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 paves the way towards inexpensive robust robot localization by proposing a system capable of estimating the longitudinal velocity of a wheeled robot on loose sandy terrain using only acoustic sensors. Which will 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. Experiments in a single wheel test bed with a wheel driving up to speeds of 0.07 m/s with a wide range of wheel slippage, show that the system is a feasible auxiliary source of odometry with 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 their combination would enhance their robustness in unstructured environments. The system is able to make a prediction from a single audio frame with a duration of 15ms in only 2ms on a user-level commercially available CPU. Additional experiments with white Gaussian noise show that high noise power (Signal to Noise Ratio of -10 dB) only affects significantly the prediction of speeds close to 0ms.**

***Index Terms*—Mobile robots, Wheels, Soil, Slippage detection, Acoustic Odometry**

I. Introduction

A. Problem

formulate problem mathematically

Robots are machines that resemble living creatures in being capable of moving independently (as by walking or rolling on wheels) and performing complex actions (such as grasping and moving objects) [13]. In accomplishing a defined mission, a robot physically interacts with its operating environment. Robot operating environments can be classified into predefined, semi-structured, and unstructured [11]. In an unstructured environment, the robot has no prior knowledge about it and has to rely on its sensory and navigation systems to operate autonomously. Hence the problem of robot localization is defined as the process of determining where a mobile robot is located in its environment [20]. 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.

Wheels or tracks still form the basis for robot locomotion, although strides have been made into exotic forms of legged robots [10]. Wheeled mobile systems are useful for practical applications compared with legged systems because of the simplicity of the mechanisms and control systems and efficiency in energy consumption [7]. 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. Which is the case for planetary missions [8], construction site applications, and agricultural robots, among others.

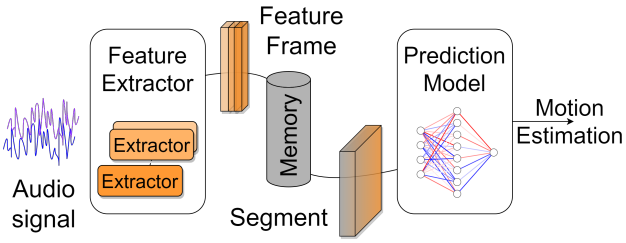


Figure 1: Proposed system

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 [28]. 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 [14]. 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. They have their 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.

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 [30, 36, 6] and to make better use of the available sensor information using deep learning and other computationally expensive

algorithms [35, 37]. Therefore, robot localization robustness has a price. Whether it is the explicit price of extra sensors or the implicit cost of extra computational resources.

B. Motivation

This work intends to minimize the price of robust ground robot localization in unstructured environments by using sensors that are relatively inexpensive, imply a low overhead, and might be needed for other functionalities of the robot: acoustic odometry. Many ground mobile robots will be equipped with sound sensors for the purpose of human-robot interaction [18], at the same time, audio signals have been extensively studied, resulting in very efficient methods for their processing.

Increasing the robustness of robot localization in unstructured environments without significantly increasing its price would make this technology more accessible. Which complies with the 9th and 10th Sustainable Development Goals: Promotes innovation without increasing the inequalities between small and big producers [40].

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 seen as a nuisance. Moreover, robot-terrain interaction sound has been used in Terrain Classification, which indicates that it carries significant information for environment understanding.

a) Audio-based Odometry: In [21], 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, [26] 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 claim that their work demonstrates an absolute error lower than 0.07 m/s and 0.02 rad/s and conclude that audio-based odometry systems should be useful auxiliary sources of odometry on the side of more traditional systems. However, the number of experiments and their evaluation is limited.

b) Sound Source localization: In [25] and [22], 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. [19] restricts the application to an in-pipe robot with a combination of orientation estimates from an inertial measurement unit and traversed distance estimations achieving as well both, self-localization and mapping of the pipeline. The distance is estimated using the time of flight of a reference sound generated with a loudspeaker at the entrance of the pipeline, which is measured with an onboard microphone. Alternatively, [15] perceives the robot's intrinsic noise to localize it using external audio sensors. Combinations with other self-localization methods are proposed in [38], where onboard sound sensors identify and remove the effect of dynamic obstacles for Visual SLAM, and [16], which localizes sounds using onboard microphones and uses them as navigation goals while using Visual SLAM for self-localization.

c) Terrain Classification: Multiple works propose to identify the terrain type of a robot's environment using onboard audio sensors. [27] proposes a deep learning framework, based on a convolutional neural network, that uses only sound from vehicle-terrain interactions to classify a wide range of indoor and outdoor terrains. This method is extended in [31], where an unsupervised classifier that learns from vehicle-terrain interaction sounds supervises a pixel-wise semantic image classifier. Similarly, [34] proposes a multi-modal self-supervised learning technique that switches between audio and image features to cluster terrain types. Extended as well by [33] using a multi-modal variational autoencoder and a Gaussian mixture model clustering algorithm on audio-visual data. It proposes as well to use gammatone-based filtering methods to extract audio features like in [26].

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

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 number of samples per frame can be tied together with the sample rate of the audio signal to define frames in terms of time. Frame duration is one of the parameters of the dataset as well as the number of features extracted per frame. **??** lists all the different parameters that define a dataset.

The term *extractor* will be used to abstract the system used to extract features from audio data. Although in this work only one type of extractor has been implemented and tested. Additional feature extractors, like the traditional Mel-frequency Cepstrum Coefficients (MFCC) or a simple Short-Time Fourier Transform, can be easily implemented as the

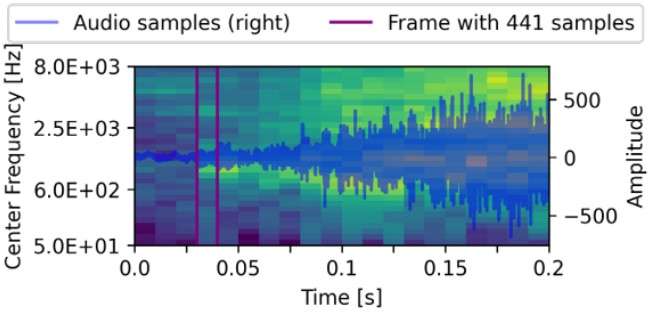


Figure 2: Audio signal (in blue) over its features extracted from frames with a duration of 10 ms with a single Gammatone filterbanks extractor.

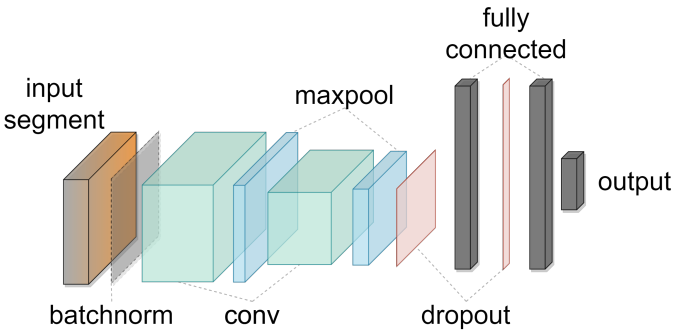


Figure 3: Convolutional Neural Network Architecture used in this work. Different layer sizes are used as well as different input segment dimensions. The output of the last fully connected layer changes with the model Task too.

Feature extraction module is decoupled from the Prediction Model.

This work opts to use a feature extractor based on Gammatone filterbanks since the performance of classification systems relying on MFCCs is greatly reduced in the presence of noise [26].

Gammatone filterbanks: An approximation to the human cochlear frequency selectivity originally introduced in [2]. Time-independent features are obtained by filtering the audio waveform with a bank of gammatone band-pass filters. The impulse response of a gammatone filter centered at frequency f_c is given by equation (1), where n indicates the order of the filter which largely determines the slope of the filter's skirts; and b is the bandwidth of the filter and largely determines the duration of the impulse response; a is the amplitude and ϕ is the phase.

$$g(t, f_c) = at^{n-1}e^{-2\pi bt}\cos 2\pi f_c t + \phi \tag{1}$$

This work implements a bank of fourth order gammatone filters with its corresponding bandwidth b of 1.019 ERB where ERB is the equivalent rectangular bandwidth scale [3]. It is based on a Matlab MEX function implemented in C by Ma et al. [9], which at the same time is based on Martin Cooke's Ph.D work [4]. The filters are distributed over a predefined frequency range linearly 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. Segments

This work proposes a shallow Convolutional Neural Network [1] 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. Different layer sizes are evaluated as well, defined in **??**. The output of the last fully connected layer depends on the Task.

1) Task: One can find here a description of the different tasks implemented and tested in models. Tasks define the goal of the model and the way its loss is computed.

a) Classification: Consists in classifying the longitudinal velocity given a set of possibilities. The different classes are ranges of longitudinal velocities, being these ranges a hyperparameter of the model. Cross entropy loss is computed between the predicted class probabilities and the class corresponding to the target longitudinal velocity and the predicted class. The output of the model is therefore a vector of probabilities corresponding to each class.

b) Ordinal classification: Consists in classifying the longitudinal velocity given a set of possibilities like in Classification, with different classes being ranges of longitudinal velocities. But the order of the class matters. This method was introduced in [5], where standard classification algorithms are extended to make use of the order of the classes. The output of this model is a vector of binary values 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.

2) Architecture: This section describes the different model architectures implemented and evaluated. A common point of them all is simplicity. It is out of the scope of this work to find an optimal architecture for acoustic odometry. But it is interesting to evaluate different simple options.

a) CNN with normalized input: This architecture is identical to the **??** except for the fact that it contains a batch normalization layer [17] as shown in figure 3.

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. As this scenario can be easily reproduced with

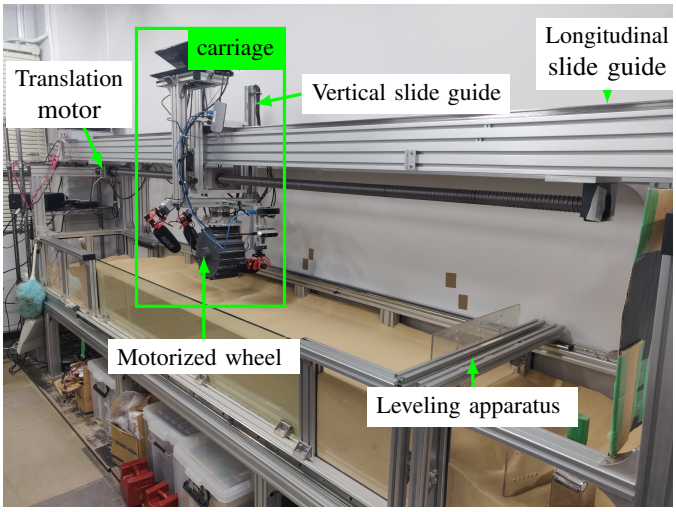
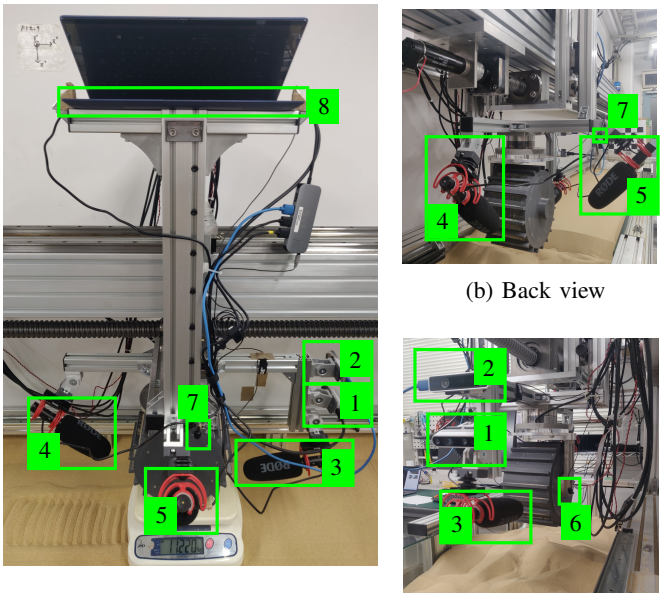


Figure 4: Single wheel testbed



(a) Lateral view

(b) Back view

(c) Front view

Figure 5: Experimental Setup: *a)* shows a lateral view of the wheel testbed carriage while being weighted. *b)* shows a back view of the wheel and sensor setup while *c)* shows a front view of it.

Index	Device
1	Intel RealSense™ Depth Camera D435i
2	Intel RealSense™ Tracking Camera T265
3	RØDE VideoMic™ NTG front
4	RØDE VideoMic™ NTG back
5	RØDE VideoMic™ NTG top
6	RØDE SmartLav+ wheel axis
7	RØDE SmartLav+ top
8	HP Elite Dragonfly built in microphone array

Table I: Devices installed in the carriage of the single wheel testbed

the available Experimental Setup, making it possible to gather a whole new Datasets. Experiments are conducted varying the different tuneable parameters in the Feature extraction module and Prediction Model and a model is selected based on its performance and computational cost.

A. Experimental Setup

This work uses a single wheel testbed, composed of a carriage unit attached to a frame over a soil bin, as it can be seen 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 the 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 by varying the angular speed of the wheel and the longitudinal velocity of the unit.

The slip ratio s of a smooth wheel is defined as the function of the horizontal velocity V_x and wheel angular velocity V_ω found in equation (2) [12], where r is the radius of the wheel. The slip ratio is bounded in the range $(-1, 1]$, where negative values correspond to wheel *skidding*.

$$s = \begin{cases} 1 - \frac{V_x}{rV_\omega} & (rV_\omega \geq V_x, 0 \leq s \leq 1) \\ \frac{rV_\omega}{V_x} - 1 & (rV_\omega < V_x, -1 \leq s < 0) \end{cases} = \frac{rV_\omega - V_x}{\max(rV_\omega, V_x)} \quad (2)$$

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.

The carriage unit is equipped with a total of six microphones and two cameras, listed in table I. Their positioning is shown in figure 5 and it takes into account ideal positions of a microphone attempting to capture the sound generated by the wheel interaction with the terrain (devices 4 and 3), feasible positions of a microphone in a mobile robot (devices 5, 6, and 7), and a reasonable position of a microphone dedicated to human-robot interaction (device 8).

B. Datasets

1) Training: used to train and validate the Prediction Model. This work defines it as a collection of annotated overlapping segments.

The annotated variables are the longitudinal velocity V_x, the wheel angular speed V_ω and the slip ratio s. However,

	Mean ATE [m]	Mean RPE [m/s]
Acoustic Odometry	$1.02e-03$	$5.10e-03$

Table II: Selected model average performance across all evaluation recordings and all devices. Average Trajectory Error (ATE) is computed between frames with a duration of 15 ms and Relative Pose Error (RPE) is computed using time windows 1s long.

as it has been explained in section II-B, these segments are relatively large in terms of time. In a way that the measured variables can oscillate significantly during the segment. This work annotates the segments with a weighted average of the measurements within its duration.

Segments are collected from a set of training recordings captured under different driving conditions, which are defined by the combination of the following controlled variables: *slip ratio* as defined in equation (2). Ranging between -0.3 and 0.6. Being free or uncontrolled slip an additional option; *load* measured in kg added to the base carriage weight (which is 11.2 kg) taking values of 0 kg, 5 kg and 10 kg; *wheel angular velocity* ranging from 5 deg/s to 30 deg/s in steps of 5 deg/s; *contact*, whether the wheel is making contact with the ground or is suspended in the air. A total of 168 different recordings that make a total of 1 hour for each of the 8 recording devices where captured for the training dataset.

The collected segments are artificially augmented using two methods: Add a random gain in the range [-10, 10]; Add white noise with a random signal-to-noise ratio in the range [0, 1] in the decibel scale.

2) Evaluation: used to evaluate the performance of the system under more realistic and challenging conditions unseen during training.

C. Selected model

This work selects a model trained on a dataset of segments of 50 frames, each of them spans over 15 ms and has 64 features from a single Gammatone filterbanks extractor. The extractor is applied on the average of the audio signal channels with a frequency range of [50, 80000] Hz and features on te Bel scale. Training data from all available recordings is selected using the `with-laptop` strategy split described in **??**, a Ordinal classification task with 28 different linearly distributed longitudinal velocity ranges and a CNN with normalized input architecture with a small size described as `S` in **??**.

IV. Results

Table II shows the average performance of the selected model across all seven evaluation recordings and devices. Performance is measured with commonly used odometry and simultaneous localization and mapping metrics as described in [29]. Additionally APE Figure 6 shows the its results on an evaluation recording characterized by high slippage driving conditions using two different devices.

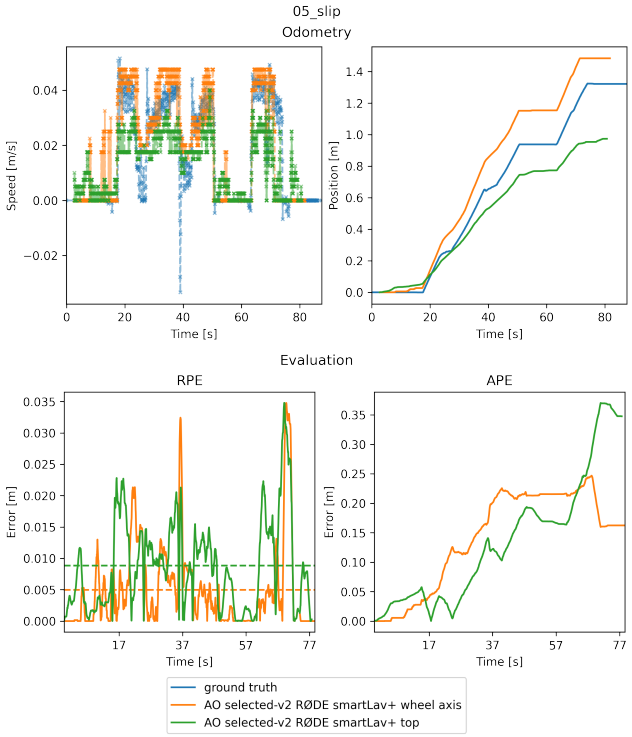


Figure 6: Selected model evaluated on a recording characterized by high slippage conditions using two different devices.WIP

A. Noise

One can see in figure 7 the behavior of the selected model against white noise. Generally, white noise increases the value of the estimated motion. Signal-to-Noise ratios of -20 dB impede the model to recognize situations where the wheel is not moving.

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 ground truth wheel angular speed; Visual SLAM using Intel® RealSense™ Tracking Camera T265, which is based on visual odometry. Figure 8 is the evaluation in an undisturbed scenario for all methods, no acoustic noise and no lightning changes or dynamic objects. It shows that the selected model can perform

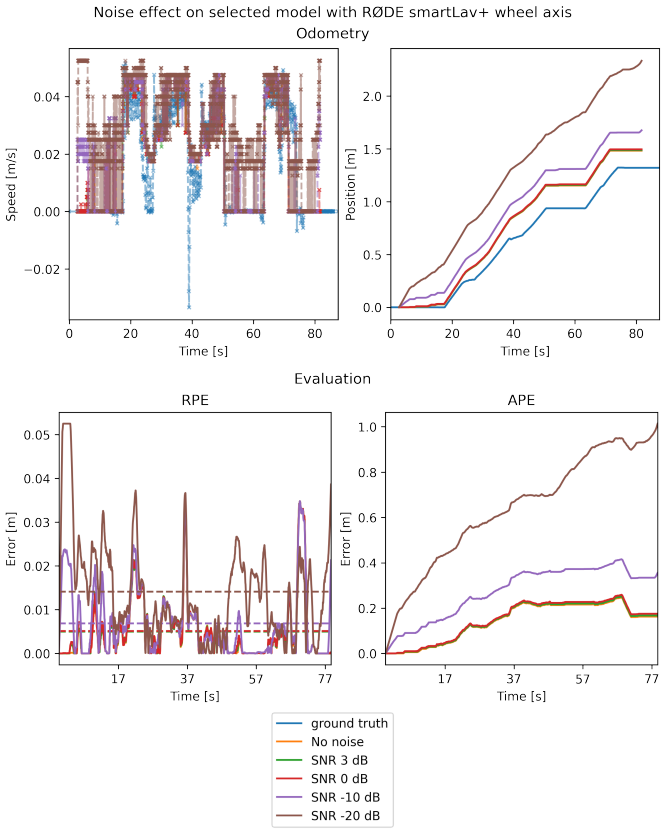


Figure 7: Selected model evaluated on a recording with added white noise with varying signal-to-noise Ratio values.WIP

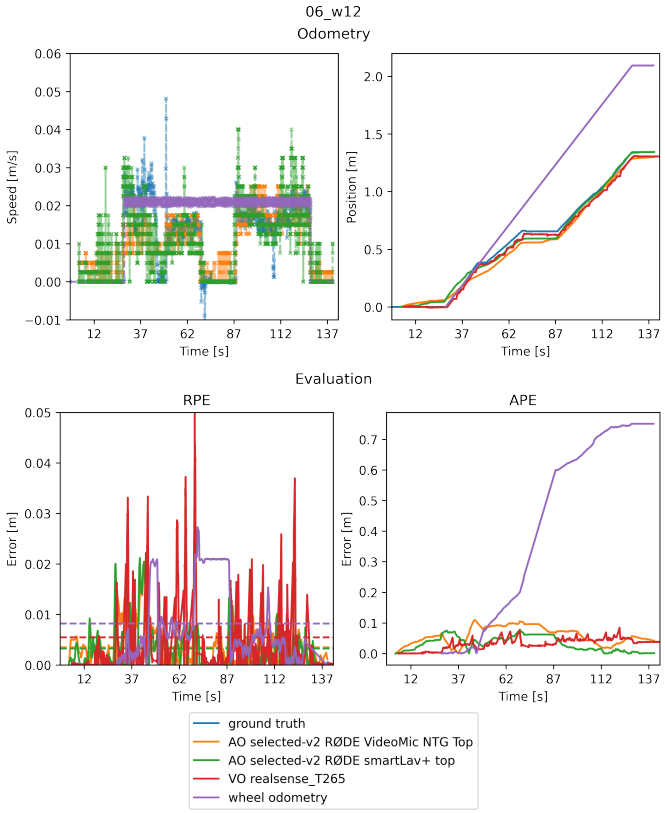


Figure 8: Selected model evaluated evaluated against Visual SLAM (based on visual odometry) and wheel odometry.

with an accuracy comparable to state-of-the-art commercial visual-based methods in this simplified scenario. Figure 9 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. Which makes Acoustic Odometry more accurate in certain scenarios where the visual odometry vulnerabilities are exploited.

D. Discussion

The proposed Acoustic Odometry system proves to be a viable auxiliary source of odometry for wheeled robots in loose sandy terrain. Figure 8 and figure 9 show that audio-based odometry is more accurate than wheel odometry, and its accuracy can be comparable to state-of-the-art commercially available visual-based methods and that it can be more accurate than visual odometry in the presence of dynamic objects and lightning.

This work only evaluates the performance of a shallow convolutional neural network with different layer sizes. But other architectures be more suitable for the task. Specially transformers [24] which have been used in speech recognition [39] and visual odometry tasks [32]. Moreover, neural networks may not even be needed, a simpler algorithm may be used with sufficient accuracy and lower computational cost.

The selected model behavior changes from one device to another. Devices that output an audio signal with a higher power tend to result in higher speed predictions, as can be seen in figure 6, while the overall predicted curve is very similar. Meaning that even if the absolute value of the predicted velocity is not accurate, the model is able to recognize slippage conditions across different devices. Even with devices not present during training, as is the case in figure 8, where none of the devices shown in the evaluation were present during training. This indicates that fine-tuning [23] a model with the device where it will be deployed might significantly increase its performance. Similarly, using wheel odometry to estimate the wheel angular speed while using the proposed system to identify the wheel slippage may improve the performance as well.

White noise affects the performance of the proposed system, as could be expected. Nevertheless, a considerably large noise power compared to the signal power does not make the selected model unusable. As it can be seen in figure 7, where a noise with a power 10 times the one of the audio signal (SNR -10 dB) only significantly affects the predictions on speeds close to 0. This indicates that white noise only significantly affects the prediction of speeds under a threshold determined by the signal-to-noise power ratio.

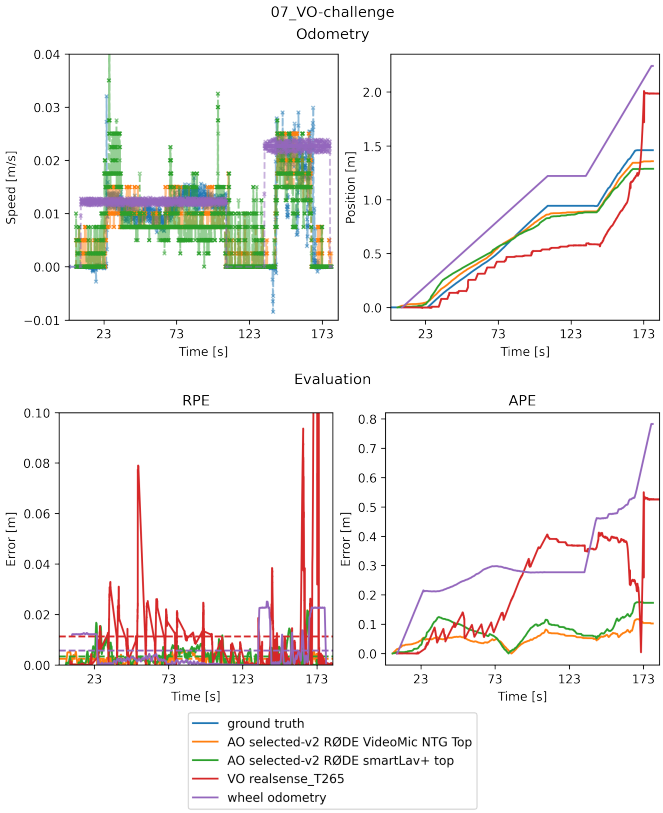


Figure 9: Selected model evaluated against Visual SLAM (based on visual odometry) in a scenario where visual odometry vulnerabilities are exploited: Dynamic objects are moved in the camera field of view and lightning conditions are changed during the recording.

Finally, this work can be compared with other odometry methods. On one hand, only [26] proposes an audio-based odometry method, claiming an average absolute error of 0.065 m/s when predicting velocities, which is comparable to the average ATE 0.001m accumulated over the 66 15ms long audio frames contained in one second. But this work also shows that the actual average RPE in a 1 second window is 0.005m. However, it remains unknown the performance of the proposed system on rotational movement and under the presence of multiple moving wheels. This comparison should also be taken with a grain of salt as no common benchmarking data exists.

On the other hand, the method presented in [6], where wheel slippage is identified using the motor current and compensated in the wheel odometry computation, still 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 devices and evaluation recordings. On the other hand, 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 and **??** discusses several improvements that could be applied to the proposed system in order to improve its performance.

V. Conclusions

This work demonstrates that it is feasible to estimate odometry from acoustic data with a computationally inexpensive system. Many future ground mobile robots will be equipped with audio sensors for human-robot interaction purposes. In that context, robot ego-noise is just noise. Being able to estimate motion using byproduct data from the human-robot interaction system can add robustness to the localization system without incurring in significant costs.

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 successfully identifies when the wheel slips. The proposed model is also evaluated in the presence of white noise and demonstrates that it still can successfully predict longitudinal velocities in the presence of high noise power. Moreover, the system is compared against wheel odometry and a commercially available visual simultaneous localization and mapping system satisfactorily.

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] K. Fukushima. “Neural network model for a mechanism of pattern recognition unaffected by shift in position - Noecognitron”. In: *Trans. IECE* A-(10).J62 (1979), pp. 658–665.

[2] J. Holdsworth et al. “Implementing a gammatone filterbank”. In: *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank* 1 (1988), pp. 1–5.

[3] Brian R Glasberg and Brian C.J Moore. “Derivation of auditory filter shapes from notched-noise data”. In: *Hearing Research* 47.1 (1990), pp. 103–138. ISSN: 0378-5955. DOI: https://doi.org/10.1016/0378-5955(90)90170-T.

[4] Martin Cooke. “Modelling auditory processing and organisation”. In: *Distinguished dissertations in computer science*. 1993.

[5] Ling Li and Hsuan-tien Lin. “Ordinal Regression by Extended Binary Classification”. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, 2006.

[6] L. Ojeda et al. “Current-Based Slippage Detection and Odometry Correction for Mobile Robots and Planetary Rovers”. In: *IEEE Transactions on Robotics* 22.2 (2006), pp. 366–378. DOI: 10.1109/TRO.2005.862480.

[7] Masayoshi Wada. “Studies on 4WD Mobile Robots Climbing Up a Step”. In: *2006 IEEE International Conference on Robotics and Biomimetics*. 2006, pp. 1529–1534. DOI: 10.1109/ROBIO.2006.340156.

[8] Faiz Ben Amar et al. “Towards an advanced mobility of wheeled robots on difficult terrain”. In: *International Journal of Factory Automation, Robotics and Soft Computing* hal-03135894 (2007), pp. 40–45.

[9] Ning Ma et al. “Exploiting correlogram structure for robust speech recognition with multiple speech sources”. In: *Speech Communication* 49 (Dec. 2007), pp. 874–891. DOI: 10.1016/j.specom.2007.05.003.

[10] Gary Boucher and Luz Maria Sanchez. “Mobile Wheeled Robot with Step Climbing Capabilities”. In: *Mobile Robots*. Ed. by XiaoQi Chen, Y.Q. Chen, and J.G. Chase. Rijeka: IntechOpen, 2009. Chap. 3. DOI: 10.5772/6988. URL: https://doi.org/10.5772/6988.

[11] X.Q. Chen, Y.Q. Chen, and J.G. Chase. “Mobiles Robots - Past Present and Future”. In: *Mobile Robots*. Ed. by XiaoQi Chen, Y.Q. Chen, and J.G. Chase. Rijeka: IntechOpen, 2009. Chap. 1. DOI: 10.5772/6986. URL: https://doi.org/10.5772/6986.

[12] Liang Ding et al. “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. DOI: 10.1109/IROS.2009.5354565.

[13] Merriam-Webster.com. *robot*. 2011. URL: https://www.merriam-webster.com/dictionary/robot (visited on 07/25/2022).

[14] Davide Scaramuzza and Friedrich Fraundorfer. “Visual Odometry [Tutorial]”. In: *IEEE Robotics and Automation Magazine* 18.4 (2011), pp. 80–92. DOI: 10.1109/MRA.2011.943233.

[15] Brian F. Allen et al. “Localizing a mobile robot with intrinsic noise”. In: *3DTV-CON 2012* hal-00732764 (Oct. 2012).

[16] Gautam Narang, Keisuke Nakamura, and Kazuhiro 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. DOI: 10.1109/SMC.2014.6974560.

[17] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. DOI: 10.48550/ARXIV.1502.03167. URL: https://arxiv.org/abs/1502.03167.

[18] Matthew R. Walter et al. “A Situationally Aware Voice-commandable Robotic Forklift Working Alongside People in Unstructured Outdoor Environments”. In: *Journal of Field Robotics* 32.4 (2015), pp. 590–628. DOI: https://doi.org/10.1002/rob.21539. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21539. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21539.

[19] Yoshiaki Bando et al. “Sound-based online localization for an in-pipe snake robot”. In: *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. 2016, pp. 207–213. DOI: 10.1109/SSRR.2016.7784300.

[20] Shoudong Huang and Gamini Dissanayake. “Robot Localization: An Introduction”. In: *Wiley Encyclopedia of Electrical and Electronics Engineering*. John Wiley and Sons, Ltd, 2016, pp. 1–10. ISBN: 9780471346081. DOI: https://doi.org/10.1002/047134608X.W8318. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/047134608X.W8318.

[21] Antonio Pico et al. “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. DOI: 10.1109/DEVLRN.2016.7846826.

[22] Daobilige Su et al. “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. DOI: 10.1109/IROS.2016.7759430.

[23] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3 (May 2016), p. 9. DOI: 10.1186/s40537-016-0043-6.

[24] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[25] Christine Evers and Patrick A. Naylor. “Acoustic SLAM”. In: *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 26.9 (Sept. 2018), pp. 1484–1498. ISSN: 2329-9290. DOI: 10.1109/TASLP.2018.2828321. URL: https://doi.org/10.1109/TASLP.2018.2828321.

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

[27] Abhinav Valada, Luciano Spinello, and Wolfram Burgard. “Deep Feature Learning for Acoustics-Based Terrain Classification”. In: *Robotics Research: Volume 2*. Ed. by Antonio Bicchi and Wolfram Burgard. Cham: Springer International Publishing, 2018, pp. 21–37. ISBN: 978-3-319-60916-4. DOI: 10.1007/978-3-319-60916-4_2. URL: https://doi.org/10.1007/978-3-319-60916-4_2.

[28] Sherif A. S. Mohamed et al. “A Survey on Odometry for Autonomous Navigation Systems”. In: *IEEE Access*

7 (2019), pp. 97466–97486. DOI: 10.1109/ACCESS.2019.2929133.

[29] David Prokhorov et al. *Measuring robustness of Visual SLAM*. 2019. DOI: 10.48550/ARXIV.1910.04755. URL: https://arxiv.org/abs/1910.04755.

[30] Michelle Valente, Cyril Joly, and Arnaud de La Fortelle. *Deep Sensor Fusion for Real-Time Odometry Estimation*. 2019. DOI: 10.48550/ARXIV.1908.00524. URL: https://arxiv.org/abs/1908.00524.

[31] Jannik Zürn, Wolfram Burgard, and Abhinav Valada. "Self-Supervised Visual Terrain Classification from Unsupervised Acoustic Feature Learning". In: (2019). DOI: 10.48550/ARXIV.1912.03227. URL: https://arxiv.org/abs/1912.03227.

[32] Xiangyu Li et al. *Transformer Guided Geometry Model for Flow-Based Unsupervised Visual Odometry*. 2020. arXiv: 2101.02143 [cs.CV].

[33] Reina Ishikawa et al. "Single-modal Incremental Terrain Clustering from Self-Supervised Audio-Visual Feature Learning". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 9399–9406. DOI: 10.1109/ICPR48806.2021.9412638.

[34] Akiyoshi Kurobe et al. "Audio-Visual Self-Supervised Terrain Type Recognition for Ground Mobile Platforms". In: *IEEE Access* 9 (2021), pp. 29970–29979. DOI: 10.1109/ACCESS.2021.3059620.

[35] Ran Long et al. "RigidFusion: Robot Localisation and Mapping in Environments With Large Dynamic Rigid Objects". In: *IEEE Robotics and Automation Letters* 6.2 (Apr. 2021), pp. 3703–3710. DOI: 10.1109/lra.2021.3066375. URL: https://doi.org/10.1109%2Flra.2021.3066375.

[36] Elizabeth Vargas et al. "Robust Underwater Visual SLAM Fusing Acoustic Sensing". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021, pp. 2140–2146. DOI: 10.1109/ICRA48506.2021.9561537.

[37] Huangying Zhan et al. *DF-VO: What Should Be Learnt for Visual Odometry?* 2021. DOI: 10.48550/ARXIV.2103.00933. URL: https://arxiv.org/abs/2103.00933.

[38] Tianwei Zhang et al. *AcousticFusion: Fusing Sound Source Localization to Visual SLAM in Dynamic Environments*. 2021. DOI: 10.48550/ARXIV.2108.01246. URL: https://arxiv.org/abs/2108.01246.

[39] Sehoon Kim et al. *Squeezeformer: An Efficient Transformer for Automatic Speech Recognition*. 2022. arXiv: 2206.00888 [eess.AS].

[40] United Nations. *Sustainable Development Goals*. URL: https://sdgs.un.org/goals (visited on 07/25/2022).