Sound Source Localization for Robotic Applications

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 an extension of our proposed sound source localization approach Motion Model Enhanced MUltiple SIgnal Classification (MME-MUSIC) for segmenting speech input.**

We evaluate the system with speech captured under real conditions in an experimental setup and show the use of our estimation in real applications.

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 and may also be occluded by obstacles. 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.

We presented a novel approach for localization of speakers in reverberant and echoic environments by use of a microphone array in [1]. We classify received audio streams as speech or non-speech using a voice activity detector (VAD). We transform the signal into the frequency domain and analyze the fourier coefficients to calculate a score. Afterwards we select the most significant bins and fed them into our direction of arrival (DoA) estimator. Further on we propose a motion model to check the calculated direction spectrum to improve the robustness.

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.

In this work we want to show the application and the use of the motion model to segment received speech and assign them to different speakers. We deliberately avoid using other techniques like mel cepstrum analysis [2]–[5] or vision-based aid [6]–[9] to illustrate the performance of a single DoA estimator.

II. Related Work

At first, research focused on imitating the binaural audio localization of animals and humans [10]–[13]. Using both the interaural phase difference (IPD) and the interaural intensity difference (IID). Further, some techniques take into account the head-related transfer function [14], [15] as well as the prior information on 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 [16]. Nonetheless these systems need a demanding hardware setup and calibration.

Other approaches use an array of microphones to overcome the challenging requirements on the hardware and to estimate the direction of arrival (DoA) of a received signal [17], [18]. It is possible to calculate the most probable DoA by estimating the time delay between the signals received by each microphone. Combining these methods with delay and sum beam forming (DSBF) as well as random sample consensus (RANSAC), more than one sound source can be

[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

localized simultaneously [19]. 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 [20]–[24].

Recently, exploiting the properties of the subspace as in Multiple Signal Classification (MUSIC) [25] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [26] 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 [27]–[30].

Several extensions have been proposed for enhancing the performance of MUSIC, e.g. using singular value decomposition [31] to reduce the computational complexity while enhancing robustness against noise. Incremental versions are introduced to reach real-time performance while enhancing robustness against noise [32], [33]. Additional research to further reduce the computational costs in the representation space is done in [34], [35].

However, even recent sound source localization systems face problems when detecting humans in indoor scenarios under non-optimal acoustic conditions. We identified significant effects that degrade the performance, namely reverberation and echo. The first one is the reflection of numerous acoustic wavelets at every surface which results in a "fading-out" effect and lower SNR. The latter one is the complete reflection and delayed reception of the original source. This leads to miss-classification.

III. System

Our system Motion Model Enhanced Multiple Signal Classification (MME-MUSIC) is based on the SEVD-MUSIC [36] approach. We enhance the process by limiting the estimation only to speech phases classified by the voice activity detector. Furthermore we reduce the number of frequency bins by selecting the most significant ones based on a score calculated in the previous step. Additionally we post-filter our results using a motion model. Lastly we exploit the decision of the model to segment the speech and assign it to the speakers. For capturing the audio we use a microphone array consisting of four acoustic sensors. An overview on the system is illustrated in Figure 2.

A. Voice Activity Detector and Band Selection

We use the VAD proposed by Ramírez *et al.* to classify the incoming signal [37]. First, we transform the audio into the frequency domain. Afterwards, we use the Longterm Speech Divergence (LTSD) approach 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.

Subsequently we use the gained information on the difference of individual frequency bins compared to noise to find the significant components. This enables the reduction of calculation costs while preserving estimation accuracy.

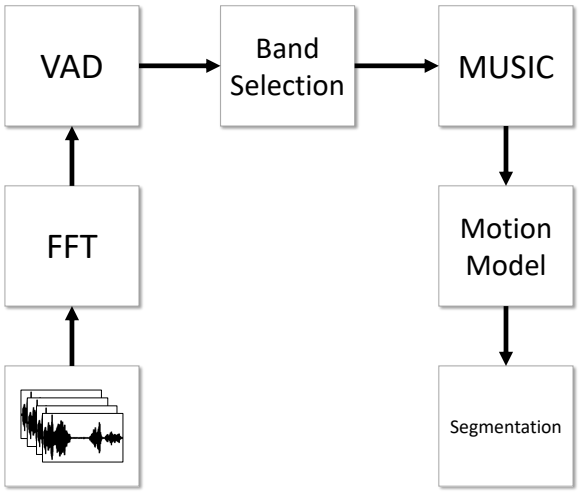


Fig. 2: System overview.

B. DoA Estimation

We assume that our received signal consists only of the direction-depending source signal and independent system noise. The approach of MUSIC exploits this dependency and decomposes the transformed audio into noise and source subspace. Ultimately it tries to find the corresponding direction vector which fulfills the constraints given by the system and the subspaces. We repeat this estimation for all selected frequency bins and accumulate a total pseudospectrum to reflect the direction dependencies.

C. Motion Model and Segmentation

We check the plausibility of the estimated angle by evaluating it with a motion model. To do this, we assume that for a given time span the source moves with mean angular velocity. We take into account a constant motion tolerance to cope with dynamic changes and measurement noise.

When receiving a new DoA from the previous steps we gather all estimation within the time span. If we can explain the measurement given our motion model, we flag them as valid. We need at least 3 valid estimations, the first ones to calculate the motion vector, the last one to verify the model.

Furthermore we exploit the verification for our segmentation. We consider a scenario with two persons speaking. If we receive new measurements which are marked as valid but based on a different motion vector than previous measurements, we assign them to a different speaker. This is a fairly naïve approach, however the performance shown in the next section is notable.

IV. Experiments

We show the application of our sound source localization in a segmentation process where two persons are having a conversation. Our system uses the estimated position to assign the speech to the corresponding speaker. The scenario is shown in Figure 3. We illustrate both cases, the speakers facing the system and each other. We assume the last one as

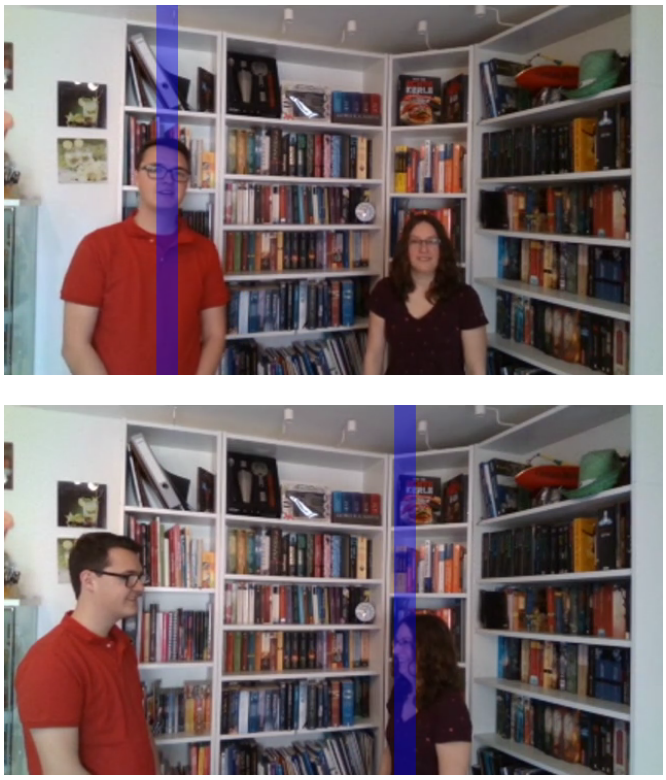


Fig. 3: Conversation between two person. Left side shows the case where both of them are speaking towards the camera. In this scenario a vision-based system may lead comparable performance. Right side shows the case where both speakers are facing each others. This is a hard task for vision classifier. As indicated by the blue bar, the auditory system succeeded in identifying the current speaker.

a hard task for camera-based systems, as the visual clues for identifying the speaker are reduced to a minimum.

We compare our approach with AFRF-MUSIC [38], which is an optimized version of SEVD-MUSIC [31] according to execution time. In contrast, our approach is also optimized for use in indoor scenarios.

For AFRF-MUSIC we add the information, that the left speaker can be localized by positive angles, the right speaker by negative, as the system has no indicator for changing sources.

We manually labeled the data for left and right speaker and compare it with the outcome of the algorithms. The results are shown in Figure 4.

For AFRF-MUSIC we get correct assignment in 79.5% of all estimated cases, for MME-MUSIC in 93.1%. In total comparing all cases where the approaches did not assign a speaker, AFRF-MUSIC performs with 61.0% and MME-MUSIC with 73.0% successfull assignments (see Table I).

TABLE I: Segmentation results.

	Total		Segmented	
Method	TP	FP	TP	FP
AFRF	61.0%	39.0%	79.5%	20.5%
MME	73.9%	26.1%	93.1%	6.9%

V. Conclusion

In this work we showed the application of our recently developed sound source localization system Motion Model Enhanced MUSIC (MME-MUSIC). We shortly introduced the pitfalls of indoor scenarios and the resulting effects on auditory systems. We developed a simple segmentation algorithm based on our approach to assign speech phases of a received signal to specific speakers. Furthermore we showed that this naïve approach is reliable enough in situations where classical approaches using vision-based systems may fail to locate the correct speaker.

With this work we want to propagate the benefit of using robot audition as an additional modality for robust robotic systems. We expect enhanced perception systems which operate robustly in complex environments.

References

[1] M. Sewtz, T. Bodenmüller, and R. Triebel, "Robust music-based sound source localization in reverberant and echoic environments," in *Intelligent Robots and Systems (IROS). IEEE/RSJ International Conference on*, 2020, submitted.

[2] M. R. Hasan, M. Jamil, M. Rahman *et al.*, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[3] S. Agrawal and D. Mishra, "Speaker verification using mel-frequency cepstrum coefficient and linear prediction coding," in *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*. IEEE, 2017, pp. 543–548.

[4] A. Charisma, M. R. Hidayat, and Y. B. Zainal, "Speaker recognition using mel-frequency cepstrum coefficients and sum square error," in *2017 3rd International Conference on Wireless and Telematics (ICWT)*. IEEE, 2017, pp. 160–163.

[5] A. Awais, S. Kun, Y. Yu, S. Hayat, A. Ahmed, and T. Tu, "Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing," in *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 2018, pp. 271–276.

[6] J. M. Rehg, K. P. Murphy, and P. W. Fieguth, "Vision-based speaker detection using bayesian networks," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 110–116.

[7] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 3. IEEE, 2000, pp. 1589–1592.

[8] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *European Conference on Computer Vision*. Springer, 2016, pp. 285–301.

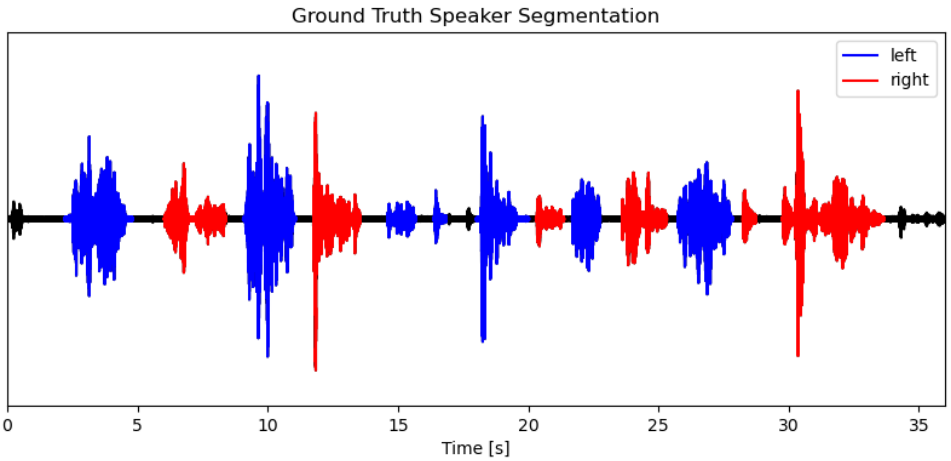
[9] K. Stefanov, J. Beskow, and G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Grounding Language Understanding GLU2017 August 25, 2017, KTH Royal Institute of Technology, Stockholm, Sweden*, 2017.

[10] 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.

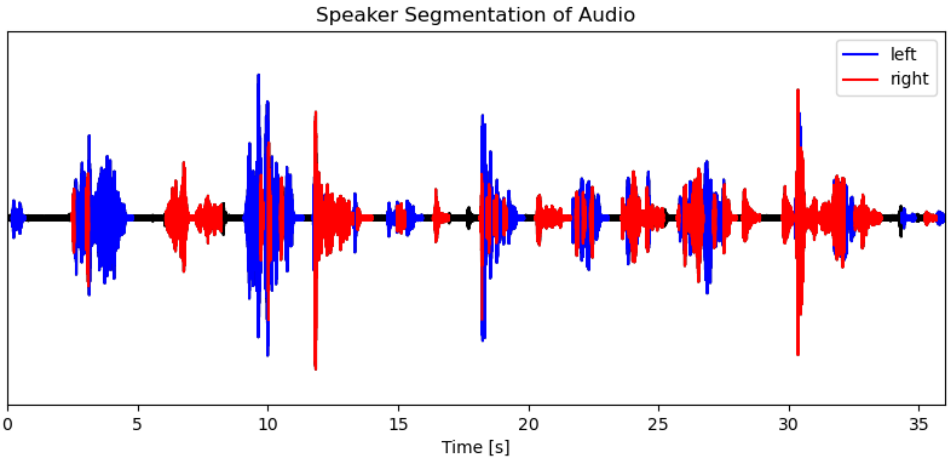
[11] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Intelligent Robots and Systems. Proceedings. IEEE/RSJ International Conference on*, vol. 2, 2003, pp. 1147–1152.

[12] 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.

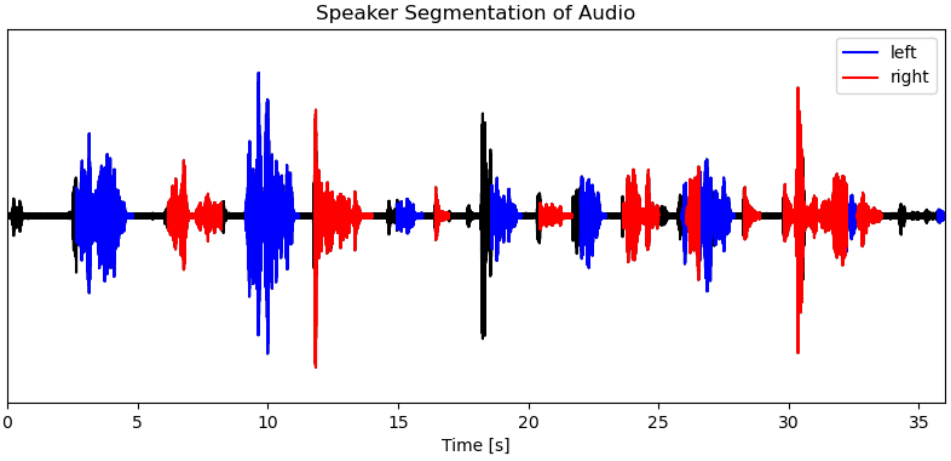
[13] L. A. Jeffress, "A place theory of sound localization." *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, p. 35, 1948.



(a) Ground Truth



(b) AFRF-MUSIC method



(c) MME-MUSIC method

Fig. 4: Experimental results of our segmentation. Top shows the manually labeled ground truth. Center shows the result using AFRF-MUSIC, a state-of-the-art and real-time capable approach. Bottom shows our approach using MME-MUSIC. It can be seen, that AFRF-MUSIC has a lot of miss-classifications. MME-MUSIC has less segmented points while having better assignments.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] ——, "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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.