Acoustic Odometry for Wheeled Robots on Loose Sandy Terrain

Andreu Gimenez Bolinches
Graduate School of Science and Technology
Keio University, Japan
andreu@keio.jp

Genya Ishigami
Graduate School of Science and Technology
Keio University, Japan
ishigami@mech.keio.ac.jp

***Abstract*—This work proposes a system capable of estimating the longitudinal velocity of a wheeled robot on loose sandy terrain using only acoustic sensors. These will likely be present anyway in many future ground mobile robots for human-robot interaction purposes. The proposed system consists of an audio feature extraction module, based on gammatone filterbanks, and a prediction module, based on a convolutional neural network. The proposed system has been tested in a single wheel test bed with a wheel driving up to speeds of 0.07 m/s in a wide range of wheel slippage resulting in an average drift of 5 mm/s. A qualitative evaluation of the proposed system against other sources of odometry shows that acoustic and visual methods vulnerabilities do not overlap, which indicates that a system based on acoustic sensors can be a feasible auxiliary source of odometry. The system is able to make a prediction from a single audio frame with a duration of 15 ms in only 2 ms on a user-level commercially available CPU.**

***Index Terms*—Audio-Visual SLAM, Robot Audition, Deep Learning Methods**

I. Introduction

A. Problem

Self-localization is one of the most fundamental competencies required by an autonomous robot as the knowledge of its location is an essential precursor to making decisions about future actions.

One of the simplest forms of self-contained localization is wheel odometry, based on wheel encoders that are mounted on a robot to track the number of revolutions each wheel has made. The number of revolutions is integrated into a dynamic model to determine the robot's current position relative to the starting point [1]. But it performs poorly in the presence of wheel slippage, accumulating position error (drift).

Another form of self-contained localization is visual odometry, which operates by incrementally estimating the pose of the vehicle through examination of the changes that motion induces on the images of its onboard cameras [2]. Similarly, laser odometry estimates the ego-motion of a vehicle by scan-matching of consecutive laser scans. Unlike wheel odometry, both visual odometry and laser odometry, are not affected by wheel slip, but they have other drawbacks: visual odometry suffers from poor illumination and low textured environments; laser might struggle in degenerated scenes where planar areas are prevalent; both are sensitive to dynamic environments.

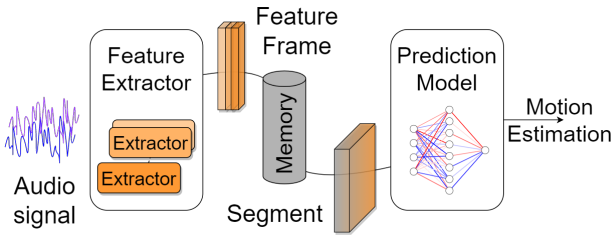


Figure 1: Proposed system

Research that improves the robustness of robot localization can be divided into two trends: To fuse information from different sensors to overcome their individual limitations [3]–[5] and to make better use of the available sensor information using deep learning and other computationally expensive algorithms [6], [7]. Therefore, there is a cost associated with increasing the robustness of robot localization. Whether it is the explicit price of extra sensors or the implicit cost of extra computational resources.

B. Motivation

One way to minimize the price of robust ground robot localization in unstructured environments is using sensors that are relatively inexpensive, imply a low overhead, and might be needed for other functionalities of the robot. Many ground mobile robots will be equipped with sound sensors for the purpose of human-robot interaction [8], at the same time, audio signals have been extensively studied, resulting in very efficient methods for their processing. Therefore, this work intends to asses the feasibility of acoustic odometry as an auxiliary source of odometry.

This work focuses on wheeled robots on loose sandy terrain. Although strides have been made into exotic forms of legged robots, wheels or tracks still form the basis for robot locomotion [9]. Wheeled mobile systems are useful for practical applications compared to legged systems because of the simplicity of the mechanisms and control systems and efficiency in energy consumption [10]. However, wheeled systems' performance depends on the traction between the wheels and the ground. If there is not enough traction, the wheel will slip and the efficiency will decrease. Traction is a special concern when the robot is expected to move over granular non-cohesive loose soil. Planetary exploration,

construction and agriculture are some of the applications where robots operate on sandy terrain.

C. Related work

Audio-based Odometry is a relatively unexplored field, but robot audition is present in self-localization and navigating tasks using Sound Source localization, where the robot's ego-noise is usually seen as a nuisance.

a) Audio-based Odometry: In [11], the authors propose a classification framework to associate ego-noise captured with an onboard microphone to a set of predefined velocity profiles. Additionally, they are able to detect a change in the inclination of the surface the robot is moving. However, the application of this framework is rather limited. On the other hand, [12] proposes a system capable of estimating ground robots' linear and angular velocities using onboard audio sensors. It uses deep neural networks to regress the motion of a vehicle from feature representations (based on Gammatone filterbanks) of the sensed audio. The authors conclude that audio-based odometry systems should be useful auxiliary sources of odometry on the side of more traditional systems. However, their evaluation is limited to a single experiment which is randomly split in a training and test dataset, without a validation split and they do not evaluate the computational cost of their solution.

b) Sound Source localization: In [13] and [14], the authors propose an algorithm to simultaneously localize a robot and map its environment (SLAM) using onboard audio sensors that perceive sound sources in its environment. Alternatively, [15] perceives the robot's intrinsic noise to localize it using external audio sensors. Combinations with other self-localization methods are proposed in [16], where onboard sound sensors identify and remove the effect of dynamic obstacles for Visual SLAM, and [17], which localizes sounds using onboard microphones and uses them as navigation goals while using Visual SLAM for self-localization.

Section II describes the proposed system while section III shows how the proposed system is trained and tunned for a wheeled robot on loose sandy terrain. Section IV evaluates the proposed system, including a discussion of the results and recommendations on future research on acoustic odometry that can be found in section IV-D. Finally, section V summarizes the most important contributions of this research.

II. Methods

This work proposes a system capable of estimating motion from audio signals. Shown in figure 1, it is a modular system that consists of two components: a Feature Extraction module composed of several *extractor* submodules and a Prediction Model. The proposed system is implemented in C++ as a Python extension module and it is publicly available in github under the Acoustic Odometry organization[1]

[1]https://github.com/AcousticOdometry

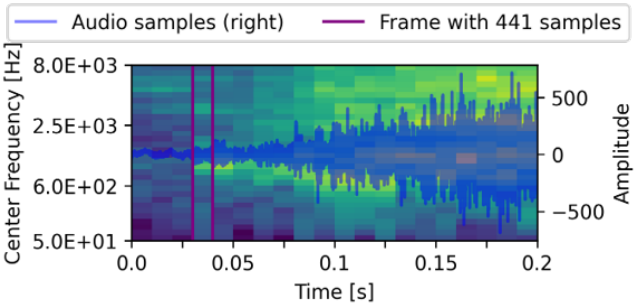


Figure 2: Audio signal (in blue) over its features extracted from frames with a duration of 10ms with a single Gammatone filterbank extractor.

A. Feature Extraction

The feature extraction module receives a multi-channel audio signal of a fixed length and outputs a *feature frame*. Which is defined as the group of vectors of features extracted from a finite number of samples. One vector per *extractor* used. Figure 2 illustrates how an audio signal is divided into frames.

The term *extractor* will be used to abstract the system used to extract features from audio data. This work opts to use a feature extractor based on Gammatone filterbanks as [12] in since the performance of classification systems relying on Mel Frequency Cepstrum Coefficients is greatly reduced in the presence of noise.

Gammatone filterbanks are an approximation to the human cochlear frequency selectivity originally introduced in [18]. Time-independent features are obtained by filtering the audio waveform with a bank of gammatone band-pass filters.

This work implements a bank of fourth order gammatone filters with its corresponding bandwidth of 1.019 ERB where ERB is the equivalent rectangular bandwidth scale [19]. The filters are linearly distributed over a predefined frequency range on the ERB scale. The number of filters used is equivalent to the number of features to be extracted.

B. Prediction Model

Consecutive feature frames from the Feature Extraction module are concatenated in order to form a *segment*. This segment is then fed to the prediction model. Which will output a motion estimation.

This work proposes a shallow Convolutional Neural Network [20] composed by two convolution layers, each of them followed by a max pooling layer, and two fully connected layers, each of them preceded by a dropout, as shown in figure 3.

The input dimensions depend on the Feature Extraction parameters, namely number of features per frame, number of extractors and number of frames per segment. Layer sizes and the output of the last fully connected layers are seen as tuneable parameters as well. Experiments are conducted with the different parameters and a model is selected based on its performance and computational cost.

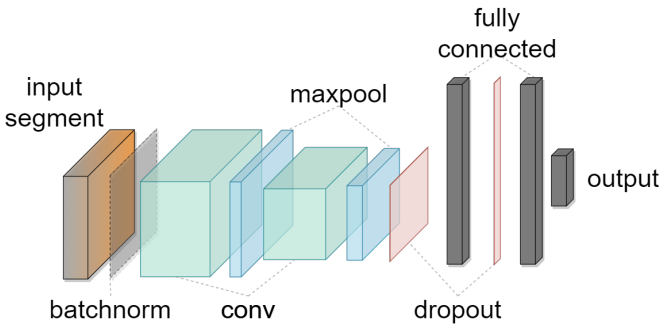
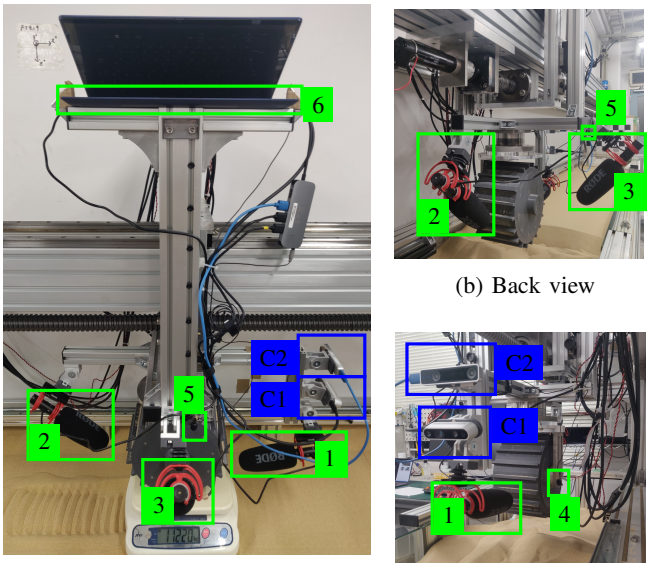


Figure 3: Shallow Convolutional Neural Network architecture proposed in this work. The selected model is composed by: a batch normalization layer; a first convolution layer with one input channel, a kernel size of 5 and 16 output channels; a second convolution layer with 32 output channels and a kernel size of 2; a first fully connected layer with an output size of 256; a second fully connected layer with an output size of 28. Each of the convolution layers is followed by a maxpool layer with a kernel size of 2 and each of the fully connected layers is preceded by a dropout layer with a probability of 50%.



(a) Lateral view

(b) Back view

(c) Front view

Figure 4: Lateral (*a*), back (*b*) and front (*c*) view of the wheel test bed carriage unit. The positioning of the microphones takes into account ideal positions of a microphone attempting to capture the sound generated by the wheel interacting with the terrain (1 and 2), feasible positions of a microphone in a mobile robot (3, 4, and 5), and a reasonable position of a microphone dedicated to human-robot interaction (6)

III. Experiments

Due to the difficulty in finding publicly available audio odometry datasets, the problem of robot localization is simplified to a single dimension over a unique terrain type: A wheel that can only move along a longitudinal axis over loose sandy terrain. This scenario can be easily reproduced with the available Experimental Setup, making it possible to gather new Datasets.

A. Experimental Setup

This work uses a single wheel testbed, composed of a carriage unit attached to a frame over a soil bin shown in figure 4. It is 3500 mm in length, 600 mm in width, and 1200 mm in height. The soil bin is filled with sandy soil with a depth of 200 mm. The surface of the sand can be leveled or sloped before each test run using a leveling apparatus attached to the sandbox.

The carriage unit can move in the longitudinal direction at an arbitrary velocity using the ball screw while the unit freely moves in its vertical direction. The wheel is placed at the bottom of the carriage wheel with an independent traction motor. Therefore, this testbed can produce an arbitrary slip ratio (as defined in [21]) by varying the angular speed of the wheel and the longitudinal velocity of the unit.

Additionally, the carriage unit can be detached from the ball screw to allow free longitudinal movement. In this way, the wheel can be driven with free slip recreating realistic driving conditions over sandy terrain while being able to keep accurate monitoring of its position.

B. Datasets

1) Training: A collection of overlapping audio frames used to train and test the Prediction Model. Audio frames are annotated with longitudinal velocity, the wheel angular speed and the slip ratio with a weighted average of the measurements within the frame duration.

Audio frames are extracted from test recordings where the wheel test bed is used to control different variables as the slip ratio and the wheel angular velocity. 168 different recordings that make a total of 1 hour for each of the 6 microphones where captured for the training dataset.

2) Evaluation: 7 recordings with a total duration of 17 minutes used to evaluate the performance of the system under more realistic and challenging conditions. Unlike in the test dataset, where the longitudinal position was controlled by the wheel test bed, the carriage unit is allowed free longitudinal movement and only the wheel drive motor is active. High wheel slippage is induced by external forces and a slopped terrain.

C. Selected model

This work selects a model trained on a dataset of segments of 50 frames, each of them spans over 15 milliseconds and has 64 features from a single Gammatone filterbank extractor. The extractor is applied on the average of the audio signal channels with a frequency range of [50, 8000] Hz and features are represented on the Bel scale. The model is trained with data from a subset of microphones (1, 2, 4 and 6 as shown in figure 4). Training data is augmented with added random SNR noise. The model classifies the longitudinal velocity given a set of 28 different linearly distributed longitudinal velocity ranges taking into account the order of the classes, ordinal classification, as introduced

	Mean ATE [m]	Mean RPE [m]
All microphones	$1.02e-03$	$5.10e-03$

Table I: Selected model average performance across all evaluation recordings and averaged across all microphones. Absolute Trajectory Error (ATE) is computed between frames with a duration of 15 ms and Relative Pose Error (RPE) [23] is computed using time windows 1 s long.

in [22]. Therefore the output of the model is a vector of probabilities that can be decoded into a class position by making use of a ranking rule. The loss is computed with the mean square error between the target class position and the predicted class position.

IV. Results

Table I shows the average performance of the selected model in the Evaluation dataset. Figure 5 shows it applied to an evaluation recording characterized by high slippage driving conditions.

A. Noise

The selected model has been evaluated in the same recording shown in figure 5 in the presence of synthetic white noise of varying signal-to-noise ratio values. Noise with the same power as the signal (SNR 0 dB) increases the mean Relative Pose Error 2.62%. SNR of -10 dB affects significantly the predictions at low speeds and when the robot is not moving increasing the RPM a 38.91%.

B. Computational cost

Tests with the selected model on a CPU show that the time to compute the feature extraction is a 40% of the total time to process a new frame, which is 2.03 ms on an Intel® Core™ i7-9750H CPU at 2.60GHz. The prediction time takes up a 55% of the total time while the remaining 5% corresponds to loading the new frame into memory. The system is able to run 7.5 times faster than real time on this particular hardware. Using hardware acceleration with CUDA the total time to process a new frame is slightly lower: 1.93 ms on a NVIDIA GeForce RTX 2060 GPU.

C. Compared with other models

The selected model is evaluated against two other odometry methods: Wheel odometry computed from the measured wheel angular velocity; and Intel® RealSense™ Tracking Camera T265, which is based on visual odometry. Figure 6 is the evaluation in an ideal scenario for all methods. Figure 7 instead corresponds to a scenario that is challenging for visual odometry. Dynamic objects are moved in the camera field of view and lightning conditions are changed during the recording. One can see in this evaluation that the vulnerabilities of audio-based odometry and visual-based odometry do not overlap.

D. Discussion

Figure 5 shows that the overall predicted curve is very similar when the selected model is applied to the to the output of different microphones. Meaning that even if the absolute value of the predicted velocity is not accurate, the model is able to recognize slippage conditions across different microphones. Even with microphones not present during training, as is the case in figure 6, where none of the microphones shown in the evaluation were used for training. This indicates that fine-tuning [24] a model might significantly increase its performance on a given microphone.

The selected model has been evaluated in presence of synthetic white noise. It's performance was not significantly affected with signal-to-noise ratio of 0 dB (noise with the same power as the signal). Nevertheless, scenarios with multiple powered wheels have not been evaluated.

In "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals" [12], the authors claim that their work demonstrates an absolute error of 0.065 m/s and 0.02 rad/s but they split their training and test set randomly. Their audio frames are generated with a sliding window of 1 s with 100 ms overlap which means that up to 20% of a test audio frame might have been present in test audio frames. Instead, this work uses a a separate set of recordings for evaluation and conditions never seen in the training dataset.

The method presented in "Current-Based Slippage Detection and Odometry Correction for Mobile Robots and Planetary Rovers" [5], where wheel slippage is identified using the motor current and compensated in the wheel odometry computation, outperforms the proposed system while being a computationally inexpensive method. They demonstrate accumulated drift of up to 1% of the traveled distance using the same range of longitudinal velocities under 0.07 m/s, while the proposed system results in an average error of 16% of the traveled distance across all microphones and evaluation recordings. However, the authors mention that current-based slippage detection fails to correctly identify a wheel slipping over rocks instead of sand. Audio-based methods do have the potential to identify wheel slippage in both scenarios.

This work comes with its drawbacks too, However the proposed system has been evaluated in only one terrain.

V. Conclusions

The proposed system is capable of estimating the longitudinal velocity of a wheeled robot on loose sandy terrain using gammatone-based features extracted from robot ego-noise data and an ordinal classification model implemented as a convolutional neural network. Said system has been trained and evaluated in a new multi-modal dataset collected using a single wheel testbed and several microphones and cameras. The evaluation contains high wheel slippage scenarios where the proposed system achieves an average drift of 5 mm per second with longitudinal velocities under 0.07 m/s. The proposed system runs 7.5 times faster than real time on a user-level hardware.

Claiming that the proposed method estimates robot odometry in its current form would be an exaggeration. The

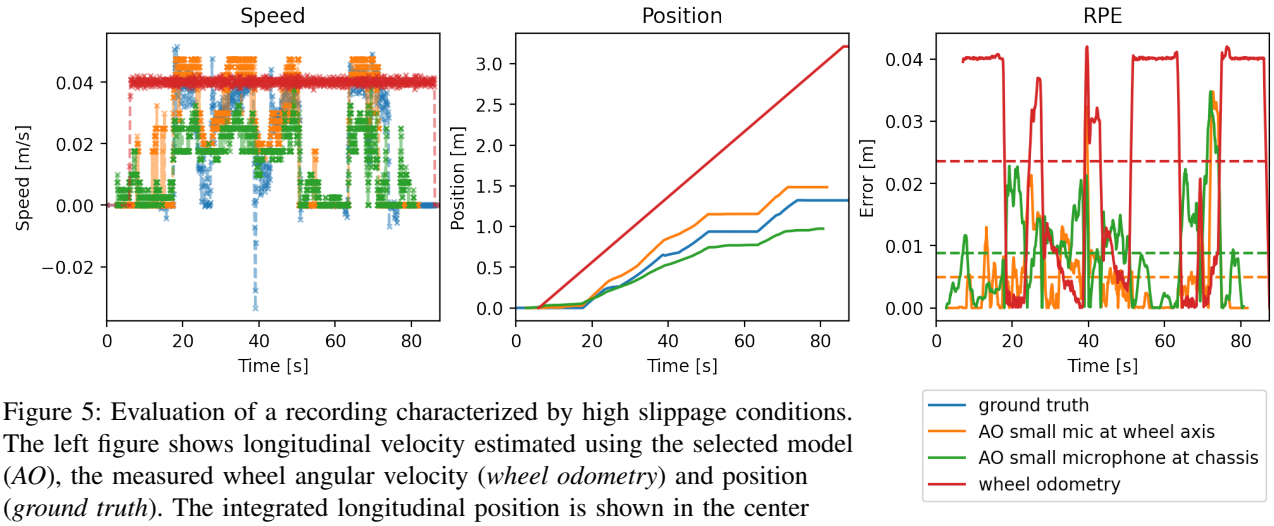


Figure 5: Evaluation of a recording characterized by high slippage conditions. The left figure shows longitudinal velocity estimated using the selected model (*AO*), the measured wheel angular velocity (*wheel odometry*) and position (*ground truth*). The integrated longitudinal position is shown in the center figure. The right figure shows the evolution of the Relative Pose Error [23] computed using time windows 1 s long.

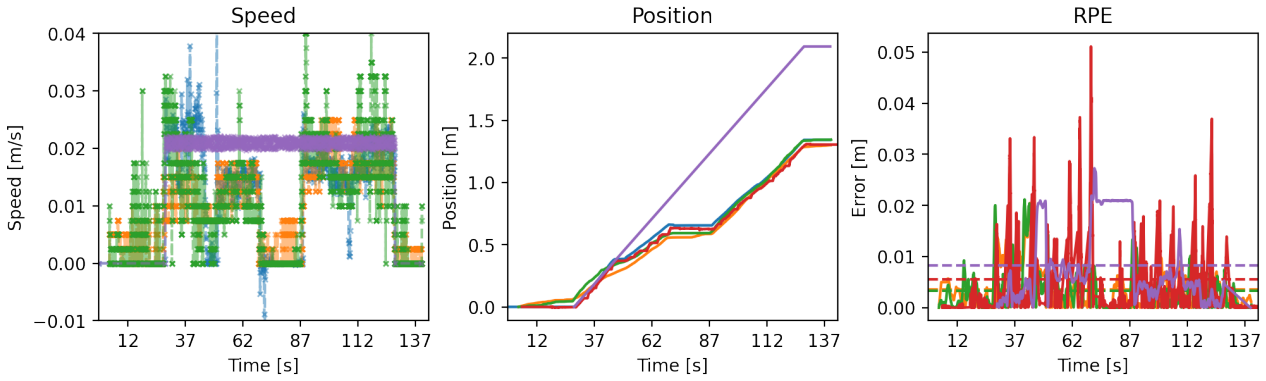


Figure 6: Selected model evaluated evaluated against wheel odometry and Intel® RealSense™ Tracking Camera T265 (based on visual odometry). One can see in the central figure that the proposed system can perform with an accuracy comparable to commercially available visual-based systems in this simplified scenario.

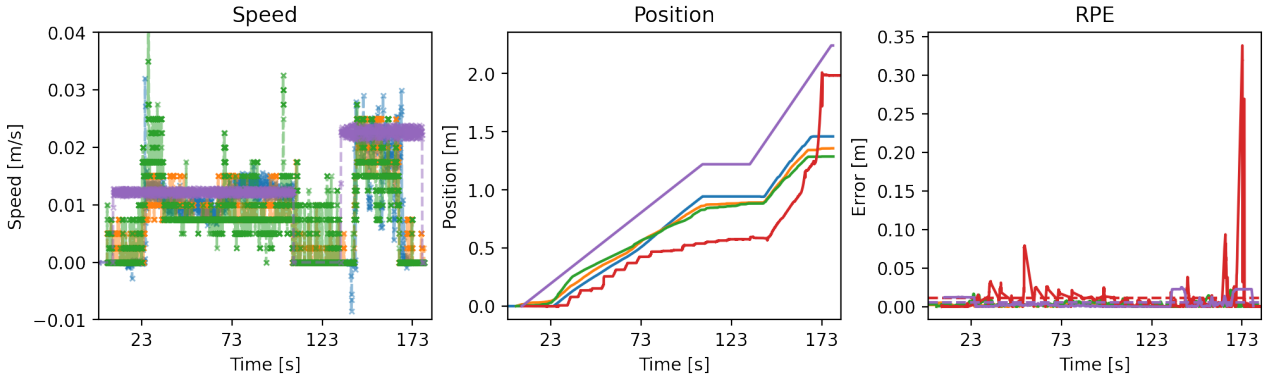


Figure 7: Selected model evaluated against Intel® RealSense™ Tracking Camera T265 (based on visual odometry) in a scenario where visual odometry vulnerabilities are exploited: Dynamic objects are moved in the camera field of view and lightning conditions are changed during the recording.

experiments presented in this paper correspond to an over-simplified scenario. Nevertheless, this work demonstrates that it is feasible to estimate longitudinal velocity of a wheeled robot on loose sandy terrain from acoustic data with a computationally inexpensive system. Which indicates that an acoustic odometry system can be a feasible auxiliary source of odometry and add robustness to the localization system without incurring in significant costs.

Finally, in recent years, single-robot simultaneous localization and mapping research is steadily moving toward systems that can build metric-semantic maps. Acoustic odometry could be combined with terrain classification in order to provide both, an auxiliary source of odometry and semantic information for a simultaneous localization and mapping system. Similarly, a system that not only estimates motion based on ego-noise but also is able to identify and subtract said ego-noise from the audio signal would be useful to improve the performance of speech recognition in collaborative ground robots.

References

[1] S. A. S. Mohamed, M.-H. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila, "A survey on odometry for autonomous navigation systems," *IEEE Access*, vol. 7, pp. 97 466–97 486, 2019.

[2] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robotics and Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[3] M. Valente, C. Joly, and A. de La Fortelle, *Deep sensor fusion for real-time odometry estimation*, 2019.

[4] E. Vargas, R. Scona, J. S. Willners, *et al.*, "Robust underwater visual slam fusing acoustic sensing," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 2140–2146.

[5] L. Ojeda, D. Cruz, G. Reina, and J. Borenstein, "Current-based slippage detection and odometry correction for mobile robots and planetary rovers," *IEEE Transactions on Robotics*, vol. 22, no. 2, pp. 366–378, 2006.

[6] R. Long, C. Rauch, T. Zhang, V. Ivan, and S. Vijayakumar, "RigidFusion: Robot localisation and mapping in environments with large dynamic rigid objects," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3703–3710, 2021.

[7] H. Zhan, C. S. Weerasekera, J.-W. Bian, R. Garg, and I. Reid, *Df-vo: What should be learnt for visual odometry?* 2021.

[8] M. R. Walter, M. Antone, E. Chuangsuwanich, *et al.*, "A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments," *Journal of Field Robotics*, vol. 32, no. 4, pp. 590–628, 2015.

[9] G. Boucher and L. M. Sanchez, "Mobile wheeled robot with step climbing capabilities," in *Mobile Robots*, X. Chen, Y. Chen, and J. Chase, Eds., Rijeka: IntechOpen, 2009, ch. 3.

[10] M. Wada, "Studies on 4wd mobile robots climbing up a step," in *2006 IEEE International Conference on Robotics and Biomimetics*, 2006, pp. 1529–1534.

[11] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, "How do i sound like? forward models for robot ego-noise prediction," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2016, pp. 246–251.

[12] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," Institute of Electrical and Electronics Engineers, 2018.

[13] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 9, pp. 1484–1498, 2018.

[14] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro, "Robust sound source mapping using three-layered selective audio rays for mobile robots," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 2771–2777.

[15] B. F. Allen, F. Picon, S. Dalibard, N. Magnenat-Thalmann, and D. Thalmann, "Localizing a mobile robot with intrinsic noise," *3DTV-CON 2012*, vol. hal-00732764, 2012.

[16] 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.

[17] G. Narang, K. Nakamura, and K. Nakadai, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual slam," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 4021–4026.

[18] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filterbank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[19] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[20] K. Fukushima, "Neural network model for a mechanism of pattern recognition unaffected by shift in position - noecognitron," *Trans. IECE*, vol. A-(10), no. J62, pp. 658–665, 1979.

[21] L. Ding, H. Gao, Z. Deng, K. Yoshida, and K. Nagatani, "Slip ratio for lugged wheel of planetary rover in deformable soil: Definition and estimation," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 3343–3348.

[22] L. Li and H.-t. Lin, "Ordinal regression by extended binary classification," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19, MIT Press, 2006.

[23] D. Prokhorov, D. Zhukov, O. Barinova, A. Vorontsova, and A. Konushin, *Measuring robustness of visual slam*, 2019.

[24] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, p. 9, 2016.