In contrast MME-MUSIC has with 39 successful measurements a rate of 95%.

The results of these examples are consistent over the complete dataset.

Comparing execution times GSVD-MUSIC takes on average 1.049 s, AFRF-MUSIC 0.158 s and MME-MUSIC 0.208 s. This is a speed up of 5.1x of MME-MUSIC compared to GSVD-MUSIC. We think that the slower execution compared to AFRF is caused by cache-misses and we expect to remove this gap by optimizing the code for that.

## V. Conclusion and Future Work

In this paper we proposed our new MME-MUSIC approach for sound source localization in reverberant environments and under echoic conditions. We presented an intelligent way to select frequencies for the DoA estimation based on SEVD-MUSIC. We exploit the frequency evaluation of our VAD system and use the information to tighten the estimation process to the source bands. The results of the estimator are evaluated by our novel motion model. This takes into account the current motion of the speaker and is able to deal with static and dynamic sources.

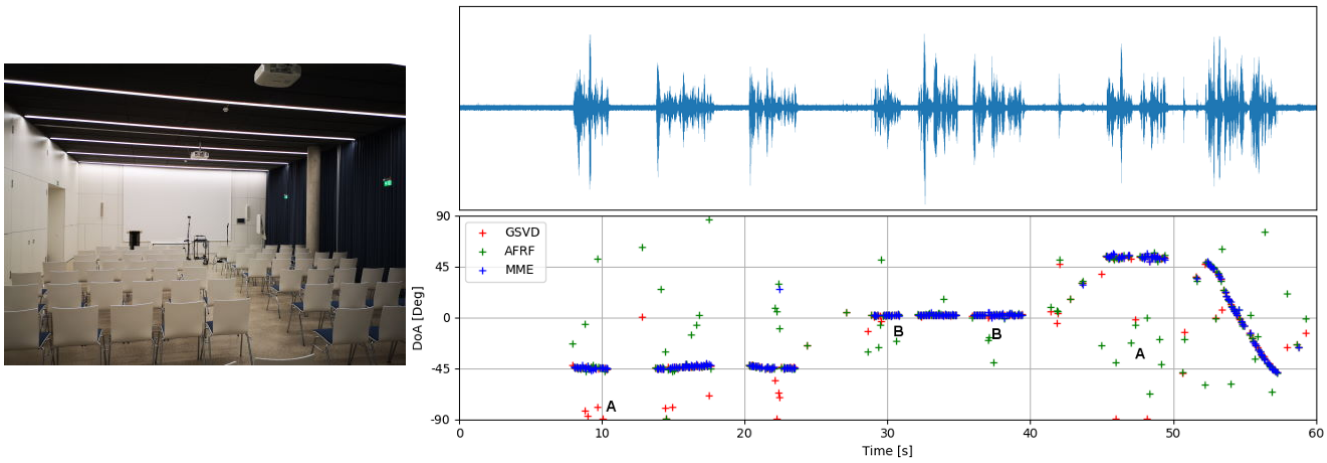


Fig. 5: Results of a recording in the lecture hall. We marked falsely estimations caused by echo with **A** and by reverberation with **B**. It can be clearly seen that MME is working better in this challenging scenario. Especially AFRF has miss-estimations at the silent endings of words.

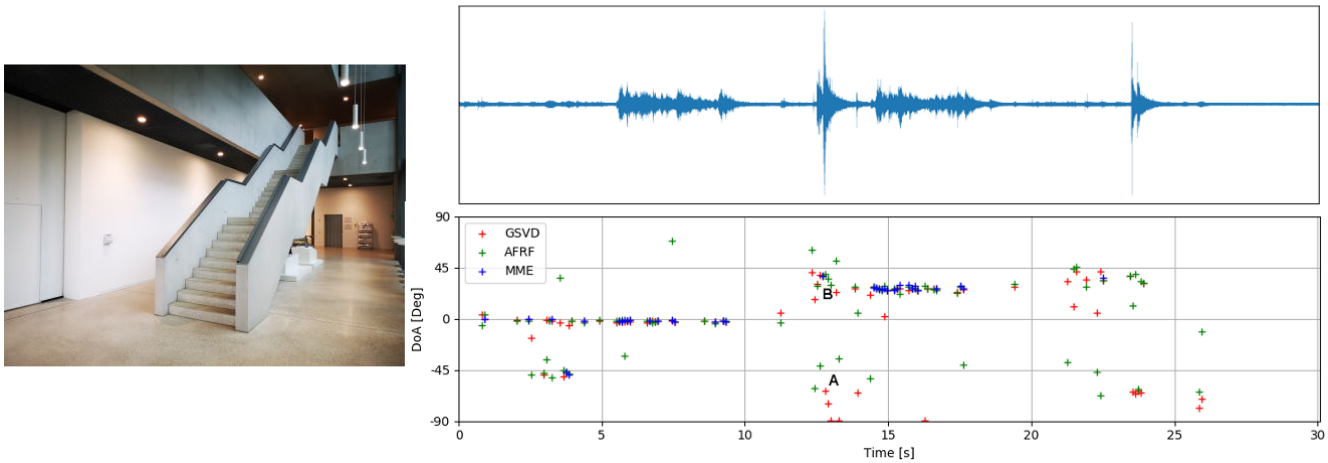


Fig. 6: Results of a recording in the entrance hall. We marked falsely estimations caused by echo with **A** and by reverberation with **B**. Because of the long reverberation time of more than 3.149s all methods have problems locating the source. At approx. $t = 13s$ a loud sound event first creates inaccurate estimations caused by reverberation, afterwards the receiving echo introduces a shadow source which confuses GSVD and AFRF.

We evaluated our approach using a four channel microphone array. We showed that our system performs well in realistic scenarios with reverberations and echo. Our MME-MUSIC approach outperforms established and state-of-the-art algorithms in these scenarios while preserving real-time execution times.

In total we expect to enable robot audition as a usable and useful technology for robotic systems by our enhancements. We plan to investigate further aspects of our work. First the use of the motion model directly in the estimation process. We believe constraining the estimator towards valid positions enhances accuracy while further reducing processing time. Second we want to extend our system to handle multiple sources at the same time. This makes it possible to use robot audition for mapping tasks or in highly complex scenarios like crowds.

For future work, we will integrate the sound source localization on our humanoid robot system Rollin' Justin. We designed a microphone array which is integrated in the head of the system [32]. Here we will have to tackle further challenges, like compensating robot intrinsic noise and extend our system to a more complex array geometry due to design limitations of the robot system. We will use our technique to robustly detect speakers in a conversation.This can be used to enhance the acceptance of a robot as the system acknowledges the speaker by turning the head towards him or to annotate received speech to a specific speaker.

## References

[1] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, p. 664–669.

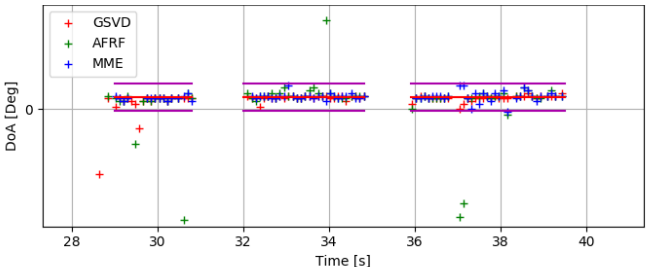


Fig. 7: Estimation results for the lecture hall dataset. The lines show the ground truth and the tolerated deviation ($\pm 2.5°$). Miss-classifications of the VAD system without ground truth data are removed.


[2] K. Nakadai, K.-i. Hidai, H. Mizoguchi, H. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," in *Artificial Intelligence. Proceedings. 17th International Joint Conference on*, 2001, pp. 1425–1432.

[3] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Intelligent Robots and Systems. Proceedings. IEEE/RSJ International Conference on*, vol. 2, 2003, pp. 1147–1152.

[4] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robots: sound localization and separation," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, 1997.

[5] L. A. Jeffress, "A place theory of sound localization." *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, p. 35, 1948.

[6] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[7] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," in *Acoustics, Speech and Signal Processing (ICASSP). Proceedings. IEEE International Conference on*, vol. 5, 2006.

[8] I. Kossyk, M. Neumann, and Z.-C. Marton, "Binaural bearing only tracking of stationary sound sources in reverberant environment," in *Humanoid Robots (Humanoids), IEEE-RAS 15th International Conference on*. IEEE, 2015, pp. 53–60.

[9] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Intelligent Robots and Systems. Proceedings. IEEE/RSJ International Conference on*, vol. 2. IEEE, 2003, pp. 1228–1233.

[10] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Robotics and Automation. Proceedings. IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 1033–1038.

[11] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Intelligent Robots and Systems. IEEE/RSJ International Conference on*. IEEE, 2006, pp. 380–385.

[12] E. Mumolo, M. Nolich, and G. Vercelli, "Algorithms for acoustic localization based on microphone array in service robotics," *Robotics and Autonomous Systems*, vol. 42, no. 2, pp. 69–88, 2003.

[13] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Deutsche Jahrestagung für Akustik (DAGA)*, 2015, pp. 1510–1513.

[14] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[15] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2015, pp. 2814–2818.

[16] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 603–609.

[17] Y. Sasaki, N. Hatao, K. Yoshii, and S. Kagami, "Nested igmm recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3930–3936.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[19] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *Acoustics, Speech, and Signal Processing. IEEE Transactions on*, vol. 37, no. 7, pp. 984–995, 1989.

[20] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *Intelligent Robots and Systems. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 2009–2014.

[21] F. Asono, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot" jijo-2"," in *Multisensor Fusion and Integration for Intelligent Systems. Proceedings. IEEE/SICE/RSJ International Conference on*. IEEE, 1999, pp. 243–248.

[22] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Intelligent Robots and Systems. IEEE/RSJ International Conference on*. Institute of Electrical and Electronics Engineers, 2009, pp. 2027–2032.

[23] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 694–699.

[24] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, p. 694–699.

[25] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, p. 3288–3293.

[26] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep 2014, p. 1902–1907.

[27] G. Chardon, "A block-sparse music algorithm for the localization and the identification of directive sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, p. 3953–3957.

[28] R. Takeda and K. Komatani, "Noise-robust music-based sound source localization using steering vector transformation for small humanoids," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, p. 26–36, Feb 2017.

[29] K. Hoshiba, K. Nakadai, M. Kumon, and H. G. Okuno, "Assessment of music-based noise-robust sound source localization with active frequency range filtering," *Journal of Robotics and Mechatronics*, vol. 30, no. 3, p. 426–435, 2018.

[30] 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*, vol. 42, no. 3-4, pp. 271–287, 2004.

[31] A. Pourmohammad and S. M. Ahadi, "N-dimensional n-microphone sound source localization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 27, 2013.

[32] M. Sewtz, T. Bodenmüller, and R. Triebel, "Design of a microphone array for rollin' justin," 2019.