HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 1–136 (2020)
hjic.mk.uni-pannon.hu

Contents

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 1–1 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-01

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
HJIC

EDITORIAL PREFACE TO THE SPECIAL ISSUE DEDICATED TO THE WORKSHOP “SELF-DRIVING VEHICLES"

ISTVÁN SZALAI[1]

[1] Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk utca 18/A,
Zalaegerszeg H-8900, HUNGARY

In the 22nd November 2019 a Workshop was held in the Automotive Proving Ground, Zalaegerszeg, Hungary. The title of Workshop was "Self-Driving Vehicles. Sensors, Algorithms, Intelligent Materials." The Workshop collected experts from the University of Pannonia, Veszprém, and the Széchenyi István University, Győr. The workshop was supported by EFOP-3.6.2-16-2017-00002.

The lectures were mostly presented by talented graduate and postgraduate students, who represent future’s professionals. They will be graduated from the two universities and hopefully will seek career in the automotive industry of Hungary, especially, in the industry of slef-driving vehicles. This area of technology is developing fast and strongly supported by governments and various players of the economy due to its potential to drive the development of state-of-the-art technologies. The workshop showed that the research groups working in these two universities can contribute to this development with novel results.

This issue of the Hungarian Journal of Industry and Chemistry contains the contributions of the lectures.

István Szalai
Professional Leader
”Research of autonomous vehicle systems related to the autonomous test track in Zalaegerszeg“ (EFOP-3.6.2-16-2017-00002)

SZÉCHENYI 2020
HUNGARIAN GOVERNMENT
European Union
European Social Fund
INVESTING IN YOUR FUTURE

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 3–10 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-02

IMPROVING THE EFFICIENCY OF NEURAL NETWORKS WITH VIRTUAL TRAINING DATA

János Hollósi*[1,2], Rudolf Krecht[2], Norbert Markó[2], and Áron Ballagi[2,3]

[1]Department of Information Technology, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY
[2]Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY
[3]Department of Automation, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY

At Széchenyi István University, an autonomous racing car for the Shell Eco-marathon is being developed. One of the main tasks is to create a neural network which segments the road surface, protective barriers and other components of the racing track. The difficulty with this task is that no suitable dataset for special objects, e.g. protective barriers, exists. Only a dataset limited in terms of its size is available, therefore, computer-generated virtual images from a virtual city environment are used to expand this dataset. In this work, the effect of computer-generated virtual images on the efficiency of different neural network architectures is examined. In the training process, real images and computer-generated virtual images are mixed in several ways. Subsequently, three different neural network architectures for road surfaces and the detection of protective barriers are trained. Past experiences determine how to mix datasets and how they can improve efficiency.

Keywords: neural network, virtual training data, autonomous vehicle

1. Introduction

Shell Eco-marathon is a unique international competition held by Royal Dutch Shell Plc. This event challenges university students to design, develop, build and drive the most energy-efficient racing cars. Our University's racing team, the SZEnergy Team, has been a successful participant in the Shell Eco-marathon for over 10 years. Two years ago, Shell introduced the Autonomous Urban-Concept (AUC) challenge, which is a separate competition for self-driving vehicles that participate in the Shell Eco-marathon. Participants in the AUC challenge have to complete five different tasks, e.g. parking in a dedicated parking rectangle, obstacle avoidance on a straight track, drive one lap of the track autonomously, etc.

Our long-term goal is to prepare for the AUC challenge. One of the main tasks is to create an intelligent system, which perceives the environment of our racing car, e.g. other vehicles, the road surface, other components of the racing track, etc. In this paper, only the segmentation of the road surface and of the protective barriers is taken into consideration. An approach based on neural networks will determine the segmentation, because such networks are one of the best tools to solve problems concerning visual information-based detection and segmentation, e.g. image segmentation. Many high-performance neural network architectures are available such as AlexNet by Krizhevsky et al. [1], VGGNet by Simonyan and Zisserman [2], GoogLeNet by Szegedy et al. [3], Fully Convolutional Networks by Shelhamer et al. [4], U-Net by Ronneberger et al. [5], ResNet by He et al. [6] and Pyramid Scene Parsing Network by Zhao et al. [7]. Training neural networks requires a large amount of training data. However, in this case, the number of training samples is insufficient, e.g. no training images of protective barriers are available and the generation and annotation of real world data is labour-intensive and time-consuming. Computer simulation environments will be used to generate training data for this task. Some attempts that apply virtually generated data to train neural networks have been made. Peng et al. [8] demonstrated CAD model-based convolutional neural network training for joint object detection. Tian et al. [9] presented a pipeline to construct virtual scenes and virtual datasets for neural networks. They proved that mixing virtual and real data to train neural networks for joint object detection helps to improve performance. Židek et al. [10] presented a new approach to joint object detection using neural networks trained by virtual model-based datasets. In this paper, an attempt is made to show the effects of computer-generated training data on the learning process of different network architectures.

The paper is structured as follows: in Section 2, the virtual simulation environment that is used for generating training data is described; in Section 3, our neural network architectures are presented; in Section 4, the training process of the networks is outlined; in Section 5, our

*Correspondence: hollosi.janos@sze.hu

results and experiences are shared; finally, in Section 6, our conclusions are stated.

2. Our virtual environment

Our aim is to create highly realistic image sets that depict racing tracks which follow the rulebook of the Shell Eco-marathon Autonomous UrbanConcept. In order to ensure repeatability and simple parameter setup, the creation of complete, textured 3D-models of the racing tracks is advised. These simulated environments can be used to create images with desired weather and lighting conditions by scanning the track environment using a camera moving at a predefined constant speed. The images created using this method can be processed further, e.g. segmentation and clustering of different types of objects such as the road surface, protective barriers and vegetation. Based on the characteristics of the predefined task, the requirements of the simulation environment can be enumerated:

- highly realistic appearance,
- easy use of textures,
- fast workflow,
- characteristics definable by parameters (parametric lights, weather conditions),
- modular environment construction,
- importability of external CAD models.

Unreal Engine 4 [11] is a games engine designed for the fast creation of modular simulated environments by the use of modular relief, vegetation and building elements. In these environments, actors based on external CAD models could be used. Fields of engineering that apply different visual sensors and cameras require very similar computer simulation technologies to the video game industry. Video games need to be highly realistic as well as efficient due to limited computational capacity. The requirements are the same for the simulation of vehicles mounted with cameras. Highly realistic computer simulations reduce the cost and duration of real-life tests and camera calibrations. It is also important to mention that by using technology implemented and/or developed by the video game industry, the support of a vast developer community is available.

Since our goal is to develop image-perceiving solutions for the Shell Eco-marathon Autonomous UrbanConcept challenge, it is important to carefully follow the rules of this competition with regard to the racing tracks. The simulated environments and racing tracks created by Unreal Engine strictly follow the rules defined by the aforementioned rulebook. These rules define that the self-driving vehicles have to compete on racing tracks equipped with protective barriers of a known height painted in alternating red and white segments. It is also defined that every racing track consists of three painted line markings, one that is green to denote the starting position, a yellow one to trigger the self-driving mode, and another that is red to mark the finish line. Because the racing tracks and tasks are well defined, it is crucial to create accurate models of the expected environments. Differences between real and simulated environments might lead to further developments in the wrong direction.

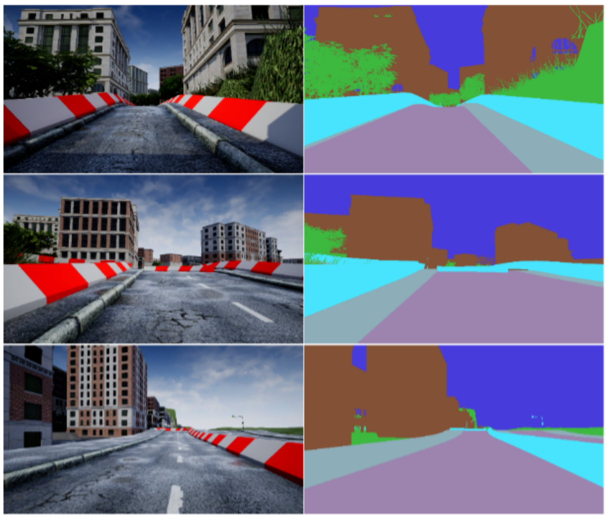
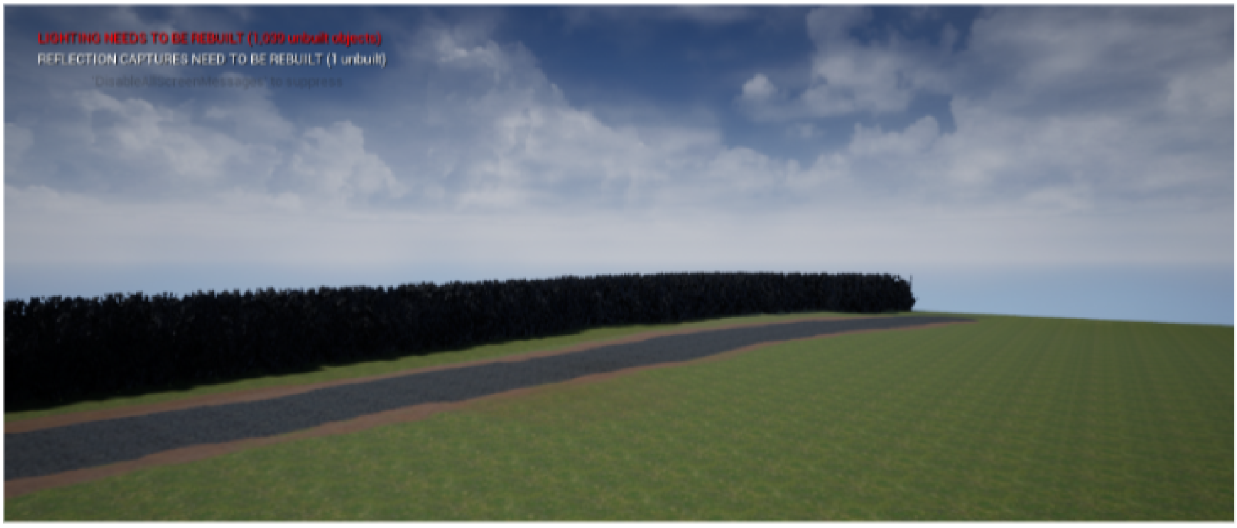


Figure 1: Example images from the training set.

Two simulated test environments were created. The first one was based on a readily available city model with streets corresponding to a typical racing track. Barrier elements were added to the roads to ensure that the racing track complies with the requirements outlined in the rulebook. This model includes defects in and textures of the road surface to ensure detection of the road surface is robust. In order to create image sets based on this environment model, a vehicle model equipped with a camera travelled around the racetrack on a pre-defined path. The camera was set to take pictures at pre-defined time intervals. The image set was annotated by using a module called AirSim. AirSim is an open-source, cross-platform simulator simulation platform built on Unreal Engine, but it also has a Unity release. This simulator module consists of a built-in Python-based API (Application Programming Interface) which was developed for image segmentation. By using this API, the necessary realistic and segmented image datasets were created. Some example pairs of images from our virtual dataset are presented in Fig. 1. In order to prepare for all the tasks defined in the rulebook, multiple models of racing tracks were created. All such models are based on the same environment model, which includes vegetation and the sky as shown in Fig. 2. The models of sections of racing track were realized according to the challenges defined in the rulebook. The CAD models representing elements of the racing track were custom-made to comply with the shapes, sizes and colors outlined in the rulebook. The sections of racing track generated can be used to simulate handling



LIGHTING NEEDS TO BE REBUILT (1,039 unbuilt objects)
REFLECTION CAPTURES NEED TO BE REBUILT (1 unbuilt)
'DisableAllScreenMessages' to suppress

Figure 2: Basic environment of racing tracks.

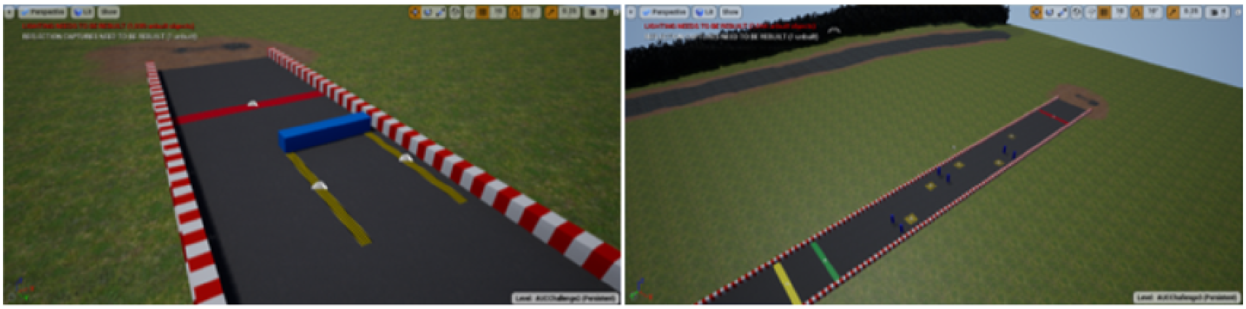


Figure 3: Parking place and slalom course.

(slaloming) and parking tasks. This virtual racing track is shown in Fig. 3. The image sets were created by a moving camera in the environment and segmentation was carried out by changing the textures.

3. Neural network architectures

Three different neural network architectures are implemented in this work: FCN, U-Net and PSPNet. All neural networks are designed for image segmentation, where the size of input images is $256 \times 512 \times 3$, and the size of output ones is $256 \times 512 \times 1$. Every network is trained for the segmentation of the road surface and protective barriers.

3.1 FCN

The Fully Convolutional Network (FCN) [4] architecture is based on fully convolutional layers, where the basic idea is to extend effective classification neural networks to conduct segmentation tasks. Our FCN architectures are shown in Fig. 4.

Let:

$$\gamma = (conv, bn, ReLu) \tag{1}$$

$$b_1 = (\gamma, \gamma, mp) \tag{2}$$

$$b_2 = (\gamma, \gamma, \gamma, mp) \tag{3}$$

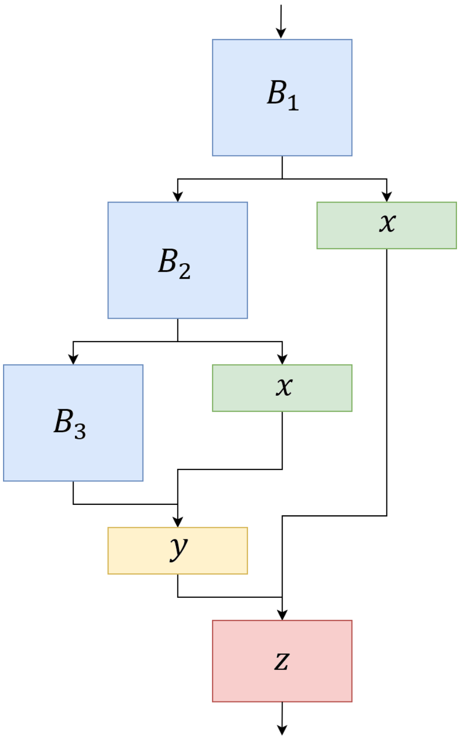


Figure 4: FCN architecture.

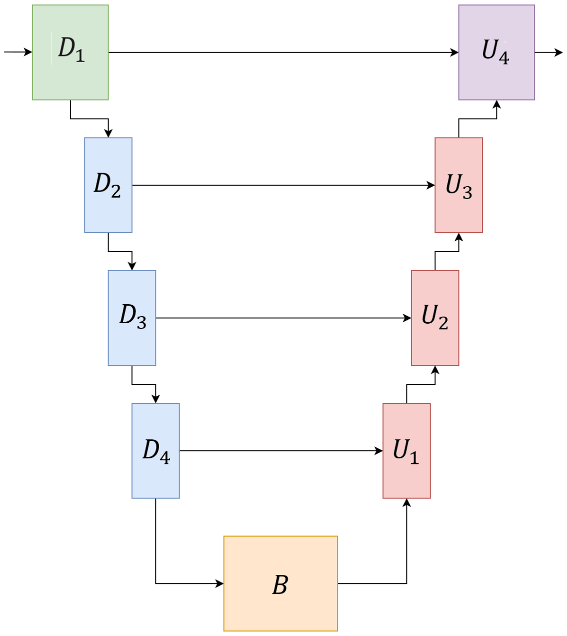


Figure 5: U-Net architecture.

where $conv$ denotes a convolutional layer, bn represents a batch normalization layer, $ReLu$ stands for a rectified linear activation unit and mp is a max pooling layer. Let:

$$B_1 = (b_1, b_1, b_2) \tag{4}$$
$$B_2 = (b_2) \tag{5}$$
$$B_3 = (b_2, \gamma, \gamma, \gamma) \tag{6}$$
$$x = (conv, bn) \tag{7}$$
$$y = (ReLu, softmax) \tag{8}$$
$$Z = (ReLu, softmax) \tag{9}$$

where $softmax$ denotes a softmax layer. In this implementation, the dimensions of all convolutional layers are 3×3, except for the three fully connected layers in block B_3. The dimensions of these convolutional layers are 7×7. In block B_1, the first two convolutional layers both contain 64 filters, the third and fourth both contain 128 filters, and the last three convolution blocks each contain 256 filters. The convolutional layers in block B_2 contain 512 filters in total. The first three convolutional layers in block B_3 contain 512 filters in total, and the fully connected layers are based on 4096 filters in total.

3.2 U-Net

The U-Net [5] neural network architecture was originally created for biomedical image segmentation. It is based on FCN, where the neural network can be divided into two main blocks, namely the downsampling and upsampling blocks. Our implementations are shown in Fig. 5. Let:

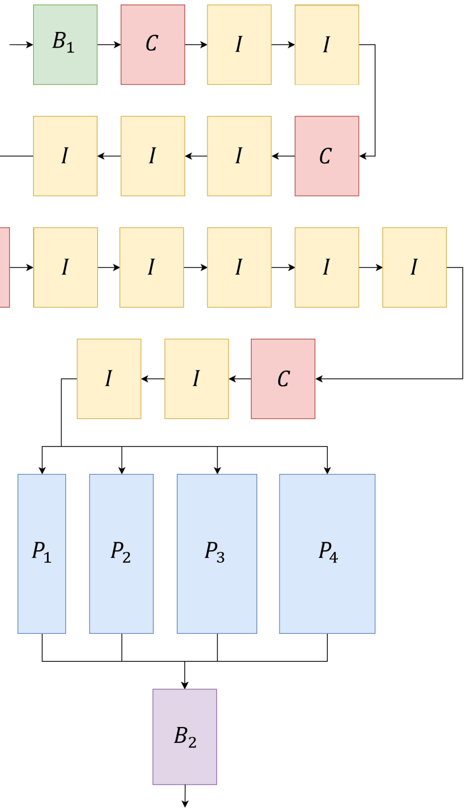


Figure 6: PSP Net architecture.

$$D_1 = D_2 = (\gamma, \gamma, maxpooling) \tag{10}$$
$$D_3 = D_4 = B = (\gamma, \gamma, \gamma, maxpooling) \tag{11}$$
$$U_1 = U_2 = U_3 = \left(conv^t, bn, ReLu, \gamma, \gamma\right) \tag{12}$$
$$U_4 = \left(conv^t, bn, ReLu, \gamma, \gamma, conv, softmax\right) \tag{13}$$

where $conv^t$ is a transposed convolution layer. In the U-Net neural network, the dimensions of all convolutions and transposed convolutions are 3×3, and 2×2, respectively. The number of convolutional filters are as follows: each convolutional layer in D_1 consists of 64, in D_2 of 128, in D_3 of 256 and in D_4 as well as B of 512 filters. The upsampling block is very similar. U_1 consists of 512, U_2 of 256, U_3 of 128 and U_4 of 64 filters.

3.3 PSPNet

The Pyramid Scene Parsing Network (PSPNet) [7] was judged to be the best architecture in the ImageNet Scene Parsing Challenge in 2016 [12]. The main building block of the PSPNet is a pyramid pooling module, where the network fuses features under four different pyramid scales. Our PSPNet-based architecture is shown in Fig. 6.

Let:

$$B_1 = (\gamma, \gamma, \gamma, maxpooling) \tag{14}$$
$$C = (\gamma, \gamma, conv, bn) + (conv, bn) \tag{15}$$
$$I = (\gamma, \gamma, conv, bn) \tag{16}$$
$$p = (avg, conv) \tag{17}$$
$$P_1 = (p) \tag{18}$$
$$P_2 = (p, p) \tag{19}$$
$$P_3 = (p, p, p) \tag{20}$$
$$P_4 = (p, p, p, p) \tag{21}$$
$$B_2 = \left(\gamma, dropout, conv, conv^t, softmax\right) \tag{22}$$

where avg denotes an average pooling layer and $dropout$ represents a dropping out unit. In block B_1, the dimensions of all convolutions are 3×3. In blocks C and I, the dimensions of every first & third and every second convolution are 3×3 and 1×1, respectively. In block B_2, the dimensions of the first convolution are 3×3 and the second 1×1. The dimensions of the transposed convolution are 16×16. Each of the first two convolutions in block B_1 consist of 64 filters, and the last one of 128. The first block C and first two I blocks contain 64, 64, 256 filters, respectively, while the second block C and the following three I blocks consist of 128, 128 and 512 filters, respectively. The third block C and the following five I blocks contain 256, 256 and 1024, respectively, and the fourth block C along with the last two I blocks consist of 512, 512 and 2048 filters, respectively.

4. Training with virtual data

An attempt was made to improve the accuracy of neural networks using computer-generated virtual training data that originates from the virtual city environment. Some mixed datasets were compiled which contain real-world images and computer-generated virtual images. The real-world images originate from the Cityscapes Dataset, a large-scale dataset for semantic segmentation [13]. The dataset contains 5000 annotated images with fine annotations created in 50 different cities under various weather conditions. 30 object classes are included, e.g. roads, sidewalks, people, vehicles, traffic lights, terrain, sky, etc. but in this research, only road surface segmentation is examined. The computer-generated images originate from the simulation environment described in Section 2.

For road surface segmentation, five different datasets are created from the Cityscapes Dataset and our collection of virtual images. Table 1 shows how these two collections were mixed. Our goals are to use a minimum amount of data from a real-world dataset, and when the number of virtual images is changed, to observe how the efficiency of the neural networks is affected. Dataset A only contains real-world images, therefore, this is regarded as the basic dataset, while the others were compared to it. Dataset B already contains the same number of virtual images as real-world images. Here, observations of how the introduction of virtual images changes the initial degree of efficiency are sought. Dataset C contains three times more virtual images than Dataset B. If the number of virtual images is much higher than the number of real-world images, the efficiency may be reduced. A future paper of ours will investigate this. In Datasets D and E the number of real-world images was increased. For the segmentation of protective barriers, only virtual training data were used. How efficiently the neural network recognizes real objects, if only trained by virtual data, will now be shown.

Table 1: Number of images in our mixed datasets

Dataset name	Training set		Validation set	
	Virtual	Real-world	Virtual	Real-world
A	0	500	0	125
B	500	500	0	250
C	1500	500	0	500
D	1500	1000	0	625
E	1500	1500	0	750

The effect of increasing the number of real-world images on efficiency was investigated. Adam optimization was used for training with a learning rate of 10^{-4} and a learning rate decay of 5×10^{-4}. As the objective function, categorical crossentropy is used:

$$L\left(y, \hat{y}\right) = -y\times\log\left(\hat{y}\right) \tag{23}$$

and the dice coefficient measured:

$$dc\left(y, \hat{y}\right) = 1-\frac{2\times y\times\hat{y}+1}{y+\hat{y}+1} \tag{24}$$

where $y \in \{0, 1\}$ is the ground truth and $0 \leq \hat{y} \leq 1$ is the result of the neural network.

5. Results

An attempt was made to examine the efficiency of road surface detection, while the composition of the dataset was modified. For examining changes in efficiency, the most useful datasets were A, C and E. Dataset A is the basic dataset, which only contains a small set of real-world images. Dataset C is based on Dataset A, but contains three times as many virtual images as real-world images. Dataset C shows how performance changes, when virtual world images are integrated into a small dataset. In Dataset E, the size of the collection was expanded. This dataset shows how much greater the efficiency of a larger mixed dataset is. Fig. 7 shows the validation accuracy over the training process of road surface detection, while Fig. 8 shows the best dice coefficient values for road surface segmentation. FCN is much simpler than both U-Net and PSPNet neural network architectures.

Hence the efficiency of the FCN on Dataset A is a little less than for the other networks. U-Net and PSPNet are very robust and complex, therefore, mixed datasets do not significantly increase the efficiency of these architectures. However, for simpler networks like FCN, this method improves the efficiency. Fig. 9 shows the perfor-

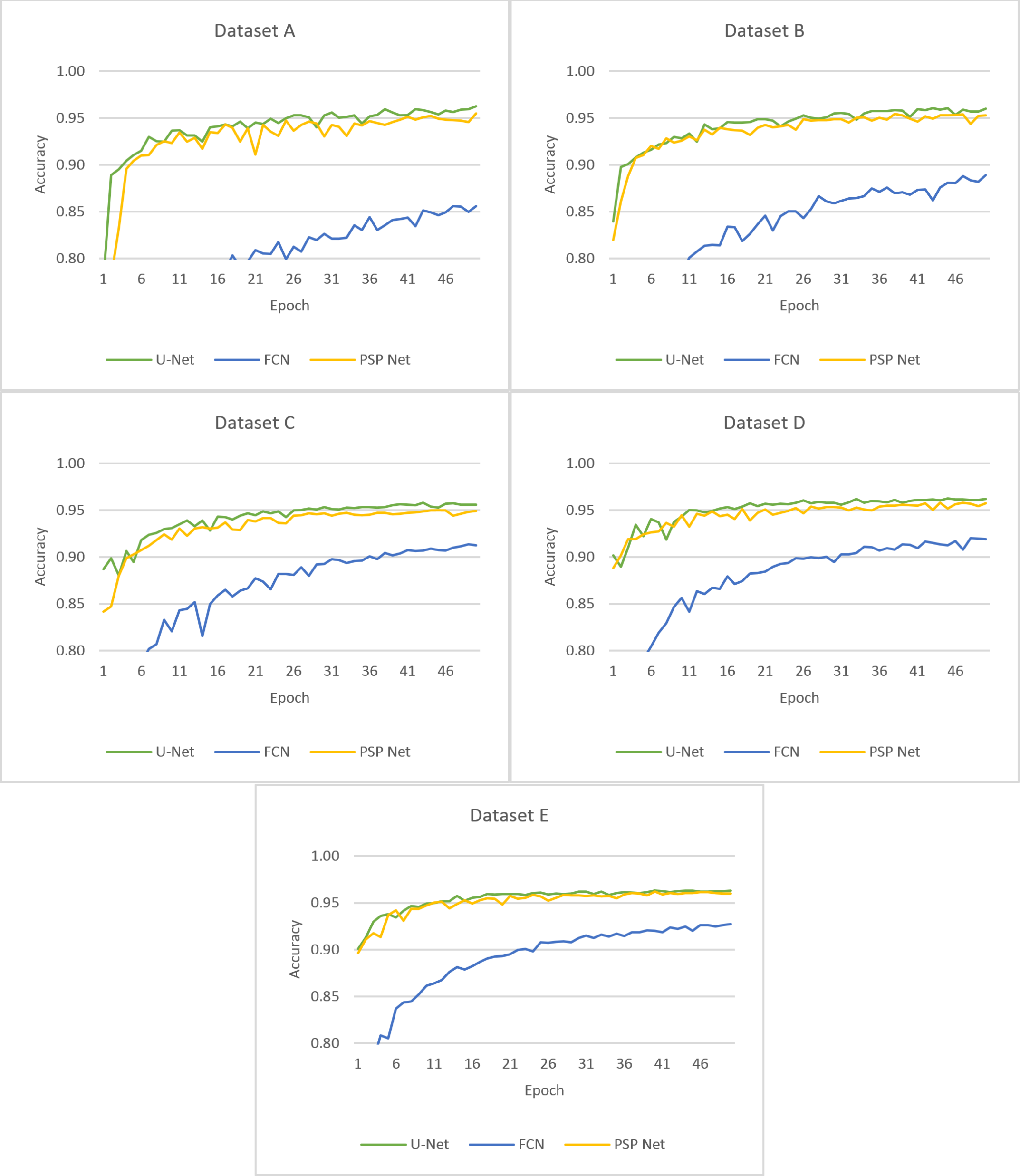


Figure 7: Road surface segmentation performance

mances with regard to the segmentation of protective barriers. Only virtual images were used to train the neural networks that determine the segmentation of protective barriers. This would not have been possible in the case of road surface segmentation, because the road surface is too complex. The texture of the protective barriers is very simple, therefore, it is possible to recognize it from virtual images alone.

It is our intention to use an environment detection system in a low-budget racing car, where the hardware resources available are limited and detection must occur in real time with a high degree of detection accuracy. Therefore, the neural network should be designed to be as simple as possible. If the neural network architecture is too simple, it is more difficult to train for complex recognition tasks. Moreover, the dataset concerning the racing track, protective barriers, etc. is not large. In this case, it is helpful to be able to train simpler neural networks, e.g. FCN, with virtual datasets to achieve higher degrees of efficiency. Experience has shown that the efficiency of road surface detection is improved by using three times as many virtual images, while for protective barrier detection it is sufficient to only use virtual images.

6. Conclusion

This paper presents how to use computer-generated virtual images to train artificial neural networks when the amount of available real-world images is limited. Three different neural network architectures, namely FCN, U-Net and PSPNet, were investigated and these networks trained with mixed datasets. It was shown that virtual im-

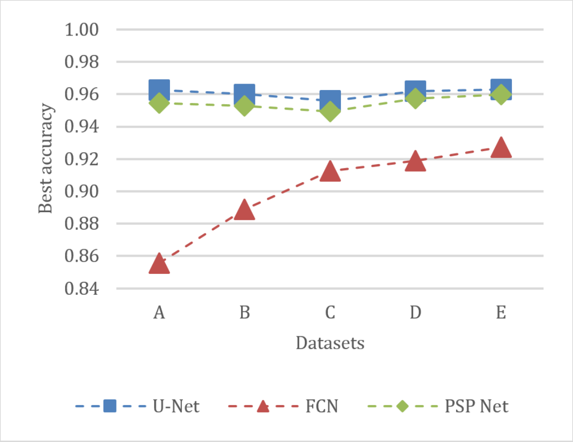


Figure 8: Best accuracy of road segmentation

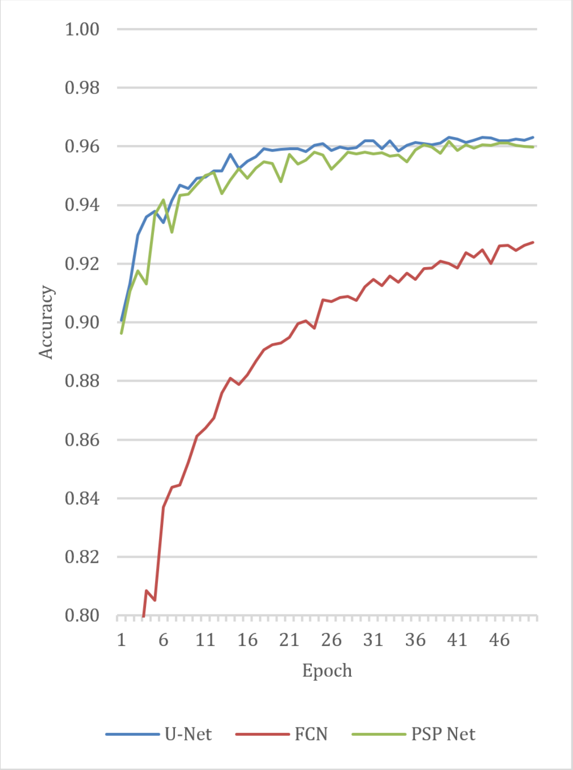


Figure 9: Barrier segmentation performance

ages improve the efficiency of neural networks. Our research demonstrates that when the texture of the objects is simple, e.g. that of protective barriers, it is sufficient to only use virtual image-based training datasets. This work may help us to create an efficient environment detector for the Shell Eco-marathon, where special objects have to be detected in the absence of real-world datasets.

Acknowledgements

The research was carried out as part of the “Autonomous Vehicle Systems Research related to the Autonomous Vehicle Proving Ground of Zalaegerszeg (EFOP-3.6.2-16-2017-00002)” project in the framework of the New Széchenyi Plan. The completion of this project is funded by the European Union and co-financed by the European Social Fund.

REFERENCES

[1] Krizhevsky, G. A.; Sutskever, I.; Hinton, G. E.: ImageNet classification with deep convolutional neural networks, *Commun. ACM*, 2017 **60**(6), 84–90 DOI: 10.1145/3065386

[2] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition, 3rd International Conference on Learning Representations, San Diego, USA, 2015

[3] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.: Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015 DOI: 10.1109/CVPR.2015.7298594

[4] Shelhamer, E.; Long, J.; Darrell, T.: Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017 **39**(4) 640–651 DOI: 10.1109/CVPR.2015.7298965

[5] Ronneberger, O.; Fischer, P.; Brox, T.: U-Net: convolutional networks for biomedical image segmentation, In: Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015: Lecture Notes in Computer Science, **9351** 234–241, Springer: Cham, Switzerland, 2015 DOI: 10.1007/978-3-319-24574-4_28

[6] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 770–778, 2016 DOI: 10.1109/CVPR.2016.90

[7] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J.: Pyramid Scene Parsing Network, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 6230–6239 2017 DOI: 10.1109/CVPR.2017.660

[8] Peng, X.; Sun, B.; Ali, K.; Saenko, K.: Learning deep object detectors from 3D models, IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 1278–1286, 2015 DOI: 10.1109/ICCV.2015.151

[9] Tian, Y.; Li, X.; Wang, K.; Wang, F.: Training and Testing Object Detectors with Virtual Images, *IEEE/CAA J. Autom. Sin.*, 2018 **5**(2) 539–546 DOI: 10.1109/JAS.2017.7510841

[10] Židek, K.; Lazorík, P.; Pitel, J.; Hošovskı, A.: An automated training of deep learning networks by 3D virtual models for object recognition, *Symmetry*, 2019 **11** 496–511 DOI: 10.3390/sym11040496

[11] Unrealengine.com 2020. Unreal Engine | The Most Powerful Real-Time 3D Creation Platform.

https://www.unrealengine.com/en-US/ [Accessed 14 September 2019]

[12] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.;Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *Int. J. Computer Vision*, 2015 **115** 211–252 DOI: 10.1007/s11263-015-0816-y

[13] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 3213–3223 2016 DOI: 10.1109/CVPR.2016.350

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 11–17 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-03

THEORETICAL BACKGROUND AND APPLICATION OF MULTIPLE GOAL PURSUIT TRAJECTORY FOLLOWER

Ernő Horváth*[1,2], Claudiu Pozna[3], Péter Kőrös[2], Csaba Hajdu[2], and Áron Ballagi[2]

[1] Department of Computer Engineering, Széchenyi István University, Egyetem tér 1, H-9026 Győr, HUNGARY
[2] Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, H-9026 Győr, HUNGARY
[3] Faculty of Electrical Engineering and Computer Science, Transilvania University of Brasov, B-dul Eroilor nr. 29, 500036 Brasov, ROMANIA

Autonomous and self-driving technology is a rapidly emerging field among automotive-related companies and academic research institutes. The main challenges include sensory perception, prediction, trajectory planning and trajectory execution. The current paper introduces a design strategy and the mathematical background of the optimization problem with regard to the multiple goal pure pursuit algorithm. The aim of the algorithm is to provide a low degree of computational complexity and, therefore, a fast trajectory-tracking approach. Finally, in terms of our approach, not only the theoretical questions but the application challenges will be described as well.

Keywords: autonomous, self-driving, algorithm, trajectory-tracking

1. Introduction

The current paper deals with the design strategies, theoretical background and application of trajectory tracking. The trajectory-follower approach that is discussed is called the multiple goal pure pursuit algorithm, in particular the windowed version. In mobile robots or vehicles, the trajectory design strategy seeks to identify the best possible trajectory which approximates a desired trajectory based on waypoints. This also assumes that the trajectory design phase has already created the desired trajectory.

This paper is organized as follows. The first section introduces the general problem and summarizes the state-of-the art and most relevant results. The second section deals with the design strategy of trajectory tracking and defines the problem precisely. The third section defines the optimization problem and possible solutions, while the fourth section summarizes the results. Finally, the last section includes a summary and conclusions.

The current state-of-the-art results The essence of pure pursuit is to choose a look-ahead point from the desired trajectory in front of the vehicle and steer according to it. This simple idea can be realized in many different ways, e.g., the classical pure pursuit algorithm uses a fixed look-ahead distance. This means that on the given track, a look-ahead point can be determined as a subset of waypoints, which are a fixed distance from the vehicle. Choosing a look-ahead distance is a tradeoff between control overshoots and precision. If the look-ahead distance is small, the overshoot will appear because the chosen point is so close that the vehicle cannot react instantaneously, thus continuously exceeds its target. In this case, the absolute distance (error) is smaller but the vehicle will oscillate. In contrast, if the chosen point is too far away, the vehicle will follow the point after the following corner so cuts off the bend. This behavior can be eliminated if the look-ahead distance is adjustable or changeable. One of the most obvious ways to modify the look-ahead distance is to scale the speed of the vehicle accordingly [1]. The other method of modifying the look-ahead distance involves a fuzzy controller that automatically adjusts the look-ahead distance based on path characteristics, velocity and tracking errors [2, 3]. Similarly to the fuzzy approach [4], it was shown that an autonomous vehicle can drive at a velocity of approximately 80 km/h along explicit paths using differential GPS data. To improve the classical pure pursuit algorithm and eliminate steering latency, CF-Pursuit [3] replaced the circles employed in pure pursuit with a clothoid curve to reduce fitting errors.

Although many papers have examined this domain, the available source code is rather rare. A few of the most notable versions are Autoware [5], Python [6] and Raptor Unmanned Ground Vehicle (UGV) [7]. Our research group also introduced a pure pursuit method based on multiple look-ahead points [8], in this paper its updated

*Correspondence: herno@ga.sze.hu

version will be presented.

2. The design strategy

The motivation of our research was to construct a simple but flexible kinematic trajectory tracking approach. In addition, the implementation of a *more human-like thinking* was sought, in other words, to track and follow more goal points on a road simultaneously. Drivers also take multiple considerations into account, e.g., not only is lane-keeping an important task but a cyclist in the same lane also influences planning. In our previous work the multiple goal pursuit algorithm can be viewed as a supplement to the pure pursuit algorithm. The most significant change was the calculation of curvature. Assuming the reference trajectory is given with a set of geometric points T, position P, steering angle γ and orientation θ of the vehicle, $G \subseteq T$ is defined as the set of selected goal points with a preset length $N = |G|$:

$$G = \{G_1, G_2, \ldots, G_N | G_k, (l+o) \leq k < (l+o+N)\} \tag{1}$$

An angle α_f from a domain of possible angles $[\alpha_{\min}, \alpha_{\max}]$ is assumed (presumably the wheel angle limits). A sequence of curvatures can be calculated each with a radius $p_f = -2/\alpha$ and a centre point C_f. A goal point G_k from set G to any C_f determines a line segment: the normalized difference between the length of this segment and radius p_f is the distance $d \in \mathbb{R}^+$ from the curvature. The metric for selecting a good angle is the sum of this difference for each goal point.

$$d_{\text{sum}} = \sum_{k=0}^{N} \left| \left\| \overrightarrow{C_f G_k} \right\| - \rho_f \right|. \tag{2}$$

Here d_{sum} denotes the total distance for goal point set G. The aim is to minimize this distance.

This paper introduces a new approach to the multiple goal pure pursuit algorithm, which is based on *windowing*. This algorithm is not regarded as a classical trajectory tracker since it slightly redesigns the original trajectory. With this method, a trajectory similar to the original one is created, but slight modifications render it more suitable for our vehicle to track. The design strategy consists of identifying the optimum trajectory which approximates a desired trajectory defined by a set of chosen points. This means that the aforementioned set is initial data and a minimum problem must be defined.

From the multitude of possibilities, the minimum problem will be defined starting with the following hypothesis:

1. During the approximation, the steering angle γ is constant;
2. Changing the steering angle γ is an instantaneous process.

If the previous hypotheses are correlated, it can be concluded that the initial set of points is approximated by a trajectory composed from arcs of a circle. The transition from one arc to another is instantaneous. For a particular arc of the circle, a subset (referred to here as the *working set*) is selected from the initial set.

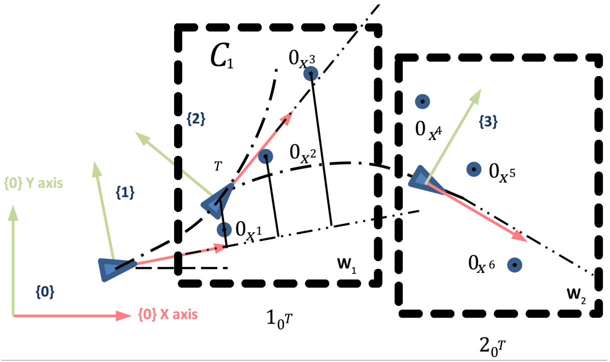


Figure 1: The windowed goal pursuit strategy

The following decisions are subjects of debate:

- The number of points included in the working set;
- The number of commune points which belong to different working sets;
- The position of the car when the next working set is defined.

The aforementioned observations are shown in Fig. 1, where a possible strategy is illustrated. The figure contains two referential systems. The first, 0, is the global (stationary) referential system; the second, $1, 2, 3$, is the mobile referential system attached to the vehicle. The blue triangles denote the abstraction of the vehicle. Two windows are labelled as $\mathbf{w}_{1,2}$. Each of them contains a *working set* of three points.

The trajectory is a composition of two arcs of a circle, C_1 and C_2. The first (C_1) is defined in referential frame 1 using the window $\mathbf{w}_1$, i.e., the points $0_{x^1}, 0_{x^2}$ and 0_{x^3}. The second (C_2) is defined in referential frame 2 using the window $\mathbf{w}_2$, i.e. the points $0_{x^4}, 0_{x^5}$ and 0_{x^6}. Each of the mentioned points is defined in referential frame 0.

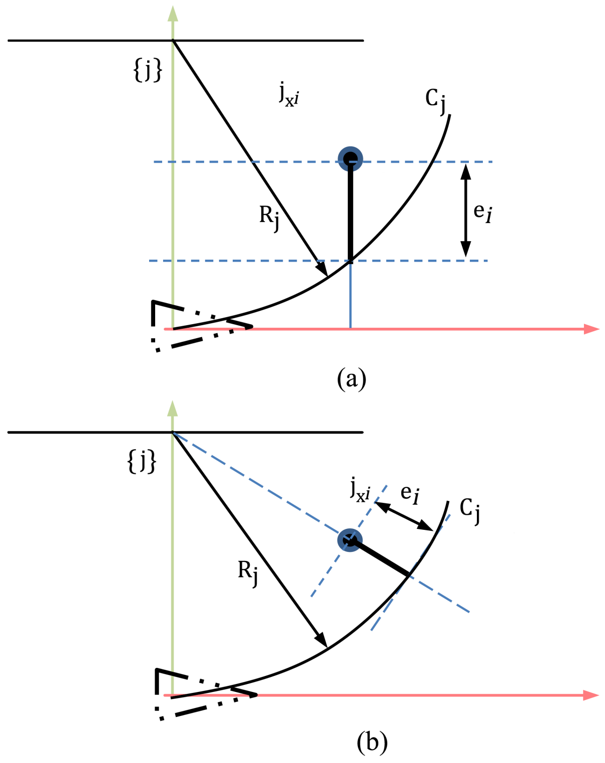
3. The optimization problem

Using m windows and n points for each window, the mathematical definition of the problem is

$$R_j = \arg\left(\min \sum_{i=1}^{m} e_i\right), \tag{3}$$

where R_j denotes the radius of arc C_j, e_i stands for the distance from point i to arc C_j, j represents the current window number $j = 1, \ldots, m$, and i is the current point number which belongs to window j.

The distance from point x_i to C_j is also debatable. Fig. 2 illustrates two possibilities: the first (Fig. 2a) only considers one coordinate of the current point. In contrast, in Fig. 2b the shortest distance is considered.



(a)

(b)

Figure 2: Possible ways of computing the distances

For the first choice:

$$e_i = y_i - C_j\left(x_i\right), \tag{4}$$

where

$$\begin{bmatrix} {}^j x_i \\ {}^j y_i \\ 1 \end{bmatrix} = {}^j_0 T\,{}^0\mathbf{x_i} = \begin{bmatrix} c\theta & -s\theta & 0 \\ s\theta & c\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} {}^0\mathbf{x_i}; \tag{5}$$

θ denotes the orientation of the referential system j; and x_j, y_j represent the coordinates of the origin of referential system j.

If this definition is assumed, it is evident that this is a quadratic regression problem with the analytical solution

$$\frac{1}{R} = \kappa = \left(\mathbf{X^t X}\right)^{-1}\left(\mathbf{X^t Y}\right), \tag{6}$$

where

$$\mathbf{X} = \begin{bmatrix} x_1^2 + y_1^2 \\ \dots \\ x_n^2 + y_n^2 \end{bmatrix} \tag{7}$$

$$\mathbf{Y} = \begin{bmatrix} 2y_1 \\ \dots \\ 2y_n \end{bmatrix} \tag{8}$$

For the second choice:

$$e_i = \left| R - \left\| \begin{bmatrix} x_i \\ y_i - R \end{bmatrix} \right\| \right| = \left| R - \sqrt{x_i^2 + \left(y_i - R\right)^2} \right| \tag{9}$$

which has a numerical solution.

3.1 Algorithm flowchart

In Fig. 3, a flowchart is presented in order to understand the approach more comprehensively.

The algorithm uses the vehicle parameters, pose, trajectory points and input parameters as inputs. The vehicle parameters are wheelbase b, steering angle γ and its position. Also, the points: ${}^0\mathbf{x}_k$ are part of the initial data. The algorithm parameters are: working set n, skyline d and curvature limit. This limit decides whether the locomotion is based on a linear or circular trajectory window. For every iteration, a new window is calculated.

4. Discussion

4.1 Simulation

The simulation uses the previous algorithm for a kinematic model which is in accordance with the starting hy-

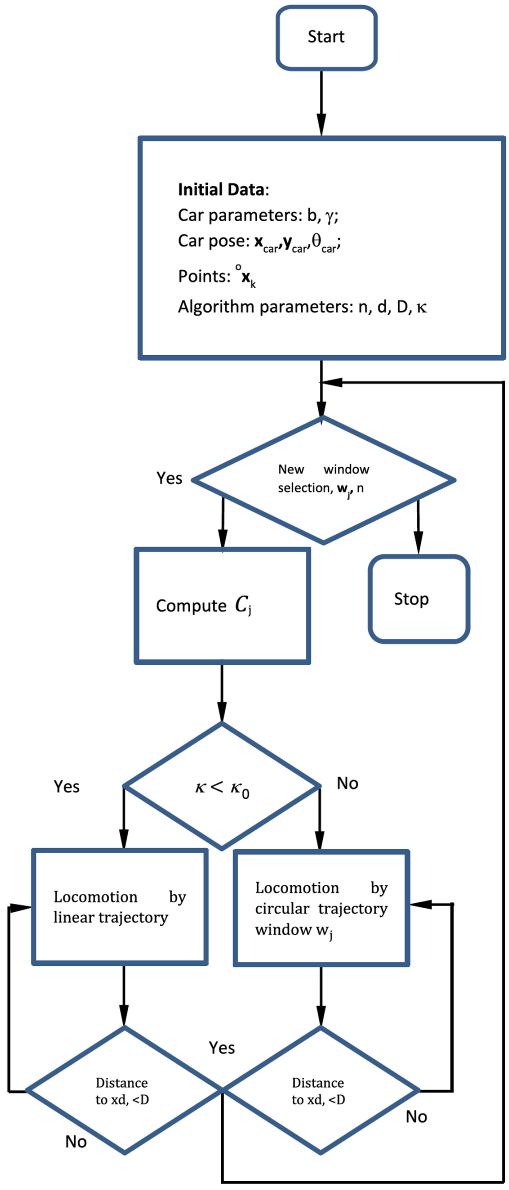

Start
Initial Data:
Car parameters: b, γ;
Car pose: $\mathbf{x}_{car}$,$\mathbf{y}_{car}$,θ_{car};
Points: ${}^o\mathbf{x}_k$
Algorithm parameters: n, d, D, κ
Yes
New window selection, $\mathbf{w}_j$, n
Compute C_j
Stop
Yes
$\kappa < \kappa_0$
No
Locomotion by linear trajectory
Locomotion by circular trajectory window w_j
Yes
Distance to xd, <D
Distance to xd, <D
No
No

Figure 3: Algorithm flowchart

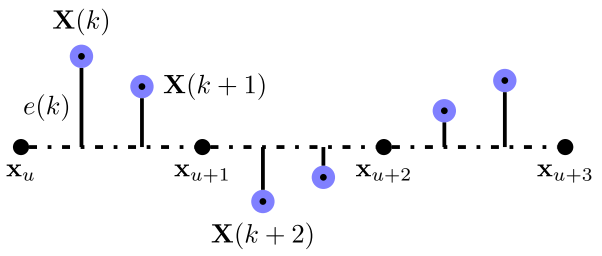


Figure 4: The definition of lateral deviation

pothesis, i.e., the steering is instantaneous. In the first simulation, the following kinematic bicycle model

$$\begin{cases} \dot{x} = v\cos(\theta) \\ \dot{y} = v\sin(\theta) \\ \dot{\theta} = \dfrac{v}{L}\tan(\gamma) \\ \gamma = ct \end{cases} \tag{10}$$

was used. To analyze the results, the lateral approximation error was defined:

- The lateral deviation error of the current position $e(k)$ is the distance between this position and the current segment of the desired trajectory (Fig. 4);
- The total deviation error e_T is the sum of the error throughout the simulation from 1 to N;
- The relative deviation error e_r is the ratio of the total error to the total number of simulated points.

$$\begin{aligned} e(k) &= \text{distance}\left(\mathbf{x}(k), \overline{{}^0\mathbf{x}_u{}^0\mathbf{x}_{u+1}}\right) = \\ &= \left|\overline{\mathbf{x}(k)\mathbf{x}_{u+1}} - \widehat{\mathbf{x}_u\mathbf{x}_{u+1}}\left(\overline{\mathbf{x}(k)\mathbf{x}_{u+1}} \cdot \overline{\mathbf{x}_u\mathbf{x}_{u+1}}\right)\right| \end{aligned} \tag{11}$$

$$e_T = \sum_{k=1}^{N} e(k) \tag{12}$$

$$e_r = \frac{e_T}{N} \tag{13}$$

These errors are dependent on the number of points included in the working set (n) referred to as the *skyline* and the position of the decision point (d). For the desired trajectory, a study relative to this dependency was conducted and the relative errors for each case were obtained (n, d).

The non-instantaneous steering was also simulated (Fig. 5). In this case, the kinematic bicycle model can be modified as follows:

$$\begin{cases} \dot{x} = v\cos(\theta) \\ \dot{y} = v\sin(\theta) \\ \dot{\theta} = \dfrac{v}{L}\tan(\gamma) \\ \gamma = \gamma_0 + \dot{\gamma}t \\ \dot{\gamma} = ct \end{cases} \tag{14}$$

The steering angle is a linear function where the steering velocity is constant during the steering process. The

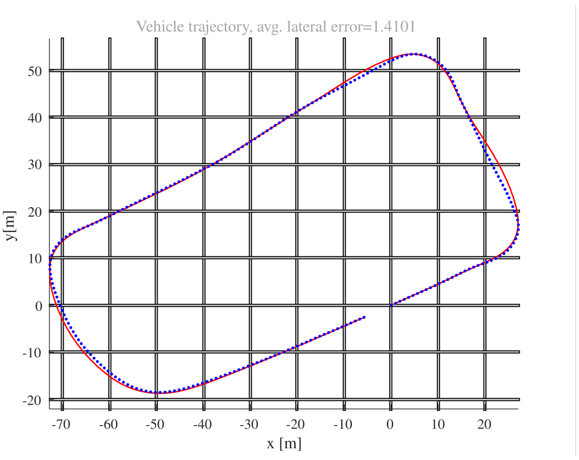


Figure 5: Simulation result: the blue dots denote the waypoints, the red line represents the trajectory generated by our model

trajectory of the robot is no longer an arc of a circle but a combination of two curves. The first of which is a clothoid (also referred to as an Euler spiral) when $\gamma = \gamma_0 + \dot{\gamma}t$ and an arc of a circle when $\gamma = ct$. In this case, the simulation result shows a lower quality of approximation. Given this phenomenon, the robot orientation angle error is defined as the difference between the orientation angle in the hypothesis where the steering is instantaneously modified to the final value and the orientation angle in the hypothesis where the steering angle is continuously modified from the initial to the final value.

$$\begin{aligned} e_\theta &= \frac{v}{L}\left(\tan(\gamma_\text{f})\,\Delta t - \int_0^{\Delta t} \tan(\gamma_0 + \dot{\gamma}\tau)d\tau\right) = \\ &= \frac{v}{\dot{\gamma}L}\left(\tan(\gamma_\text{f})\,\Delta\gamma - \log\left(\frac{\cos(\gamma_0)}{\cos(\gamma_\text{f})}\right)\right) \end{aligned} \tag{15}$$

where γ_0 denotes the initial steering angle, $\gamma_\text{f} = \arctan(L\kappa$ represents the final steering angle, $\dot{\gamma}$ stands for the steering angle velocity, and Δt is the steering time.

It is observed that Eq. 12 is in accordance with our intuitions: if the steering angle velocity increases the orientation error angle will decrease; in contrast, the velocity is proportional to the error.

A strategy where a maximum orientation error e_{θ_max} is allowed can now be contemplated and since the maximum steering velocity $\dot{\gamma}$ is an actuator parameter (which is constant), the following function for the vehicle velocity is proposed:

$$v_\text{new} = \frac{e_{\theta_\text{max}}}{e_\theta} v_\text{old}. \tag{16}$$

In conclusion, the algorithm must be improved in order to yield better results. The following possibilities are suggested (Fig. 6):

- Refine the input set of points: interpolate new points;
- Change the vehicle velocity during locomotion (Eq. 16);

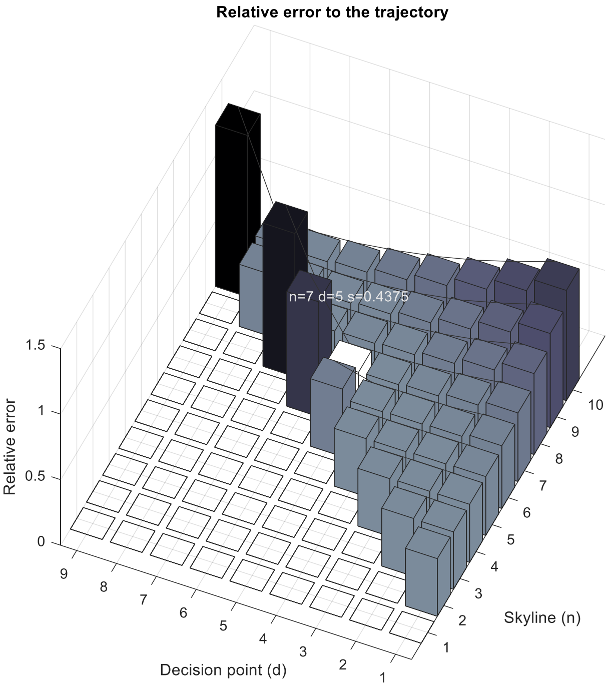


Figure 6: The proposed new strategy

- Define the new windows by including the points which succeed the decision point in the new working set.

With regard to an examined sample trajectory, a 3D graph (Fig. 7) is provided which shows the relative dependency obtained in terms of the relative error value in each case (n, d). In Fig. 7, the x axis denotes the skyline (n), the y axis represents the decision point (d), and the z axis stands for the relative error.

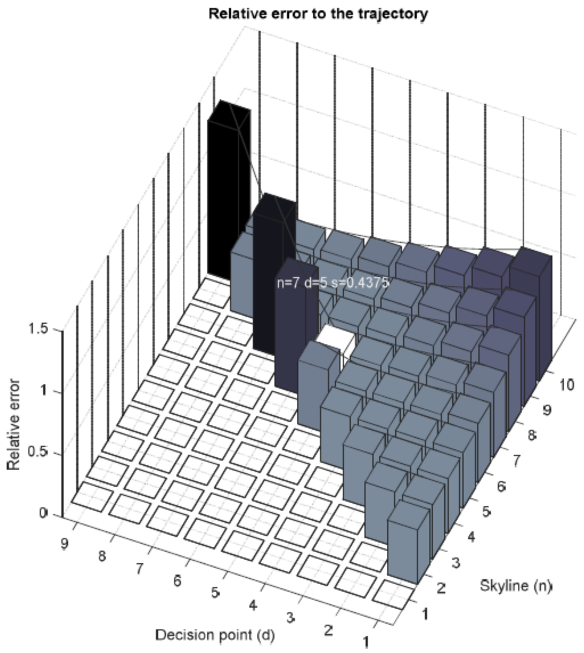


Figure 7: The relative error depends on the skyline (n) and decision point (d)

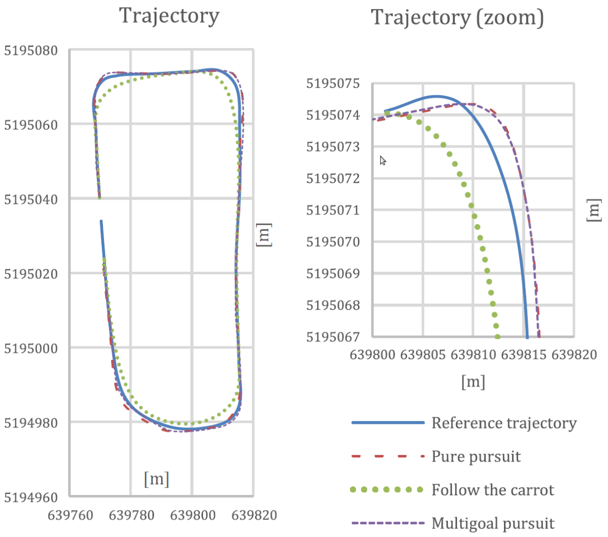


Figure 8: Typical behaviors of the examined approaches

4.2 Experiments

Firstly, the behavior of the algorithm in its own simulation environments was tested; the two presented approaches (the plain multiple goal pursuit and the windowed multiple goal version) were developed in parallel in Python and MATLAB. These simulations also used real-world trajectories and the vehicle model was first a kinematic, and later a dynamic one. Subsequently, it became necessary to even test the algorithms in Open Source Robotics Foundation's (OSRF) Gazebo for two reasons. Firstly, implementation of the algorithm as a Robot Operating System (ROS) node was sought and secondly, a more precise dynamic model of the vehicle already existed in Gazebo. Gazebo also facilitated the integration of the useful software as well as agile migration between development and target.

For these purposes, the algorithm had to be recreated in C++ not only because of the ROS compatibility but computational resources as well. Simulation-based execution yielded no significant deviation from the simulated tests. In Fig. 8, some of our results are shown regarding the examined trajectories. The x and y axes are in meters. This simulation is based on real-world measurements taken at the ZalaZone Automotive Test Track in Zalaegerszeg and the GPS coordinates are assigned using Universal Transverse Mercator (UTM) because of the minimal degree of distortion involved. The Follow the Carrot and speed ratio-based pure pursuit algorithms had to be implemented in order to experiment with them in the same environment.

In another experiment, autonomous functionalities were installed in an electric vehicle (Nissan Leaf). Fig. 9 represents the vehicle in operation. The operational domain of this vehicle was limited to the ZalaZone Auto-

Figure 9: The vehicle used in our experiments

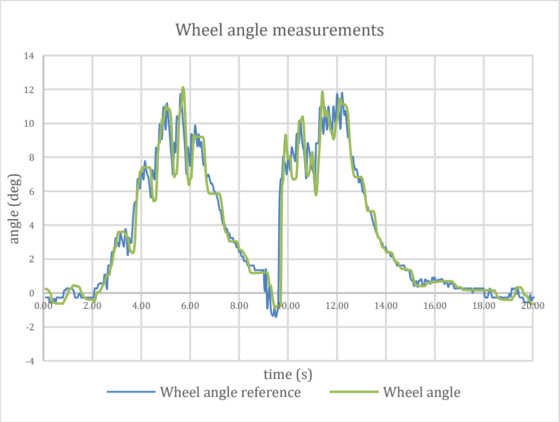


Figure 10: Changes in the wheel angle around two bends

motive Test Track and the university campus, moreover, it was required to reach a relatively slow speed of 25 km/h. The vehicle has front-wheel drive, with a wheelbase of 2.70 meters and a track width of 1.77 meters. To ensure instantaneous localization of the vehicle, two highly accurate Real Time Kinematic (RTK)-capable GPSs were used. One was KVH's Fiber Optic Gyro 3D Inertial Navigation System (GEO-FOG 3D INS), the other was Swift Navigation's Duro Inertial RTK. These sensors were also used to capture reference trajectories.

Our software was based on OSRF's ROS [9]. Our computing platform first features the NVIDIA Jetson TX2, later the NVIDIA Jetson AGX Xavier. Low-level control is realized by using a National Instruments' CompactRIO Controller, namely cRIO-9039, as the Real-time and FPGA modules - the NI-9853 CAN, NI-9403 DI, NI-9205 Voltage Input and NI-9264 Voltage Output Modules to be exact.

After the algorithm exhibited satisfactory behavior in the simulation environment, it was tested in a real-world scenario. During the experiment, the NVIDIA Jetson TX2 was used in the Nissan Leaf. The NVIDIA Jetson TX2 is a 7.5-watt embedded controller in which the Ubuntu 18.04 LTS Bionic Beaver along with the ROS Melodic Morenia were used. The embedded controller had a memory capacity of 8 GB and memory bandwidth of 59.7 GB/s as well as NVIDIA's Denver2, quad-core ARM Cortex-A57 CPU and integrated 256-core Pascal GPU installed. During our experiments, only the KVH's GEO-FOG 3D INS sensor was used as the source of localization, the wheel angle reference signal in addition to the speed reference via CAN were provided, and the wheel angle and speed were measured back in the same protocol. Nevertheless, during acquisition of the data, all the sensory information was logged by two of Velodyne's 16-channel LIDARs, Sick's 1-channel Laser Measurement Sensor LMS111 and the camera stream into ROS bag files. The following chart (Fig. 10) shows an example of how the algorithm generated reference trajectories for two bends on the routine track, moreover, the measured wheel angle is shown.

The waypoints for the algorithm were provided by driving through a predetermined path then saving as well as filtering the position and speed data via the waypoint_saver node.

In this experiment, only the multiple goal pursuit algorithm with one parameter set was used. This set is better compared to Autoware's pure pursuit algorithm in terms of lateral deviation, but the speed ratio-based version worked even better. As a result, the multiple goal pursuit algorithm was able to control the vehicle with the aforementioned embedded controller at slow speeds of approximately 25 km/h. The algorithm is able to perform smooth and continuous movements along a predefined trajectory, even maneuvering around bends and along edges.

5. Conclusion

The current paper described the development of a trajectory-tracking approach, namely the multiple goal pursuit, and summarized the mathematical and theoretical background needed to understand its working principles. Enhancements to and variations in the basic approach are also described with regard to their benefits and weaknesses. Furthermore, a brief insight into the development process is given. Firstly, an initial version was developed in our own simulation environment, later the code became an ROS node and the simulation environment was replaced by Gazebo. Finally, real-world tests showed the viability of the new algorithm, which was tested by an autonomously guided vehicle in a closed but real-world traffic environment.

One of the benefits of this algorithm is that it can be tuned more precisely around bends, moreover, *it involves more human-like thinking*, that is, tracks and follows more waypoints on the road simultaneously. Furthermore, this approach provides quick and reliable reference signals for car-like robot kinematics, thus can follow the trajectory with relatively small errors.

The discussed algorithm was implemented in MATLAB and C++ as a Robot Operating System node and our measurements were based on debug data generated by this node. This implementation is currently publicly available on GitHub in addition to other measurements, videos, results and source codes: `https://github.com/szenergy/szenergy-public-resources`.

Acknowledgements

The research was carried out as part of the “Autonomous Vehicle Systems Research related to the Autonomous Vehicle Proving Ground of Zalaegerszeg (EFOP-3.6.2-16-2017-00002)" project in the framework of the New Széchenyi Plan. The completion of this project is funded by the European Union and co-financed by the European Social Fund.

REFERENCES

[1] Paden, B.; Čáp, M.; Yong, S. Z.; Yershov, D.; Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles, *IEEE Trans. Intell. Veh.*, 2016, **1**(1), 33–35 DOI: 10.1109/TIV.2016.2578706

[2] Ollero, A.; García-Cerezo, A.; Martinez, J.: Fuzzy supervisory path tracking of mobile reports, *Control Eng. Practice*, 1994, **2**(2), 313–319 DOI: 10.1016/0967-0661(94)90213-5

[3] Samuel, M.; Hussein, M.; Mohamad, M. B.: A review of some pure-pursuit based path tracking techniques for control of autonomous vehicle, *Int. J. Computer Applications* 2016, **135**(1) 35–38 DOI: 10.5120/ijca2016908314

[4] Rodríguez-Castaño, A.; Heredia, G.; Ollero, A.: Analysis of a GPS-based fuzzy supervised path tracking system for large unmanned vehicles, *IFAC Proceedings Volumes*, 2000, **33**(25) 125–130 DOI: 10.1016/S1474-6670(17)39327-8

[5] Kato, S.; Tokunaga, S.; Maruyama, Y.; Maeda, S.; Hirabayashi, M.; Kitsukawa, Y.; Monrroy, A.; Ando, T.; Fujii, Y.; Azumi, T.: Autoware on board: Enabling autonomous vehicles with embedded systems, in Proceedings of the 9th ACM/IEEE International Conference on Cyber-Physical Systems, 2018, 287–296 DOI: 10.1109/ICCPS.2018.00035

[6] Sakai, A.; Ingram, D.; Dinius, J.; Chawla, K.; Raffin, A.; Paques, A.: PythonRobotics: a Python code collection of robotics algorithms, 2018, arXiv: 1808.10703

[7] Giesbrecht, J.; Mackay, D.; Collier, J.; Verret, S.: Path tracking for unmanned ground vehicle navigation: Implementation and adaptation of the pure pursuit algorithm, Defence Research and Development Canada (DRDC), Suffield, Alberta, Canada, 2005. https://apps.dtic.mil/sti/pdfs/ADA599492.pdf

[8] Horváth, E.; Hajdu, Cs.; Kőrös, P.: Novel pure-pursuit trajectory following approaches and their practical, in 10th IEEE International Conference on Infocommunications (CogInfoCom), Naples, Italy, 2019.

[9] Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; Ng, A. Y.: ROS: an open-source Robot Operating System, ICRA Workshop on Open Source Software, 2009 **3**(2)

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 19–23 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-04

HIGH-POWER MODULAR INVERTER DEVELOPMENT FOR ELECTRIC MOTOR TESTING PURPOSES

ZOLTÁN SZELI*[1] AND GÁBOR SZAKÁLLAS[1]

[1]Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY

This paper presents the development process of a high-power modular inverter system. The goal was to develop a universal inverter system (motor controller + power electronics) in which the performance of the power stage is scalable. The design concepts, hardware architecture, system components, built-in features and protections are described regarding the power stage.

Keywords: modular, universal, desaturation, active clamping, protection

1. Introduction

Power electronics and their control algorithms are very important parts of electric vehicles. The whole system can operate reliably, efficiently and with good dynamics only if all elements of the vehicle are well coordinated. Before the different electric drivetrains can finally be applied, they should be tested under different circumstances on electric motor or drivetrain test benches. Otherwise it would be difficult to test them in-vehicle as they are already built-in. For such testing processes, a flexible test system is a good solution that can be easily modified at both hardware and software levels, especially for institutions like our research center where various types of electric motors from different manufacturers are dealt with. Their inverter systems are usually inaccessible without the help of experts.

2. Design Concept

The purpose of the development was to design a universal and scalable (at the power stage) inverter system. Thus, the same system can be used for three-phase electrical machines with different power requirements and applications. The first version is capable of 600 A at a nominal system voltage of 400 V. As the motor controller is equipped with standard interfaces, it is able to adapt power stages to different power levels. For the design of the hardware for the motor controller and power stage, state-of-the-art solutions were used.

Furthermore, different motor control algorithms are dealt with at the research center, so it was decided to use them with the motor controller to ensure the software environment is capable of further developments and different types of motors can be easily adapted to it. Motor control software can be implemented using model-based simulation (MATLAB/Simulink) with automatic code generation. As a result, the behaviour of the motor can be determined in advance by simulation and easily adapted to our specific environment.

The system supports different communication protocols, e.g. CAN, RS232, Ethernet and FlexRay, which can be used for control, data acquisition and status information polling. The system is usually used to test three-phase electric machines or can be installed in electric vehicles.

3. Results and Analysis

The inverter system (Fig. 1) consists of four separate printed circuit boards (PCBs), namely Control Board,

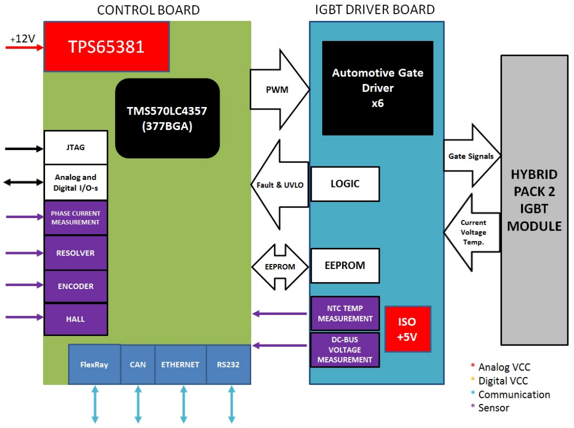


Figure 1: Hardware structure

*Correspondence: szgabor@ga.sze.hu

Power Stage, F-S-800 Insulated Gate Bipolar Transistor (IGBT) Interface Board, and a Current Measurement Board for current transducers. Only commercially available and automotive qualified components were used for the hardware. With regard to the different circuits on the boards, the SPICE simulator was used to support the design.

The main features:

- Undervoltage protection
- Overcurrent protection
- Isolated sensor supply
- Isolated voltage measurement
- Isolated DC/DC power supply for the IGBT drivers
- Temperature measurements
- Built-in protections:
  - Surge protector
  - Short circuit
  - Desaturation
  - Advanced Active Clamping

3.1 Power stage

The power stage was designed to be able to drive a maximum of two IGBT modules simultaneously (one IGBT module can also be used by us). Thus, the output power of the inverter is scalable and can easily be doubled should a different IGBT module be used.

3.2 Isolated power supply for the gate drivers

The power stage includes 6 separate Flyback DC/DC converters that produce an isolated +15 V / −8 V asymmetrical power supply for IGBT gate drivers as seen in Fig. 2.

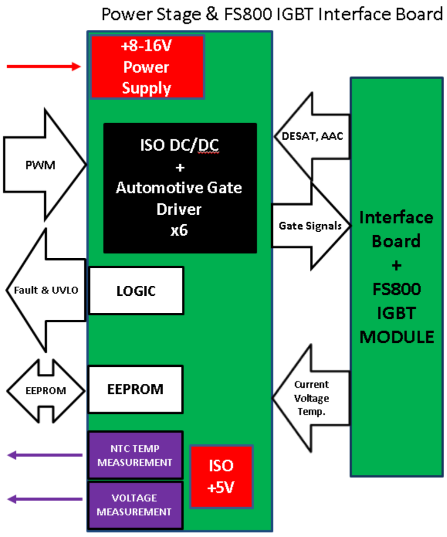


Figure 2: Power stage

Design parameters according to the system requirements and the Flyback Converter datasheet [1]:

- 8-18 V DC input voltage
- −8 V / +15 V output voltage
- 10 W output power (due to IGBTs in parallel, $f_{\text{sw}} = 20$ kHz)
- 3, 956 W power consumption per IGBT module was calculated from [2]

$$P_{\text{Gdr}} = Q_G \left(V_{\text{GE(on)}} - V_{\text{GE(off)}} \right) f_{\text{sw}} \qquad (1)$$

The output power selected was chosen for 2 IGBT modules in parallel.

The isolated power supply for the gate drivers works as expected in the real circuit, as can be seen in Figs. 3 and 4.

3.3 Gate drive circuit

The IGBT modeules were powered by Infineon's 1ED020I12FA2 galvanically isolated single channel IGBT gate driver circuits, which were placed on a separate PCB. The maximum output current of the 1ED020I12FA2 driver was ±2 A. Because of this low peak current, significantly increased gate resistors were required, which greatly degraded the switching properties of the IGBT driver. Therefore, it was necessary to

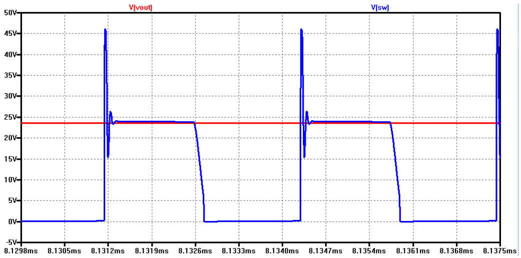
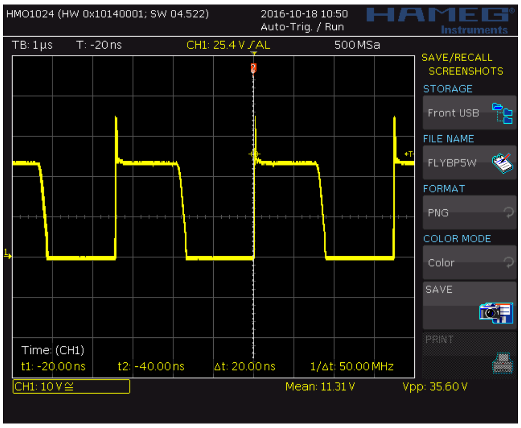


Figure 3: Simulated Flyback DC/DC switching waveform ($P_{\text{out}} = 5$ W)


HMO1024 (HW 0x10140001; SW 04.522) 2016-10-18 10:50 Auto-Trig. / Run

TB: 1µs T: -20ns CH1: 25.4 V ⁄AL 500MSa

Time: (CH1)
t1: -20.00ns t2: -40.00ns Δt: 20.00ns 1/Δt: 50.00MHz

CH1: 10 V ≅ Mean: 11.31 V Vpp: 35.60 V

STORAGE: Front USB
FILE NAME: FLYBP5W
FORMAT: PNG
COLOR MODE: Color

Figure 4: Flyback DC/DC switching oscilloscope waveform ($P_{\text{out}} = 5$ W)

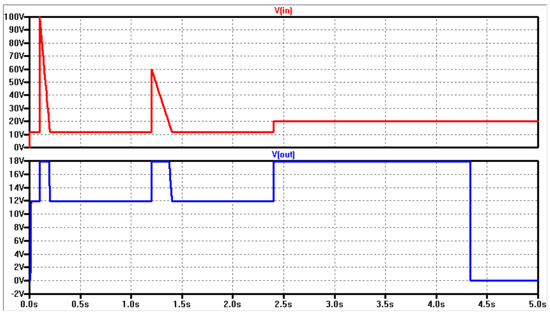


Figure 5: Operation of the surge stopper protection

use a so-called Bipolar Totem-Pole amplifier between the output of the driver and the IGBT gate terminal. The designed Bipolar Totem-Pole circuit was capable of driving the IGBT with the use of a resistor of $R_{\mathrm{G_{off}}} = 0.82\ \Omega$.

3.4 Built-in features

Surge stopper The voltage supply (+8 V – +16 V) was connected to the load through an overvoltage protection circuit. This circuit protects the other parts of the power stage from high voltage spikes and controls the output voltage in case of input overvoltages (Load Dumps) as shown in Fig. 5. It is also equipped with overcurrent and reverse polarity protection as well as an adjustable low-voltage fault detection limit.

Desaturation protection The IGBT driver is a power semiconductor that is normally used to switch operation in the closed and saturation regions. The closed region can be considered to be a broken wire, and in the saturation region it functions as a short circuit. Between the IGBT collector and emitter terminals, the saturation voltage, V_{CE}, can be measured [3]. This voltage depends on the temperature of the semiconductor, its collector current and the voltage applied to the gate. Fig. 6 shows the dependence of the saturation voltage of the FS800R07A2E3 IGBT module with regard to its collector current on the temperature of the semiconductor.

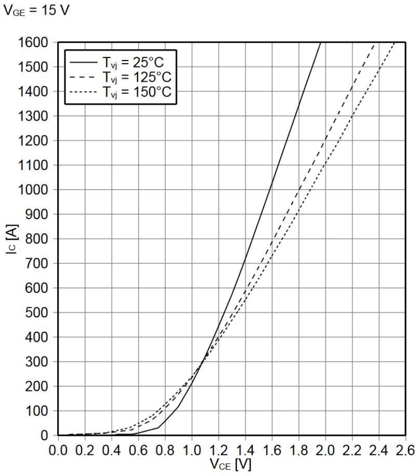


Figure 6: Output characteristics of the FS800R07A2E3 IGBT module

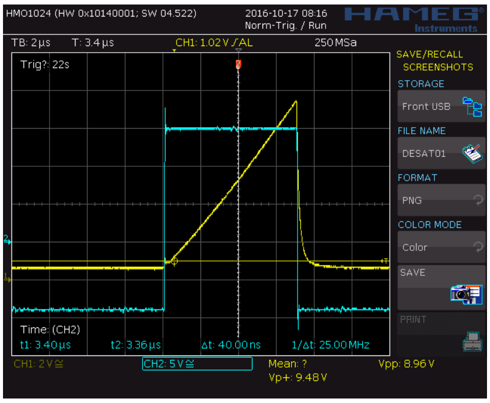


Figure 7: Operation of the desaturation protection

The desaturation voltage can be adjusted using Zener or fast diodes, so in our case, the threshold voltage, V_{CE}, selected of 1.3 V used in the switching will be subject to a 600 A collector current limitation per module at 25 °C.

The IGBT driver circuit is continuously monitoring the V_{CE}, using a comparator, and if that exceeds the reference voltage, that is 9 V, of the integrated circuit, the driver forces its output to VEE2 (-8 V) within a maximum T_{DESATOUT} time of 430 ns as shown in Fig. 7. With this kind of protection, short-circuit failures can be avoided and a temperature-dependent current limitation realized. The V_{CE} can be monitored only after a period of time equal to $T_{\mathrm{BLANKING}} + T_{\mathrm{DESATOUT}}$(1.8 μs+430 ns= 2.23 μs) has elapsed following the arrival of the control signal [4].

The fault signal is only active when the IGBT driver is switched on.

Advanced Active Clamping When working with IGBTs, parasitic inductances inside them and the converter circuits must be taken into consideration. Their effects cannot be completely eliminated for physical reasons and their influence on the system behaviour cannot be neglected. Fig. 8 shows the parasitic inductances contained in a commutation circuit.

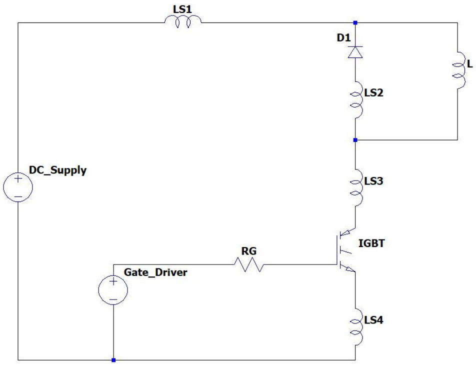


Figure 8: Parasitic inductances in IGBTs

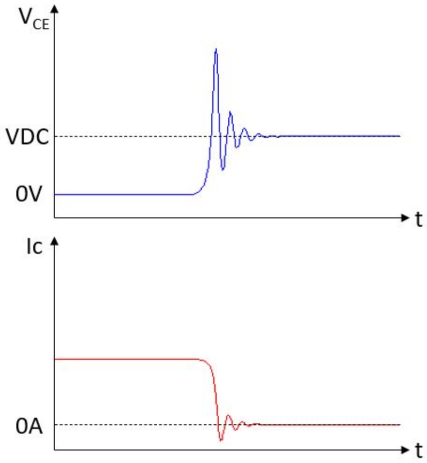


Figure 9: Characteristics of the IGBT turn-off process

Voltage transients are produced at the collector of the IGBTs during their turn-off process, which are caused by the change in current passing through them as shown in Fig. 9. The commutation speed and thus, the turn-off overvoltage at an IGBT can, in principle, be affected by the turn-off gate resistance $R_{\mathrm{G_{off}}}$. This technique is used most often in low-power devices where $R_{\mathrm{G_{off}}}$ must be selected for overload conditions such as turn-off of the double rated current, short circuits and a temporarily increased link circuit voltage. During normal operations, this causes increased turn-off switching losses and turn-off delays, which reduce the efficiency of the modules [5].

The FS800R07A2E3 IGBT module has a parasitic inductance of 14 nH.

To reduce the voltage spikes, a clamping circuit was used that allows charging of the gate-emitter capacitance and flow charges to the base of the proposed Bipolar Totem-Pole over the specified voltage range through a transient-voltage-suppression diode connected between the collector and the gate, driving the IGBT into the normal active mode and reducing the current-change rate, $\mathrm{d}I_{\mathrm{C}}/\mathrm{d}t$ as seen in Figs. 10 and 11.

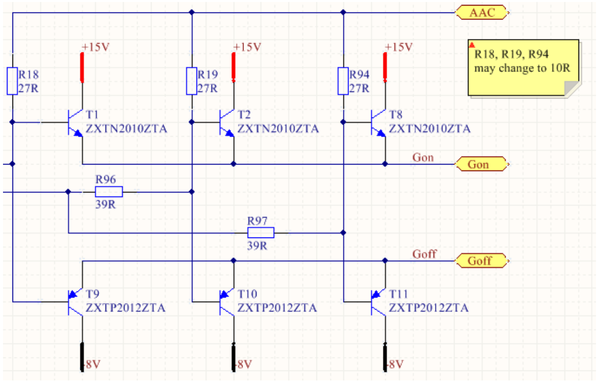


Figure 10: The proposed Bipolar Totem-Pole layout with feedback

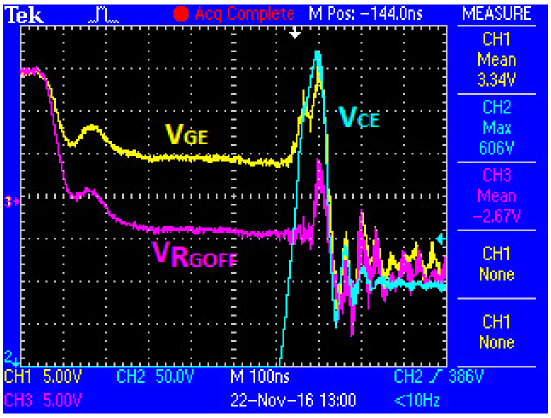


Figure 11: Operation of the Advanced Active Clamping circuit @ 340 V/ 800 A (single IGBT module)

This type of control of the Bipolar Totem-Pole circuit reduces the power dissipation as a result of the resistance $R_{\mathrm{G_{off}}}$ whilst operating.

3.5 Cooling system

To design the cooling system of the inverter, thermal simulations were carried out. The maximum power dissipation of the IGBT module was 1500 W, and its operating temperature range was between -40 and 150 °C. The thermal resistance between the liquid coolant (50% water, 50% glycol) and the surface of the heatsink was 0.097 K/W.

The IGBT manufacturing drawings were used to design the 3D model as seen in Fig. 12.

4. Conclusion

The functional testing of an inverter was carried out at the Electric Motor Test Bench Laboratory of Széchenyi István University. The developed system and the test environment can be seen in Figs. 13 and 14. A permanent-magnet synchronous motor (PMSM) was connected to it and different peripherals and built-in functions tested. On the other side of the test bench, the same type of PMSM connected to a 400 V / 500 A, downscaled (single IGBT module) version of the inverter was used. At first, the tests were performed at a speed of 1600 rpm and at low load

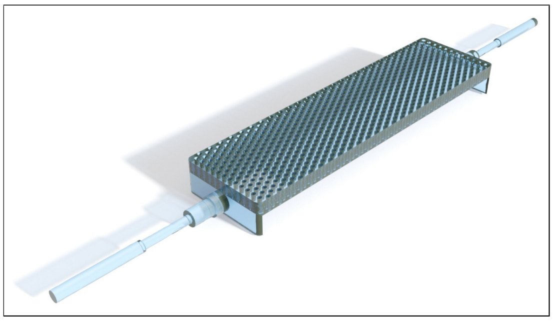


Figure 12: The designed coolant space inside the heatsink

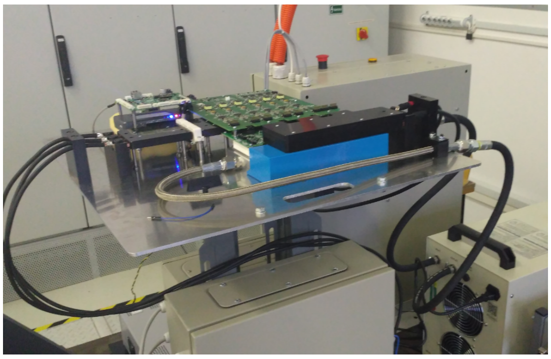


Figure 13: The assembled inverter system

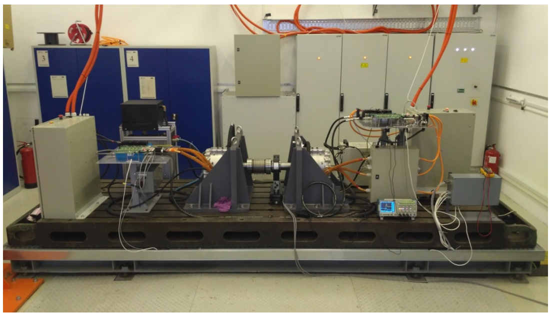


Figure 14: Eletric motor test bench with the mounted inverters

torques. The inverters were operated alternately, one in torque-control mode, the other in speed-control mode. In a future paper, these inverters will be tested under heavy-load conditions, when the motor control algorithms work reliably.

Acknowledgements

The research was carried out as part of the "Autonomous Vehicle Systems Research related to the Autonomous Vehicle Proving Ground of Zalaegerszeg (EFOP-3.6.2-16-2017-00002)" project in the framework of the New Széchenyi Plan. The completion of this project is funded by the European Union and co-financed by the European Social Fund.

REFERENCES

[1] Linear Technology, ”Flyback, Forward and Isolated Controllers“, LT8302 Datasheet, Sept. 2016 Rev. C

[2] Infineon Technologies AG, “Automotive IGBT Module Explanation of Technical Information”, Application Note AN2010-09, 2010

[3] Avago Technologies, “Desaturation Fault Detection”, Application Note AN5324, 2012

[4] Infineon Technologies AG, ”Galvanic Isolated Gate Driver“, 1ED020I12FA2 Datasheet, May 2013 Rev. 2.0

[5] Rüedi,H.; Thalheim, J.; Garcia, O.: Advantages of Advanced Active Clamping, *Power Electronics Europe*, 2009 **8**, 27–29

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 25–31 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-05

LIDAR-BASED COLLISION-FREE SPACE ESTIMATION APPROACH

Miklós Unger*[1], Ernő Horváth[1], and Csaba Hajdu[1]

[1]Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY

As autonomous technologies flourish within the vehicle industry, an increasing number of academic autonomous competitions are appearing. One of them is the Shell Eco-marathon Autonomous Urban Concept competition (SEM AUC) which seeks to provide hands-on experience for the academic community to design, build and test their own driverless vehicles within a realistic infrastructure. The team at our university participates in this competition and our concept is to rely on simple and robust algorithms. This paper presents a simple collision-free space estimation algorithm for the LiDAR sensor.

Keywords: Autonomous vehicle, Voronoi, Delaunay

1. Introduction

Nowadays autonomous technology is not only important in the automotive industry but increasingly in the fields of academia and reasearch as well. A sign of this tendency is the increasing number of academic competitions, one of which is the Shell Eco-marathon Autonomous Urban Concept competition (SEM AUC).

1.1 Shell Eco-marathon

The Shell Eco-marathon is a competition for engineering students to design and build the most ultra-efficient car. The participating teams come from all over the world. Two main classes of vehicles compete, namely Urban Concepts and prototypes. Both classes can participate in the Autonomous category where five different challenges await the teams as follows:

- Complex track section
- Maneuverability
- Obstacle avoidance
- The unknown challenge
- Autonomous distance

In this paper, a simple solution for the *autonomous distance* challenge is presented.

1.2 Our team

The SZEnergy Team is comprised of students from Széchenyi István University, assisted by tutors and researchers from the fields of the development and construction of electric vehicles. Since 2008, the Team has been competing each year in the Shell Eco-marathon, the biggest fuel-efficiency competition in the world. Our Team participates in the electric vehicle category of the Urban Concept class by performing autonomous challenges and regular races using the same car. Since 2019, a new car called SZEmission has been used as shown in Fig. 1.

Figure 1: Our car - SZEmission.

2. The challenge

Challenges of the AUC will be undertaken on the same standard track as the other competitions at the SEM or on similar but modified track sections. The track is 970 m long and includes inclines since the track partially consists of closed-off roads (Fig. 2). For all five challenges, protective barriers are installed on both sides of the track.

The barriers are 0.5 m high separated by small gaps, though the size of these gaps is unspecified. Contact with these barriers is forbidden, cars are limited to a top speed of 25 km/h and every car is driven separately on the track. The minimum width of the track is 6 meters unless otherwise specified.

[*] Correspondence: unger.miklos@ga.sze.hu

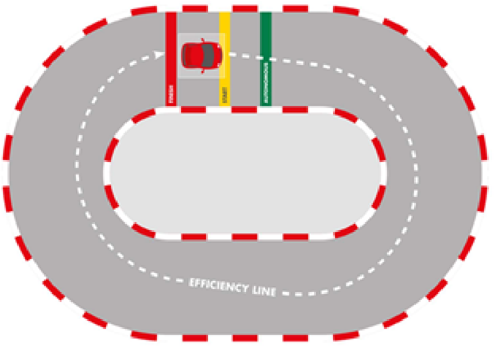


Figure 2: Simple layout of the challenge.

Our goal is to create an algorithm that despite the noisy surroundings robustly finds the coordinates of free route between the barriers without touching them based on the output of 3D LIDAR signals. The route must also be unambiguous.

3. LiDARs

LiDAR stands for Light Detection and Ranging or Laser Imaging Detection and Ranging and is a sensor commonly used in autonomous vehicles to map the environment. The sensor contains a transmitter and receiver. The transmitter fires laser beams towards the ground which interact with the objects of the environment then are deflected back to the receiver. The distance can be computed using a simple formula, where d_{o} denotes the distance to the object, c_{l} represents the speed of light, and Δt stands for the time of flight:

$$d_{\text{o}} = \frac{c_l \Delta t}{2}, \tag{1}$$

where the time of flight is measured indirectly by determining the phase shift between the transmitted and received signals, f_{mod} denotes the frequency modulation, φ_{r} represents the measured difference between the transmitted and received wavelengths [1], and

$$\Delta t = \frac{\varphi_{\text{r}}}{2\pi f_{\text{mod}}}. \tag{2}$$

LiDAR sensors on autonomous vehicles Autonomous vehicles use these rotating beam sensors to scan the environment and detect obstacles. Two main types of LiDARs, the 2D and 3D, are used. 2D LiDARs, also referred to as LiDARs with one channel, provide information about the environment in one plane only, hence the objects above and below the LiDARs are not detected. 3D LiDARs use more channels (16, 32, 64 or even 128), thus yield all x, y and z coordinates of the surroundings.

Velodyne's Puck (VLP-16) is a LiDAR with 16 channels that give a 360° three-dimensional view.

4. Simulation with ROS

4.1 ROS in general

Autonomous vehicles can be regarded as four-wheeled robots. In our project, an ROS (Robot Operating System) is used. It is an open-source, flexible framework for writing robot software. The most important components of it are as follows:

Topics Topics are named buses, in which data is exchanged using ROS messages. Each topic has a specific name, moreover, one node publishes data to a topic and another node reads the data from the topic by subscribing to it.

Messages Every topic consists of a type of message, moreover, topics send and receive data in the form of ROS messages. ROS messages form a data structure used by ROS nodes to exchange data. Different topics send different types of messages, namely 2D LiDARs use LaserScan's, while 3D LiDARs use Point Cloud's.

Rosbags Bags are a useful utility for the recording and playback of ROS topics. While working on autonomous vehicles, some situations may arise where it is necessary to work without actual hardware. Using rosbags, sensor data can be recorded and bag files copied to other computers to inspect data by playing it back.

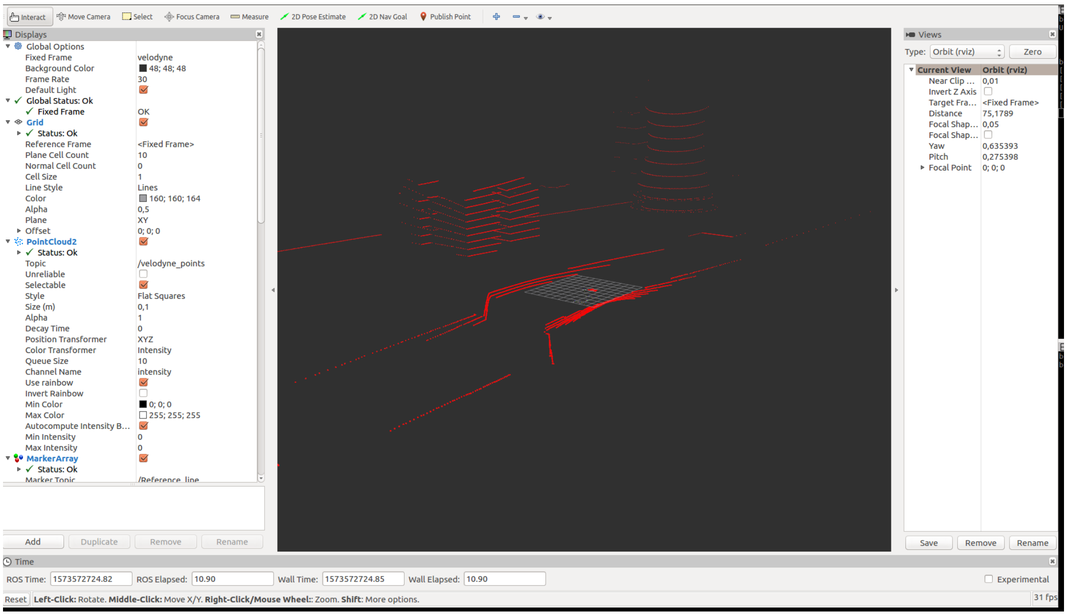
Gazebo Gazebo stands for open source robotic simulators tightly integrated with ROS. In this environment, various types of robots, indoor sensors and outdoor elements could be implemented. The values of sensors can be accessed by the ROS through topics.

Rviz Rviz is a 3D visualizer in ROS to visualize 2D and 3D values from ROS topics and parameters which helps to visualize data such as robot models, robot 3D transform data (TF), point clouds, laser and image data, as well as a variety of sensor data [2].

4.2 Simulation

Simulations are based on models with which we try to copy the physics and geometry of the real world. The more factors that are built into them, the more likely what is expected to happen will actually occur in a real-world test.

Our car acts as a robot, modelled on a CAD program. In this phase of the algorithm, the layout of the car is not important so a Nissan Leaf was used as our race car. In Gazebo, numerous types of sensors are available, the parameters of which can be defined as real-world LiDAR signals as mentioned in Section 3. Unfortunately, the exact layout of the racetrack is unknown, so one was created. Our track consists of straight sections, curves with



Global Options	
Fixed Frame	velodyne
Background Color	■ 48; 48; 48
Frame Rate	30
Default Light	☑
✓ Global Status: Ok	
✓ Fixed Frame	OK
Grid	☑
✓ Status: Ok	
Reference Frame	<Fixed Frame>
Plane Cell Count	10
Normal Cell Count	0
Cell Size	1
Line Style	Lines
Color	▨ 160; 160; 164
Alpha	0,5
Plane	XY
Offset	0; 0; 0
PointCloud2	☑
✓ Status: Ok	
Topic	/velodyne_points
Unreliable	□
Selectable	☑
Style	Flat Squares
Size (m)	0,1
Alpha	1
Decay Time	0
Position Transformer	XYZ
Color Transformer	Intensity
Queue Size	10
Channel Name	intensity
Use rainbow	☑
Invert Rainbow	□
Min Color	■ 0; 0; 0
Max Color	□ 255; 255; 255
Autocompute Intensity B...	☑
Min Intensity	0
Max Intensity	0
MarkerArray	☑
✓ Status: Ok	
Marker Topic	/Reference_line

Type: Orbit (rviz)

Current View	Orbit (rviz)
Near Clip ...	0,01
Invert Z Axis	□
Target Fra...	<Fixed Frame>
Distance	75,1789
Focal Shap...	0,05
Focal Shap...	□
Yaw	0,635393
Pitch	0,275398
Focal Point	0; 0; 0

ROS Time: 1573572724.82 ROS Elapsed: 10.90 Wall Time: 1573572724.85 Wall Elapsed: 10.90 □ Experimental

Left-Click: Rotate. **Middle-Click:** Move X/Y. **Right-Click/Mouse Wheel:**: Zoom. **Shift:** More options. 31 fps

Figure 3: Simulation in Rviz. Different topics can be visualized at the same time. It can be seen that the 3D LiDAR point cloud originates from two 2D beams, one above the other.

different radii as well as white and red protective barriers on both sides.

As 3D sensors are used, not only the track but its surroundings are also important. Therefore, buildings, trees in addition to cones between and beyond the two barriers were implemented. When the simulation was completed, it was possible to drive a lap of the track in the simulation and collect the ROS topics data via rosbags. Moreover, with Rviz, the following points could be visualized as presented in Fig. 3.

5. Steps of the algorithm

5.1 Filtering

Our LiDAR sensor gives information over 360 degrees which is very useful in many cases, however, in this challenge, only information about points in front of the car is needed. If information about displacement of the sensors on the car is available, it can be assumed that the front of the car and all points smaller than the linear equation

$$x = \frac{-b + y}{m} \tag{3}$$

will be deleted, where b denotes the y-intercept and m represents the gradient.

5.2 Clustering

As from our simulation only raw data (only x, y and z coordinates) are obtained, which points belong to each object has to be defined. Clustering is the task of dividing the population or data points into a number of groups such that the data points in a group are more similar to other data points in the same group than to data points in other groups. It is basically a collection of objects arranged according to how similar or dissimilar they are to each other.

For us the most effective method is DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a well-known data clustering algorithm that is commonly used in data mining and machine learning. Based on a set of points, DBSCAN groups together points that are in close proximity to each other based on a distance measurement (usually the Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions [3]. The DBSCAN algorithm basically requires two parameters:

eps specifies how close points should be to each other to be considered as parts of a cluster. This means that if the distance between two points is smaller or equal to this value (eps), these points are considered to be neighbors.

minPoints the minimum number of points to form a dense region. For example, if the minPoints parameter is set as 5, then at least 5 points are needed to form a dense region.

After clustering, all of the points belong to a group as shown in Fig. 4.

5.3 Convex hull

Even though the point cloud of a 3D LiDAR stands for more than one thousand points, only a proportion of them is needed, thus points should be filtered. The clustered points have to be packed by each group into a convex hull. The convex hull of a set of points is defined as the

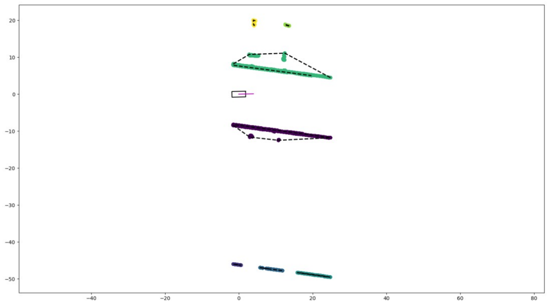


Figure 4: The different clusters denoted by different colors.

smallest convex polygon that encloses all of the points in the set [4].

For the purposes of clarification, only the vertices of a convex polygon are used as shown in Fig. 5. Vertices are side points which form the simplified shape of a polygon.

5.4 Delaunay triangulation and Voronoi diagram

The Delaunay triangulation for a given set P of discrete points in a plane is a triangulation $\mathrm{DT}(P)$ such that no point in P is inside the circumcircle of any triangle in $\mathrm{DT}(P)$ [5]. Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation as presented in Fig. 6.

The circumcenters of Delaunay triangles are the vertices of the Voronoi diagram. In the 2D case, the Voronoi vertices are connected via edges that can be derived from adjacency relations of the Delaunay triangles. If two triangles share an edge in the Delaunay triangulation, their circumcenters are to be connected with an edge in the Voronoi tessellation as shown in Fig. 7.

For a set of points in 2D space, the Voronoi diagram creates cells whose edges are the same distance from two neighboring points. This property grants that all reference points are exactly between two detected items and the edges of Voronoi cells are ideal for the representation of a reference line. Unfortunately, real environments are never ideal and this method would provide more route options in the presence of noise as presented in Fig. 8.

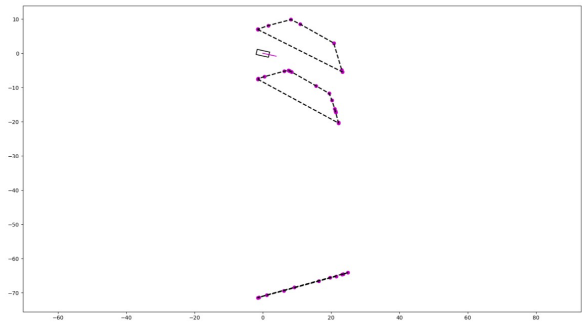


Figure 5: To clarify the large number of points, only the vertices of the convex hull were used.

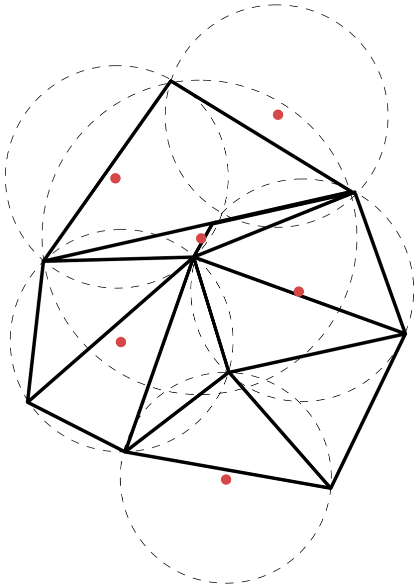


Figure 6: The Delaunay triangulation with all the circumcircles and their centers (marked in red).

5.5 RANSAC

At this stage of the algorithm, no information is available about which side of the car the detected points are on. As the reference lines should only be between the two middle barriers, this information is important. For this process, an estimator must be implemented.

RANSAC (random sample consensus) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers which do not influence the values of the estimated parameters. The input to the RANSAC algorithm is a set of observed data values, a method of fitting some kind of model to the observations and several confidence parameters. RANSAC achieves its goal by repeating the following steps [6]:

1. Selecting a random subset of the original data, re-

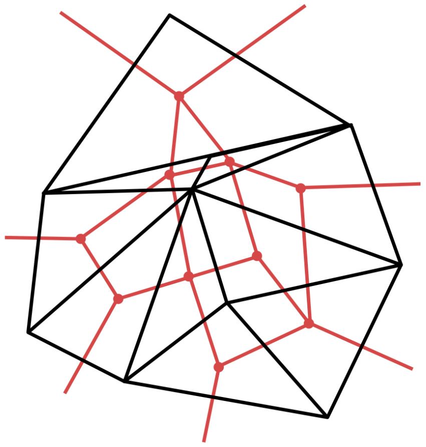


Figure 7: A Voronoi diagram is constructed by connecting the centers of the circumstances.

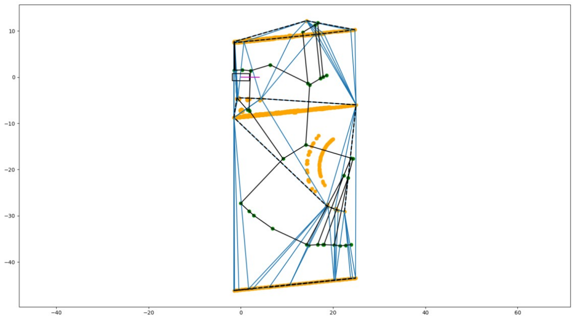


Figure 8: A real scenario from the algorithm. The detected points are denoted in orange, the green points represent the centers of the circumstances, and the black lines stand for the Voronoi edges which also represent the reference lines.

ferred to as the hypothetical inliers;

2. Fitting a model to the set of hypothetical inliers;
3. Then testing all other data against the fitted model. Those points that fit the estimated model well, according to a model-specific loss function, are considered to be part of the consensus set.
4. The estimated model is reasonably good if a sufficient number of points are classified as part of the consensus set.
5. Finally, the model may be improved by re-estimating it using all members of the consensus set.

If RANSAC is used for the set of Voronoi midpoints, a line is obtained between the protective barriers on either side of the track (Fig. 9). Now the linear equation for y with its gradient can be written:

$$y = mx + b, \tag{4}$$

where b denotes the y-intercept and m represents the gradient.

Points with higher and lower values of y belong to the left- and right-hand sides, respectively.

As RANSAC is an estimator, sometimes it yields wrong results as shown in Fig. 10.

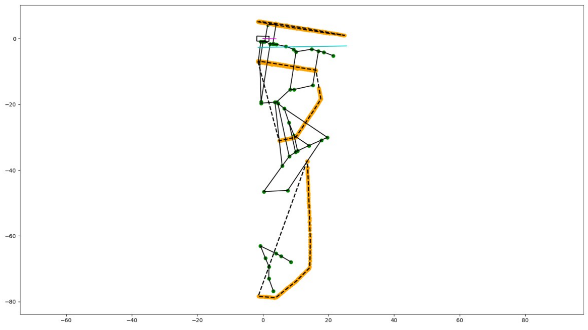


Figure 9: The RANSAC line (denoted in blue) splits the space into two parts.

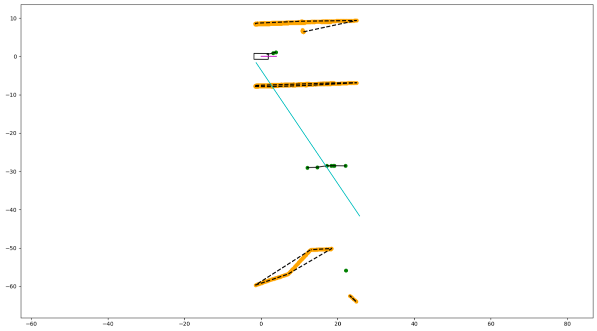


Figure 10: The estimator sometimes yields wrong results.

To avoid producing wrong results, some criteria must be declared:

- The starting point of the line should be between the protective barriers.
- The line must be sufficiently long over the full range of the x-axis.
- If the RANSAC line intersects any of the lines of the convex hull, it must be rotated until it no longer intersects them.

5.6 Delete unnecessary reference lines

One of our main criteria is that the reference line should exclusively be between the two barriers. To ensure that our car moves in such a way, all unnecessary Voronoi midpoints must be eliminated. Now all the points that belong to each side can be collected, however, since they are unsorted, an Andrew's monotone chain convex hull algorithm was used [7]. This algorithm sorted the points into a lexicographical order (first by x-coordinates, and should any be equal, by y-coordinates) and then constructed upper and lower hulls of the points as seen in Fig. 11.

Some naive approaches concerning how to use the upper and lower sides of the convex hulls exist. Firstly, an attempt was made to copy the shape of the upper hull with a single line using Equation (Eq. 4), but was not possible when used at corners.

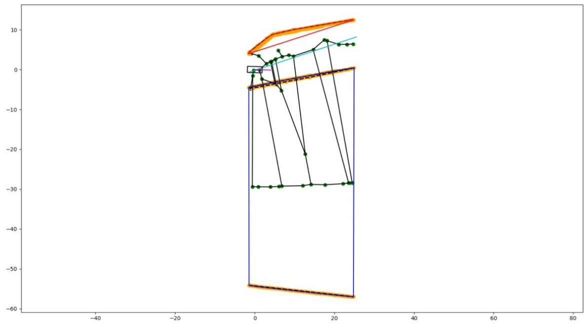


Figure 11: The left- and right-hand sides were packed into another convex hull.

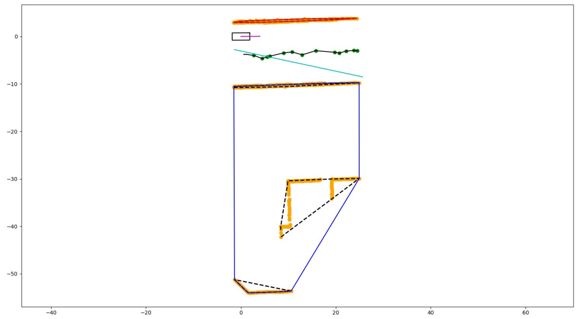


Figure 12: The reference line along straight sections affected by noise.

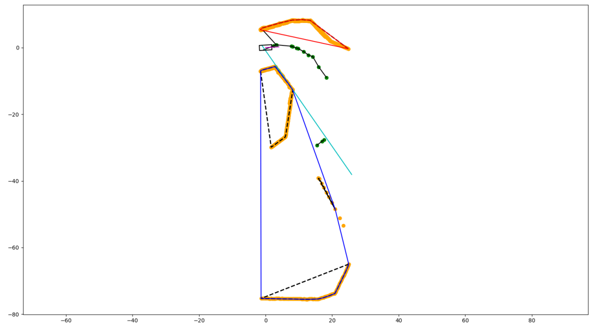


Figure 13: The reference line at corners.

Another idea was that since the borderline was not a single straight line, it could be divided into as many sections as the number of vertices found in the upper hull, however, this resulted in a wavy line which in some cases also deleted inlier Voronoi points. Then it was realized that the easiest solution would be to delete all the Voronoi midpoints which lie inside the left and right convex hulls.

6. Results and future works

By following the aforementioned steps, the algorithm drew an unambiguous reference line by connecting Voronoi midpoints which lie in the middle of the road of the required territory. Furthermore, this method also works along straight sections and at corners as shown in Figs. 12 and 13.

For safety reasons, vehicle speed is limited to 25 km/h. Our algorithm is designed to operate at this speed range. The algorithm is based on Voronoi cells so its complexity is $O(n)$ [8]. In our test environment the algorithm could estimate the free space with 90 % confidence.

With the aid of Voronoi diagrams, every detected object has an effect on our reference line. Fortunately, the majority of these points are filtered by the steps outlined in Section 5.6, however, outliers between the barriers cannot be dealt with. These points have a detrimental effect on the unambiguity of the reference line and lead to the creation of several nodes resulting in some multiple-choice decisions that need to be made by our car. To make our algorithm robust, these points should be eliminated.

Now a single RANSAC line divides the sides of the road and works well for the layouts of simple corners, however, one simple line cannot satisfy the criteria for more difficult layouts, e.g. chicanes, because the RANSAC line is unable to intersect any lines of the convex hull. To avoid this, the line should be divided into smaller sections that need to follow on from each other and determine if any parts of the linked line intersect the line of the convex hull.

Our path-following algorithm requires that the reference points which follow on from each other are the same distance apart. For this purpose, our reference points must be joined into one line and as many points as the algorithm requires must be generated.

This algorithm has only been examined using test data and has yet to be tested in real environments.

7. Conclusion

In 2020, our team will participate in the autonomous distance challenge of the Shell Eco-marathon. The results of this algorithm are promising but a significant amount of research and development has yet to be done until this algorithm can operate and lead our car around a single lap. One of the benefits of the proposed method is that it involves widely used and, over time, highly improved geometric algorithms like Delaunay triangulation and Voronoi diagrams.

Acknowledgements

The research was carried out as part of the “Autonomous Vehicle Systems Research related to the Autonomous Vehicle Proving Ground of Zalaegerszeg (EFOP-3.6.2-16-2017-00002)” project in the framework of the New Széchenyi Plan. The completion of this project is funded by the European Union and co-financed by the European Social Fund.

REFERENCES

[1] Verőné Wojtaszek, M.: Fotointerpretáció és távérzékelés 3., A lézer alapú távérzékelés. 2010 Nyugat-magyarországi Egyetem

[2] Lentin, J.: ROS robotic projects. (Packt Publishing Ltd., Birmingham, UK) 2017, ISBN: 978-1-783-55471-3

[3] Kriegel, H.-P.; Kröger, P.; Sander, J.; Zimek, A.: Density-based clustering, *WIREs Data Mining Knowl. Discov.*, 2011, **1** 231–240 DOI: 10.1002/widm.30

[4] Chazelle, B.: An optimal convex hull algorithm in any fixed dimension, *Discrete Comput. Geom.*, 1993, **10** 377–409 DOI: 10.1007/BF02573985

[5] de Berg, M.; Cheong, O.; van Kreveld, M.; Overmars, M.: Computational geometry: Algorithms and applications (Springer-Verlag, Berlin, Germany) 2008, DOI: 10.1007/978-3-540-77974-2

[6] Fischler, M. A.; Bolles, R. C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Commun. ACM*, 1981, **24**(6) 381–395 DOI: 10.1145/358669.358692

[7] Andrew, A. M.: Another efficient algorithm for convex hulls in two dimensions, *Inf. Process. Lett.*, 1979, **9**(5) 216–219 DOI: 10.1016/0020-0190(79)90072-3

[8] Seidel, R.; Adamy, U.; On the exaxt worst case query complexity of planar point location, *J. Algorithms*, 2000, **37** 189–217 DOI: 10.1006/jagm.2000.1101

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 33–41 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-06

IDENTIFICATION OF THE MATERIAL PROPERTIES OF AN 18650 LI-ION BATTERY FOR IMPROVING THE ELECTROCHEMICAL MODEL USED IN CELL TESTING

BENCE CSOMÓS*[1] AND DÉNES FODOR[1]

[1]Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

The aim of this paper is to present an application of the generalized Warburg element and Constant Phase Element (CPE) for non-Fickian diffusion modeling. These distributed elements are intended to provide a better fit of low-frequency impedance data than the standard finite-length Warburg element in the case of most batteries. In addition, the current study demonstrates the ambiguity of the finite-length Warburg element if impedance data is insufficient within the very-low-frequency impedance spectrum. In order to select the appropriate Randles circuit for non-Fickian diffusion modeling, several configurations have been investigated. Based on the best fit of impedance data, the State-of-Charge (SoC) dependency of the Randles circuit parameters has also been analyzed. This study concerns a Samsung ICR18650-26F 2600 mAh battery cell which was subjected to Electrochemical Impedance Spectroscopy (EIS) measurements between 10 mHz and 100 kHz as a function of SoC. The results were plotted and compared in the form of Nyquist plots. The Randles circuit parameters such as the resistances R_s and R_ct, double-layer C_dl, leaky capacitance CPE and Warburg coefficients were estimated using ZView software. The present paper shows that CPE – and its QPE form – is a recommended choice to yield the best fit in terms of non-Fickian diffusion impedance. In addition, using CPE is a better alternative to avoid problems with initial values and multiple local solutions, which may exist in the case of the Warburg element. The resultant Randles circuit parameters and their SoC characteristics can be effectively used in further electrochemical modeling.

Keywords: Li-Ion Battery, Electrochemistry, Material, Battery model, Parameter estimation

1. Introduction

The State of Health (SoH) of a battery plays an important role in electric applications since it has a great influence on the available capacity and power of a battery [1]. SoH deteriorates with battery usage and the rate of aging is related to the operating history of the battery. Therefore, it is recommended to track the state variables of a cell throughout its life cycle and adapt the SoH prediction according to the current condition of the battery cell.

A common and reliable electrochemical model used in Finite Element Analysis (FEA) is based on the work of Newman et al. [2]. It consists of charge and mass balance equations in both solid (electrode) and liquid (electrolyte) materials, which describe the main operating characteristics of the cell. Even though these formulae could describe the behavior of the cell in 3D, due to their high degree of nonlinearity and complexity, Pseudo-2D (P2D) modeling is favorable in terms of FEA [3]. It is also sufficiently representative to model wearout for automotive applications [4]. In Fig. 1, a typical Pseudo-2D structure of a Li-Ion cell can be seen.

Since battery manufacturing is still a developing sector, the electrochemical composition of a cell constantly changes. Therefore, few battery-chemistry standards and complete databases can describe a given cell structure. Insufficiently reliable and valid battery data inhibits battery modeling since a cell must always be inspected to determine its electrochemical parameters before modeling. The standard way to obtain these electrochemical data is usually through an equivalent circuit modeling process with which the electrochemical properties of the cell can be extracted from EIS measurements.

2. Diffusion modeling techniques

It is possible to calculate battery-specific data using several techniques, which can basically be grouped into two types in terms of the measurement approach applied:

- direct measurements, which typically require disassembly of the cell, special preparations or an experimental open-cell. These measurements can be, for example, different types of Electron Microscopy (EM), Computed Tomography (CT), titration, post-mortem analysis, etc.

*Correspondence: csomos.bence@gmail.com

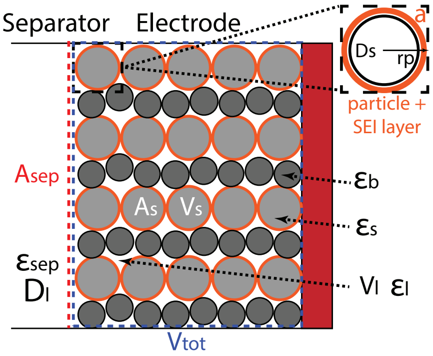


Figure 1: The relationship between the core material parameters and components of the cell. A_{sep} is addressed to the area of the separator. $\varepsilon_{\mathrm{sep}}$, ε_{b}, ε_{s}, and ε_{l} denote the porosities of the separator, binder, solid matrix and void fraction, respectively. D_{l} represents the salt diffusion coefficient in the electrolyte. D_{s} stands for the diffusion coefficient in the solid electrode. V_{l}, V_{s} and V_{tot} are the volumes of the liquid, solid material and whole electrode, respectively. r_{p} and a denote the average radius of each electrode particle and the specific surface area of the electrode.

- indirect measurements that do not require disassembly of the cell, e.g. current impulse excitation, Electrochemical Impedance Spectroscopy (EIS), galvanometry, potentiometry, chronoamperometry, etc. [5]

EIS is a well-established and suitable method in the analysis with regard to battery kinetics and has a solid background in the literature [6]. Another advantage of EIS is that it does not require special preparation of the cell that would be extortionate and time-consuming.

Electrochemical parameters are formulated from EIS data in the form of resistive, capacitive, inductive or distributed elements such as the Constant Phase Element (CPE) or Warburg element.

2.1 Standard equivalent circuits

In order to obtain battery-specific data, a Transmission-Line Model (TLM) was applied that is introduced and expounded on in [7]. It provides a generalized modeling solution for transport processes in porous electrodes by utilizing a finite number of serially connected resistor-capacitor (RC) pairs in parallel, which resolve the ion and electron transport appropriately. The network obtained in this way is considered to be ambiguous since various arrangements can be reduced to the same circuit resulting in identical resistance and capacitance of the circuit [8]. Although the number of RC pairs used can increase the resolution of the transport process and thus improve its accuracy, the increase in the number of elements results in stability as well as local and global optima problems due to equivocality. Since this is a major disadvantage of extended TLM networks, they are usually avoided.

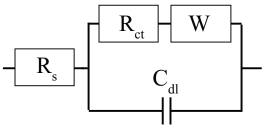


Figure 2: Randles equivalent circuit consisting of serial resistance R_{s}, charge-transfer resistance R_{ct}, double-layer capacity C_{dl} and distributed impedance element W.

A simplified form of TLM is the Randles equivalent circuit model (Fig. 2) which can be obtained if the theoretical RC line of infinite length is reduced to a single Z-distributed element. In other words, the Z-distributed element models the limiting case of TLM. Its fundamental principles and mathematical background are comprehensively described in Barsoukov's book [9]. The standard Randles circuit model couples together the individual characteristics of electrodes and electrolytes into one corresponding circuit element. The Randles circuit model consists of relaxation components such as the serial resistance of the electrolyte R_{s}, charge transfer resistance R_{ct}, a double-layer capacitor C_{dl} and distributed impedance for modeling the low-frequency diffusive behavior of the cell. Usually, the double-layer effect exhibits non-ideal capacitive behavior so a CPE – the equivalent to a "leaky" capacitor – should be used instead of a C_{dl}. All of these detailed parameters of both electrodes are grouped together, thus they represent the combined behavior of the two electrodes.

Ion transport that occurs in the low-frequency bandwidth of the impedance spectrum can involve diffusion, migration and convection, however, only diffusion is of interest because distributed elements are responsible for diffusion modeling. Diffusion modeling is the bottleneck of model development since it is related to a complex and interleaved electrochemical process that mainly characterizes the long-term operation of the battery. Ion transport can be, by and large, divided into two parts, namely lithium diffusion in the solid active electrode material and electrolyte. The "tail" at the beginning of the low-frequency bandwidth in the Nyquist plot shows diffusion in the electrolyte while the transition from the tail through an arc to a straightened end of the impedance curve depicts the solid-phase diffusion inside the electrode matrix (Fig. 3). In fact, the cathode is composed of a poor ion-conducting material. A general rule of thumb is that its diffusion coefficient is $2-4$ orders of magnitude smaller than the diffusion coefficient of lithium ions in the electrolyte, that is, the salt diffusion coefficient of the electrolyte D_{l} falls within the range of $10^{-10}-10^{-11}$ m^2/s while the diffusion coefficient of lithium ions D_{s} in the solid matrix falls between 10^{-13} and 10^{-14} m^2/s. In other words, the time constant for diffusion of Li-Ion transport is smaller in the electrolyte ($\sim 10-100$ s) than in the solid matrix (> 100 s). Consequently, the two types of diffusion need to be separated and should be modeled based on their different characteristic impedances.

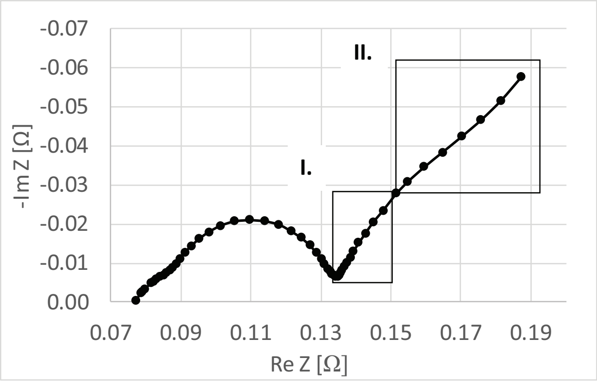


Figure 3: A typical Nyquist plot of a Li-Ion cell; I. denotes a "tail" that mainly represents diffusion in the electrolyte; II. shows a transition from liquid-phase to solid-phase diffusion.

Generally, diffusion is modeled by the classical Warburg impedance in accordance with the following assumptions: diffusion is Fickian (planar diffusion); the electrolyte is supporting, symmetric and binary; the cell remains in a quasi-equilibrium state during excitation; and no reaction occurs in the bulk of the electrolyte. Under these premises, the standard Warburg impedance has an exponent of 0.5 that implies its $45°$ phase angle. If diffusion occurs in an infinite reservoir where the concentration can decrease to zero, infinite-length Warburg impedance can be assumed, otherwise diffusion is restricted and finite-length reflective or transmissive Warburg impedance can be assumed depending on whether the equivalent circuit is terminated by an open circuit or a resistor, respectively. The former and latter cases can be mathematically expressed by extending the standard infinite-length Warburg impedance with the hyperbolic functions tanh and coth, respectively. All three types of Warburg elements exhibit the same $45°$ gradient at the beginning of the low-frequency bandwidth in the Nyquist plot.

In terms of impedance, using finite-length Warburg elements is unsuitable if an insufficient number of data points in the low-frequency bandwidth of the impedance spectrum are available to fit the hyperbolic functions well. This occurs when the EIS measurements typically run up until 10 mHz but some cells do not show a clear and distinct effect of diffusion in the solid phase. In this case, only the tail part of the impedance spectra can be reasonably modeled. Due to a lack of low-frequency data points, the finite-length Warburg elements cannot be effectively applied. On the other hand, the tail part of the diffusion impedance can be modeled by CPE according to [11, 12] which is a similar but more robust alternative to the Warburg elements.

All the transfer functions of the standard types of distributed elements are summarized in Table 1. The transfer function of CPE can be expressed in two different forms according to the position of its time constant for diffusion τ_D . If τ_D is emphasized from its square root, it is referred to as QPE, which sometimes provides a more stable regression than CPE.

Table 1: Transfer functions of standard distributed elements used in Fickian diffusion modeling [10]. σ denotes the Warburg coefficient, ω represents the excitation frequency, τ_D stands for the diffusion time constant, Q is the QPE time constant, j denotes the imaginary unit and R_w represents the Warburg resistance.

Name	Impedance
infinite-length Warburg	$Z_{\mathrm{ilw}}(j\omega)=\frac{R_{\mathrm{w}}}{\sqrt{j\omega\tau_{\mathrm{D}}}}=\sigma\frac{1}{\sqrt{\omega}}(1-j)$
finite-length reflective Warburg	$Z_{\mathrm{flrw}}(\omega)=R_{\mathrm{w}}\coth\frac{(\sqrt{j\omega\tau_{\mathrm{D}}})}{\sqrt{j\omega\tau_{\mathrm{D}}}}$
finite-length transmissive Warburg	$Z_{\mathrm{fltw}}(\omega)=R_{\mathrm{w}}\tanh\frac{(\sqrt{j\omega\tau_{\mathrm{D}}})}{\sqrt{j\omega\tau_{\mathrm{D}}}}$
CPE	$Z_{\mathrm{CPE}}(\omega)=\frac{1}{\sqrt{j\omega\tau_{\mathrm{D}}}}$
QPE	$Z_{\mathrm{QPE}}(\omega)=\frac{1}{Q\sqrt{j\omega}}$

Up to this point, only Fickian diffusion has been considered that can be clearly identified by its square-root-like frequency dependency in terms of the transfer function of the distributed elements. However, the impedance fit becomes more interesting when the cell impedance exhibits non-Fickian behavior, that is, the phase angle is not $45°$ due to diffusion non-idealities. These phenomena were investigated and are mostly related to multi-phase and multi-scale diffusion in porous electrodes [13, 14], diffusion coupled with migration [15, 16] and/or diffusion in non-conventional space [17–19]. Non-Fickian diffusion can be treated by fractional order circuits which consist of various types of typical configurations. In the literature [20], Warburg impedance is referred to as "generalized" if it reflects the fractional intent. In this case, the Warburg exponent is generalized and denoted by γ, the dispersion parameter.

2.2 Equivalent circuit development for modeling non-Fickian diffusion

In order to obtain the most accurate fit of the impedance curves, several configurations of the Randles circuit were developed. These setups are presented in Fig. 4. The key differences between them can be explored in terms of both the position and type of the distributed elements in the circuit. In some papers from the literature, the distributed element is in series with R_{ct} [21], while in others, it is placed in series with the parallel R_{ct}-C_{dl} or R_{ct}-CPE branch. The distributed element can be either the Warburg element or CPE/QPE. Birkl et al. [22] shows that it is possible to model diffusion with an RC pair instead of a distributed element in the Randles circuit model. All of these configurations have been exhaustively expounded

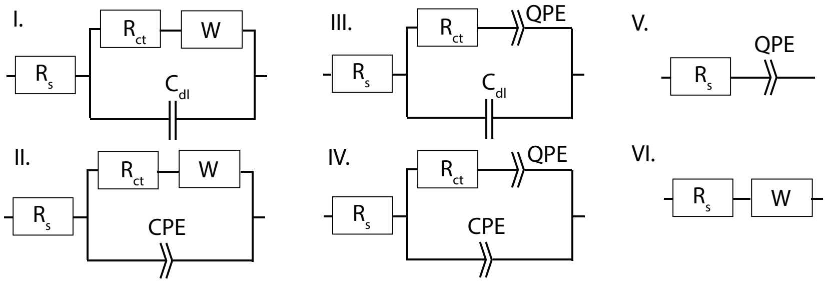


Figure 4: Different configurations of Randles circuits to study regression performance and fitness. W stands for Warburg element. The model of circuits I-IV, relaxation and diffusion are presented together while V-VI only focus on the tail part.

on in [23].

Based on these results, numerous Randles circuits have been evaluated to study the regression performance and fitness. In Fig. 4, the arrangements of circuits I-IV model relaxation and diffusion simultaneously while V and VI only account for the tail part. The main reason for separation is that the stability and robustness of the impedance regression could be increased using this technique.

The standard Warburg element and CPE/QPE had to be adjusted to match with the non-Fickian diffusion. At first, since only the tail part of the diffusion impedance was modeled, the hyperbolic part of the finite-length Warburg elements was neglected. Hence, the Warburg impedance was simplified to the infinite-length form. Therefore, the Warburg impedance Z_w had to be transformed into a generalized form by replacing the square root in the denominator with γ. As a result, the generalized Warburg impedance could be written in the following form:

$$Z_\mathrm{w}(\omega) = \frac{R_\mathrm{w}}{(j\omega\tau_\mathrm{D})^\gamma} \tag{1}$$

where R_w denotes the Warburg resistance, ω stands for the excitation frequency, τ_D represents the diffusion time constant, j is the imaginary unit and $0 \leq \gamma \leq 1$. Now, the j^γ term should be practically separated into real and imaginary components to reveal the contribution of γ to each part. Using Euler's formula, Z_w was unbundled and grouped into real and imaginary parts:

$$Z_\mathrm{w}(\omega) = \frac{R_\mathrm{w}}{\tau_\mathrm{D}^\gamma \omega^\gamma} \cos\left(\frac{\pi}{2}\gamma\right) - \frac{R_\mathrm{w}}{\tau_\mathrm{D}^\gamma \omega^\gamma} j \sin\left(\frac{\pi}{2}\gamma\right) \tag{2}$$

τ_D used in Eq. 2 was then expressed as

$$\tau_\mathrm{D} = \frac{L_\mathrm{eff}^{1/\gamma}}{D_\mathrm{eff}} = \frac{\epsilon_\mathrm{1,sep}^{\beta/\gamma} L_0^{1/\gamma}}{\epsilon_\mathrm{1,sep}^{\beta} D_{1,0}} \tag{3}$$

where L_eff denotes the effective diffusion length, D_eff represents the effective diffusion coefficient, $\epsilon_\mathrm{1,sep}$ stands for the liquid fraction in the separator, and β is the Bruggeman coefficient. Since τ_D is emphasized from the denominator of Eq. 1, the Warburg coefficient σ could be expressed as a fraction of R_w and τ_D according to

$$\sigma = \frac{R_\mathrm{w}}{\tau_\mathrm{D}^\gamma}. \tag{4}$$

On the other hand, the transfer functions of CPE and QPE had to be transformed into

$$Z_\mathrm{CPE}(\omega) = \frac{1}{(j\omega\tau_\mathrm{D})^\gamma} \tag{5}$$

$$Z_\mathrm{QPE}(\omega) = \frac{1}{\tau_\mathrm{D}^\gamma (j\omega)^\gamma} \tag{6}$$

The updated transfer functions of the distributed elements enabled the non-Fickian diffusion to be properly fitted.

3. Experimental setup

This study is devoted to a commercial Samsung ICR 18650-26F cell with a nominal capacity of 2600 mAh. It consists of a double-sided Nickel-Manganese-Cobalt (NMC) cathode and graphite anode according to the manufacturer's datasheet.

The Samsung ICR 18650-26F cell was evaluated by EIS within the 10 mHz – 100 kHz bandwidth at different States-of-Charge (SoC). The test was run at ambient temperature, namely 25 °C, which was considered to be constant throughout. A Solartron SI1287 (Electrochemical Interface) and a Schlumberger SI 1255 (HF Frequency Response Analyzer) were used for data acquisition. The Nyquist plot of the spectrum and the parametric fitting were produced by ZPlot and ZView software, respectively.

An auxiliary measurement had to be performed along with EIS to determine the cell's Open Circuit Potential (OCP) characteristic at a 0.1 C-rated load current. The load current was generated by a TENMA 72-13210 Programmable DC Load in Constant Current (CC) mode. The data was recorded in NI PXI hardware that ran LabVIEW-based data acquisition software.

4. Analysis of the measurement results

The purpose of the EIS analysis of the cell was to evaluate the fitness of the different Randles circuit models according to non-Fickian cell impedance data. On the other hand, the cell underwent another EIS measurement to detect how the parameters of the Randles circuit model changed during discharge. The EIS measurements presented in Fig. 5 and 6 were made between 10 mHz and

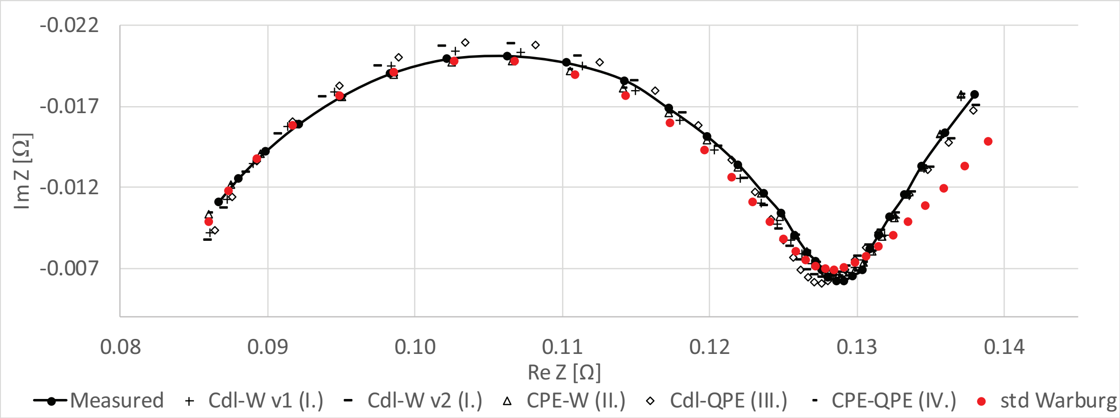


Figure 5: Nyquist plot of the measured impedance and different fit functions (seen in Fig. 4 I-IV). v1 and v2 indicate that Warburg-based models were run on the basis of two different sets of initial parameters. The fit was performed between 13 Hz and 10 mHz at ambient temperature with a fully charged cell. The "std Warburg" was based on model I in Fig. 4 but with a Warburg exponent of 0.5.

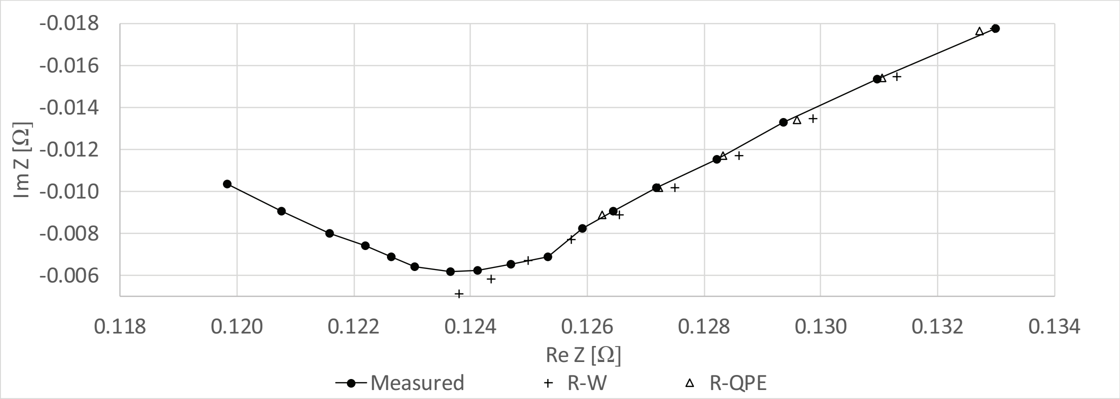


Figure 6: Nyquist plot of the measured impedance and different fit functions (seen in models V-VI of Fig. 4. The fit was performed between 80 mHz and 10 mHz (only the tail part) at ambient temperature with a fully charged cell. The left part of the solid line is associated with the end of the relaxation semicircle and helps to locate the position of the tail.

13 Hz where relaxation and diffusion occur. In the case of R-W and R-QPE pairs, only the tail part was modeled. In Fig. 5, only a slight difference between the fits of the model is observed and the CPE-QPE pair shows the best match. Despite the insignificant difference between the non-Fickian models, a substantial improvement in accuracy can be observed especially with regard to fitting the tail when the standard Warburg element was replaced by any of the generalized Warburg- or CPE/QPE-based models.

The changes in the Randles circuit parameters during discharge of the cell was also investigated further where only the best matching CPE-QPE pair was used for fitting. The resultant impedance plot is presented in Fig. 7. The EIS measurements were performed at 20 % SoC level increments. Every step of the discharge was followed by a 12-hour-long period of relaxation before the EIS measurement was made in order to provide sufficient time for the cell to reach its stationary state. The impedance curves show that the degree of relaxation was high when the rate of the electrode reaction decreased or if the transport of Li ions became limited. This occurred, for example, as the rate of Li diffusion decreased away from the interface. The evolution of changes in the distinctive impedance curves could be tracked by changes in the parameters of the Randles circuits. Fig. 7 shows that R_s and R_ct significantly increased as SoC decreased due to the decreasing amount of Li-Ion particles engaged in the charge transfer process. Furthermore, the electrolyte resistance R_s increased due to the decreasing ionic conductivity of the solution. All of these results agree with a well-known phenomenon, namely that the overall resistance of the cell increases during discharge. The shape and position of the plateau at the beginning of the semicircle was apparently due to a Solid Electrolyte Interphase (SEI) layer which formed on the anode particles that was unaffected during discharge. The presence of an SEI layer ascertains that the calendar and cycle lives of the cell both reduced. Since the gradient of the tails show significant similarities between 100 % and 20 % SoC, the diffusion time constant and diffusion coefficient of the electrolyte should change slightly during discharge. At 5 % SoC, the

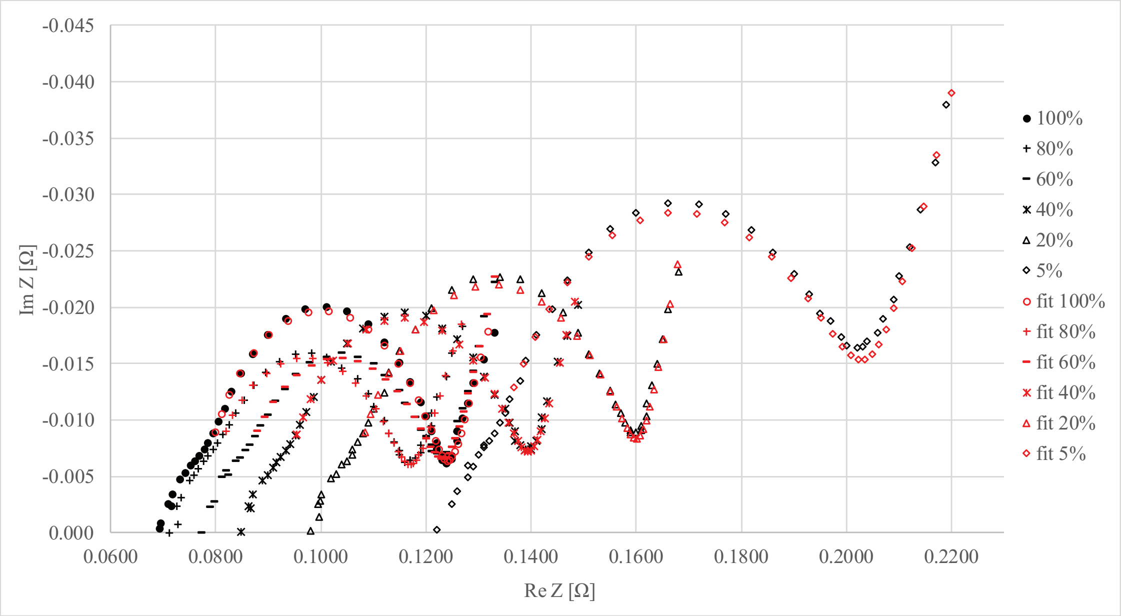


Figure 7: Nyquist plots of the Samsung ICR 18650-26F 2600 mAh Li-Ion cell at different levels of SoC. The temperature was assumed to be constant at 25 °C. The regression bandwidth was limited to between 13 Hz and 10 mHz. The small plateau at approximately 250 Hz is a consequence of a Solid Electrolyte Interphase (SEI) layer that formed on the anode particles. This shows that the calendar and cycle lives of the cell were reduced. The fit was made by the model of the CPE-QPE pair.

utilizable Li-Ions in the electrode became exhausted leading to a significant decrease in both the rate of diffusion and reaction. On the other hand, the reaction rates did not vary extensively in the normal operating region since the “valley” between the semicircle and tail possesses a similar imaginary component of the impedance.

In order to carry out cell characterization in the time domain, a quasi-equilibrium discharge was run using a 0.1 C-rated load current at a constant temperature of 25 °C. The OCV against SoC curve is presented in Fig. 8 that exhibits typical discharge characteristics with a small plateau around 30 % SoC and rapidly decreases below 10 % SoC.

4.1 Determining Randles circuit parameters

Evaluation of EIS data in Fig. 5 and the impedance regression were carried out by ZView software that applies the non-linear least squares method to fit and calculate the Randles circuit parameters. The results summarized in Table 2 show that slight changes in R_s, R_ct and C_dl were observed with this setup. This was due to fitting on an almost ideal semicircle that exhibits a simple RC characteristic in the impedance spectrum. With regard to the estimation of diffusion time constants, diffusion in a real electrochemical battery cell is usually limited due to the relatively thin electrodes. Consequently, from a practical point of view, ZView only has finite-length Warburg elements at its disposal. Since finite-length Warburg elements should require data from the very low bandwidth that is unavailable in the present case, it is favorable to check the applicability of this type of element in the current case.

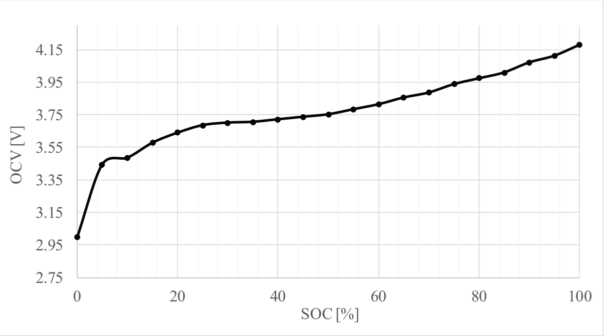


Figure 8: The Open Circuit Voltage (OCV) against State-of-Charge (SoC) characteristic of the cell.

For this purpose, the Warburg parameters were estimated on the basis of two different sets of initial values as is denoted by v1 and v2 in Fig. 5. Both cases yielded a similar fit but extremely different Warburg parameters. Therefore, the estimation of Warburg parameters ran into multiple local solutions that erroneously characterize the same system with different diffusion time constants. This problem could be efficiently handled by using either the CPE+QPE pair or just QPE instead. The results are presented in Table 3. Given the γ_w values, the Warburg exponent is clearly far from 0.5, hence the preliminary assumption of exhibiting non-Fickian diffusion was substantiated. The maximum phase error between the standard Warburg-element- and CPE+QPE pair-based Randles circuit models was 2.3 % at 10 mHz. Since Randles circuit parameters are very sensitive due to the origin of their exponential functions, even a small improvement in

Table 2: The estimated Randles parameters where diffusion was only modeled by a Warburg element and a QPE in two different configurations. The QPE+CPE modeling technique was an effective way to avoid the uncertainties of finite-length Warburg elements due to a lack of data points in the very low bandwidth. The cell was fully charged and kept at 25 °C.

Randles parameter	QPE+CPE	W	QPE
R_s [Ω]	0.154	0.1968	0.1967
R_ct [Ω]	0.047	-	-
C_dl [F]	-	-	-
T_CPE [s]	2.038	-	-
γ_CPE [-]	0.893	-	-
γ_QPE [-]	0.6868	-	0.595
R_w [Ωm^2]	-	0.241	-
τ_w [s]	-	893	235.7
γ_w [-]	-	0.595	-
τ_D [s$^{0.5}$/Ωm^2]	339.5	893	235.7
σ [Ωm^2/s$^{0.5}$]	0.0015	0.00424	0.0021

Table 3: The estimated Randles parameters based on the configurations seen in Fig. 4 where diffusion and double-layer effect have been modeled by Warburg-element, QPE and CPE in different coupled configurations. In the table heading, W and C stand for Warburg and double-layer capacitor, respectively. The cell has been at fully charged state and kept at 25 °C.

Randles	C+W	C+W	C+QPE	CPE+W
parameter	I	II		
R_s [Ω]	0.156	0.156	0.156	0.153
R_ct [Ω]	0.041	0.041	0.041	0.047
C_dl [F]	1.731	1.731	1.73	-
T_CPE [s]	-	-	-	2.038
γ [-]	-	-	0.579	0.893
R_w [Ωm^2]	0.0855	0.168	-	0.263
τ_w [s]	179	569.1	-	693.2
γ_w [-]	0.578	0.5792	-	0.687
τ_D [s$^{0.5}$/Ωm^2]	179	569.1	233.9	693.2
σ [Ωm^2/s$^{0.5}$]	0.0043	0.0043	0.0023	0.0029

fitting can significantly increase the accuracy in further electrochemical calculations based on these data.

The Randles circuit parameters were measured when the cell was fully charged but changed as the SoC level of the cell was altered as was seen in Fig. 7. The estimated parameters at different SoC levels are summarized in Table 4 and their trends presented in Fig. 9. The changes in R_s and R_ct exhibited an exponential-like decreasing tendency against SoC while R_ct slightly increased at approximately 100 % SoC. The increase in R_s was due to the effect of a reduction in the ionic conductivity, while an increase in R_ct was due to the decreasing rate of Li-Ion transfer through the electrode-electrolyte interface. In Fig. 10, the characteristics of changes in the CPE capacity and diffusion time constant can be seen. CPE slightly increased with discharge and its overall variance was about 0.4 F. This behavior along with the increase in R_ct can be attributed to the relaxation effect. Furthermore, less charge was available on the anode surface as the cell became fully discharged. The remarkable increase in diffusion time constants at approximately 100 % SoC implies that it was intended that diffusion coefficients should decrease at the end of the discharge. This phenomenon plays a significant role in increasing the overall cell resistance especially when one of the electrodes is exhausted in Li.

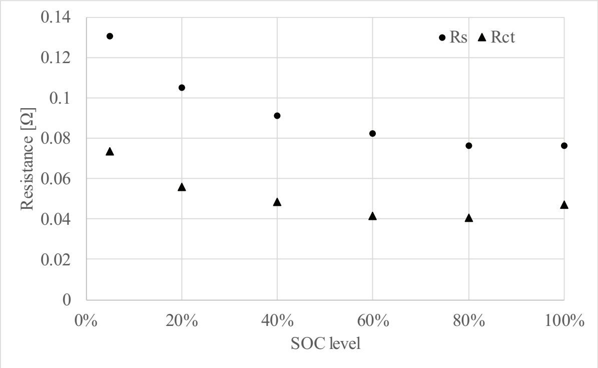


Figure 9: The trend of R_s and R_ct changes with SOC level. The cell shows a well-known increasing overall resistance during discharge.

5. Conclusions

The current work demonstrated an improved method of diffusion modeling. Several configurations of Randles circuits were studied in order to obtain the best fit of the impedance characteristics of a battery. The inappropriate fit of standard Warburg elements with regard to non-Fickian diffusion was corrected by applying generalized Warburg elements and CPEs. The proposed generalized model compensates well for the phase error between the measured and modeled impedances.

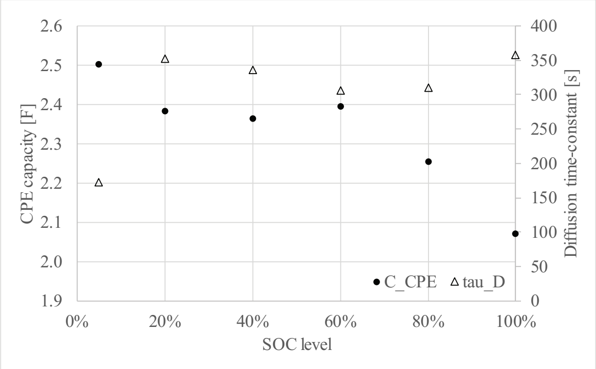


Figure 10: Charactersitics of changes in CPE capacity constants and diffusion time-constant. Diffusion time-constant remarkably increases during discharge that implies diffusion coefficients to decrease. The CPE does not change significantly.

Table 4: The estimated Randles circuit parameters at given SoC levels based on the CPE-QPE Randles circuit model.

SoC levels	100 %	80 %	60 %	40 %	20%	5 %
R_s	0.0764	0.0763	0.0823	0.0909	0.105	0.1305
R_{ct}	0.047	0.0406	0.0413	0.0487	0.056	0.0734
C_{CPE}	2.07	2.255	2.394	2.365	2.383	2.503
τ_D	358	310	306	336	353	172

The finite-length Warburg element is a classical and generally applied tool in diffusion modeling, but it should be used carefully in some cases. The proposed work also shows that the finite-length Warburg element yields multiple local solutions depending on its initial values if measured impedance data is insufficient within the very low bandwidth of impedance. This leads to ambiguous results in terms of diffusion-related parameters which should be avoided, however, this problem can definitely be solved by applying CPE or QPE instead in diffusion models.

The Randles circuit parameters were estimated and can be used in further calculations to determine the electrochemical parameters of batteries [24].

Symbols

Binary diffusion coefficient of electrolyte	$D_{1,0}$	m^2/s
Warburg coefficient	σ	$\Omega/\sqrt{s}$
Diffusion time constant	τ_D	s
AC Excitation frequency	ω	Hz
Warburg exponent	γ	–
Electrolyte resistance	R_s	Ω
Charge transfer resistance	R_{ct}	Ω
Warburg resistance	R_w	Ω
Double-layer capacitance	C_{dl}	F

Acknowledgements

The research was supported by EFOP-3.6.2-16-2017-00002 “Research of Autonomous Vehicle Systems related to ZalaZone autonome proving ground”. We would like to thank Zoltán Lukács and Tamás Kristóf to support conducting EIS measurements in the laboratory of Faculty of Physical Chemistry and sharing their knowledge in electrochemical modeling.

REFERENCES

[1] Redondo-Iglesias, E.; Venet, P.; Pelissier, S.: Efficiency Degradation Model of Lithium-Ion Batteries for Electric Vehicles, *IEEE Transactions on Industry Applications*, 2019, **55**(2), 1932–1940 DOI: 10.1109/TIA.2018.2877166

[2] Doyle, M.; Fuller, T.F.; Newman, J.: 1-Modeling of Galvanostatic Charge and Discharge, *J. Electrochem. Soc.*, 1993, **140**(6), 1526–1533 DOI: 10.1149/1.2221597

[3] Mei, W.; Chen, H.; Sun, J.; Wang, Q.: The effect of electrode design parameters on battery performance and optimization of electrode thickness based on the electrochemical-thermal coupling model, *Sustain. Energ. Fuels*, 2019, **3**(1), 148–165 DOI: 10.1039/c8se00503f

[4] Lawder, M.T.; Northrop, P.W.; Subramanian, V.R.: Model-based SEI layer growth and capacity fade analysis for EV and PHEV batteries and drive cycles, *J. Electrochem. Soc.*, 2014, **161**(14), A2099–A2108 DOI: 10.1149/2.1161412jes

[5] Orazem, M. E.; Tribollet, B.: Electrochemical Impedance Spectroscopy, *John Wiley & Sons Inc.*, 2008 ISBN: 9781119363682

[6] Diard, J-P.; Gorrec, L.B.; Montella, C.: Handbook of Electrochemical Impedance Spectroscopy, 2017, 2–40 http://www.bio-logic.info

[7] Falconi, A.: Electrochemical Li-Ion battery modeling for electric vehicles. Material chemistry. Communaute Universite Grenoble Alpes, 2017, tel-01676976

[8] Lasia, A.: Electrochemical Impedance Spectroscopy and its Applications, Springer, 2014 DOI: 10.1007/978-1-4614-8933-7

[9] Barsoukov, E.; Macdonald, J.R.: Impedance Spectroscopy, John Wiley & Sons, 2005 DOI: 10.1016/j.snb.2007.02.003

[10] Harrington, D.A.: Electrochemical Impedance Spectroscopy (thesis), 2004

[11] Huang, J.: Diffusion impedance of electroactive materials, electrolytic solutions and porous electrodes: Warburg impedance and beyond, *Electrochim. Acta*, 2018, **281**, 170–188 DOI: 10.1016/j.electacta.2018.05.136

[12] Guha, A.; Patra, A.: Online Estimation of the Electrochemical Impedance Spectrum and Remaining Useful Life of Lithium-Ion Batteries, *IEEE Transactions on Instrumentation and Measurement*, 2018, **67**(8), 1836–1849 DOI: 10.1109/TIM.2018.2809138

[13] Huang, J.; Ge, H.; Li, Z.; Zhang, J.: An Agglomerate Model for the Impedance of Secondary Particle in Lithium-Ion Battery Electrode, *J. Electrochem. Soc.*, 2014, **161**(8), E3202–E3215 DOI: 10.1149/2.027408jes

[14] Huang, J.; Li, Z.; Zhang, J.; Song, S.; Lou, Z.; Wu, N.: An Analytical Three-Scale Impedance Model for Porous Electrode with Agglomerates in Lithium-Ion Batteries, *J. Electrochem. Soc.*, 2015, **162**(4), A585–A595 DOI: 10.1149/2.0241504jes

[15] Franceschetti, D.R.; Macdonald, J.R.: Diffusion of neutral and charged species under small-signal a.c.

conditions, *J. Electroanal. Chem.*, 1979, **101**(3), 307–316 DOI: 10.1016/S0022-0728(79)80042-X

[16] Lelidis, I.; Ross Macdonald, J.; Barbero, G.: Poisson-Nernst-Planck model with Chang-Jaffe, diffusion, and ohmic boundary conditions, *J. Phys. D: Appl. Phys.*, 2016, **49**(2), 25503 DOI: 10.1088/0022-3727/49/2/025503

[17] Sapoval, B.; Chazalviel, J.-N.; Peyrière, J.: Electrical response of fractal and porous interfaces, *Phys. Rev. A*, 1988, **38**(11), 5867–5887 DOI: 10.1103/PhysRevA.38.5867

[18] Jacobsen, T.; West, K.: Diffusion Impedance in Planar, Cylindrical and Spherical Symmetry, *Electrochim. Acta*, 1995, **40**(2), 255–262 DOI: 10.1016/0013-4686(94)E0192-3

[19] Bisquert, J.; Garcia-Belmonte, G.; Bueno, P.; Longo, E.; Bulhões, L.O.: Impedance of constant phase element (CPE)-blocked diffusion in film electrodes, *J. Electroanal. Chem.*, 1998, **452**(2), 229–234 DOI: 10.1016/S0022-0728(98)00115-6

[20] Ramos-Barrado, J.R.; Galán Montenegro, P.; Criado Cambón, C.: A generalized Warburg impedance for a nonvanishing relaxation process, *J. Chem. Phys.*, 1996, **105**(7), 2813–2815 DOI: 10.1063/1.472806

[21] Qu, D.: The study of the proton diffusion process in the porous MnO2 electrode, *Electrochim. Acta*, 2004, **49**(4), 657–665 DOI: 10.1016/j.electacta.2003.08.030

[22] Birkl, C.R.; Howey, D.A.: Model identification and parameter estimation for LiFePO4 batteries, *IET Conference Publications*, 2013, **2013**(621 CP), 1–6 DOI: 10.1049/cp.2013.1889

[23] Zou, C.; Zhang, L.; Hu, X.; Wang, Z.; Wik, T.; Pecht, M.: A review of fractional-order techniques applied to lithium-ion batteries, lead-acid batteries, and supercapacitors, *J. Power Sources*, 2018, **390**(June), 286–296 DOI: 10.1016/j.jpowsour.2018.04.033

[24] Nguyen, T.Q.; Breitkopf, C.: Determination of diffusion coefficients using impedance spectroscopy data, *J. Electrochem. Soc.*, 2018, **165**(14), E826–E831 DOI: 10.1149/2.1151814jes

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 43–49 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-07

CHALLENGES OF LOCALIZATION ALGORITHMS FOR AUTONOMOUS DRIVING

HUNOR MEDVE*[1] AND DÉNES FODOR[1]

[1] Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

One could easily believe that the technology surrounding us is already easily capable of determining the current location of a vehicle. Whilst many devices, technologies, mathematical models and methods are available in the automotive world, the complexity of the localization problem still cannot be underestimated. The expectation is to determine in real time with a high degree of accuracy the location of a vehicle in order to make correct autonomous decisions and avoid dangerous and potentially damaging situations. Various research directions have been undertaken since the birth of autonomous driving from the well-known satellite navigation-based systems that rely on offline maps to the more sophisticated approaches that use odometry and existing sensor data using sensor fusion. The aim of the current work is to review what has been achieved so far in this field and the challenges ahead, e.g. the need for a change in paradigm as today's global positioning systems are not intended for machines but humans and are based on the abstraction of human thinking and human decision-making processes.

Keywords: autonomous driving, localization, information fusion, filtering

1. Introduction

Vehicle localization is one of the four functions of autonomous vehicle navigation, namely mapping, localization, motion and interaction, which are the answers to the four basic questions concerning navigation: Where am I? Where can I move to? How can I do it? How do I interact? If a vehicle is to navigate as expected, these functions need to operate correctly [1]. Historically, the purpose of in-car localization was driver assistance in the form of helping the driver to navigate. Such systems that are currently in use provide information, with some degree of accuracy, to the driver and then the driver makes decisions based on the information, which can either be accepted and acted upon or rejected in the form of proceeding in another direction. In the case of autonomous driving, it is quite clear that simply rejecting position information since the main control algorithm is not an option as this is the only item of data to be used, therefore, it must be used and a decision made based on it. This raises the question of certainty.

The requirement to operate safely anywhere and at anytime makes the performance measures far stricter than ever before. The performance measures are [2]:

Accuracy the degree of conformity of position information provided by the localization system relative to actual values.

Integrity a measure of trust that can be implemented in the information from the localization, which is the likelihood of undetected failures given the specified accuracy of the system.

Continuity of service the probability of the system continuously providing information without nonscheduled interruptions during the intended working period.

Availability the percentage of time during which the service is available for use taking into account all the outages irrespective of their origins. The service is available if the requirements concerning accuracy, integrity and continuity are satisfied.

Over the last 10-15 years, the number of sensors and related advanced driver-assistance systems in passenger vehicles has increased. The primary task of each of these sensors and services is to observe a segment of the surroundings and its status, then assist the driver in that regard. Since the data from a single sensor does not contain all the information about the vehicle's surroundings, further information concerning its absolute location cannot be extracted based on a single sensor. In fact, the sensors provide complementary information and through information fusion the vehicle's absolute location and status can be obtained. This is shown in Fig. 1.

The main groups of information sources are the following:

*Correspondence: medve.hunor@mk.uni-pannon.hu

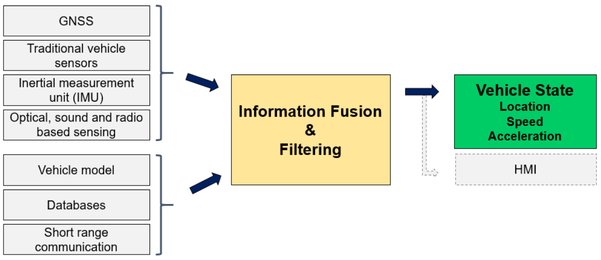

GNSS
Traditional vehicle sensors
Inertial measurement unit (IMU)
Optical, sound and radio based sensing
Vehicle model
Databases
Short range communication
Information Fusion & Filtering
Vehicle State
Location
Speed
Acceleration
HMI

Figure 1: The concept of information fusion

- Global Navigation Satellite Systems (GNSS)
- Traditional vehicle sensors:
  - Odometer
  - Wheel speed sensor
  - Steering angle sensor
- Inertial Measurement Unit (IMU):
  - Accelerometer
  - Gyroscope
- Optical, sound- and radio-based sensors:
  - Radar
  - Ultrasonic sensors
  - Vision sensors
  - LiDAR (Light Detection and Ranging)
- Vehicle models with various levels of complexity
- Databases; offline or cloud-based:
  - Maps
  - Traffic situation
- Dedicated short-range communication:
  - Vehicle-to-vehicle (V2V)
  - Vehicle-to-infrastructure (V2I).

It is important to note that none of these information sources are ideal and error-free. The errors will propagate through the sensor fusion algorithm, moreover, affect the end result and the previously described figures of merit.

The following sections address each device family then the fusion methods are analysed.

2. Information Sources

2.1 Global Navigation Satellite Systems

The global navigation satellite system (GNSS) is a radio positioning-based technology using satellite infrastructure that aims to achieve global coverage. Historically, satellite-based sytems have been considered as the core element of localization. Currently, a number of systems are in operation, the major ones are GPS (USA), BeiDou (China), GLONASS (Russia) and Galileo (EU).

Every satellite broadcasts a specific signal and its position. The spectral range of the signals is $1.2 - 1.6$ GHz, utilising frequency bands of between 2 and 40 MHz. Any user equipped with a GNSS receiver receives the signal and measures the signal propagation delay, then estimates the range of distance from it. By using signals from at least four satellites, the receivers can reduce the estimate to intercept the ranges from each satellite, which basically provides a potential location within the range in terms of geospatial coordinates. It is important to note that the position information is useful only if used together with maps which put the information in context.

Even though the accuracy of receivers is increased by various augmentation systems, issues resulting from poor satellite constellations, signal blockage and multipath propagation in urban environments cannot be excluded. For this reason, safety-critical applications cannot solely rely on GNSS technology.

Although satellite-based systems are far from perfect, they are and will continue to be the single most important information source of any localization algorithm.

2.2 Vehicle Model

Models representing the dynamic model of a vehicle's range from the simple spring-mass model to a complex multibody multi-level model. A well-known and used model is the single track model [3] with a number simplifications, however, it provides a reasonable solution for modeling lateral dynamics, therefore, it forms the core of the electronic stability program (ESP) of many vehicle manufacturers. The inputs of the single track model are lateral acceleration, longitudinal speed and yaw rate, which are provided by the relevant sensors as discussed in Section 2.3.

Another element in a complex model is the tire model which is assumed to be the only part in contact with the road. These models, e.g. Pacejka's Magic Formula [4], are often semi-empirical.

It is important to note that a more detailed model requires more parameters which, in the case of inaccurate identification, may impact the overall accuracy of the model's output.

2.3 Traditional Vehicle Sensors and Inertial Measurement Unit (IMU)

A wide range of traditional vehicle sensors have already been installed in most vehicles, moreover, analogue measurements are already being processed digitally. Most of them provide basic information to human drivers directly, e.g. the odometer, whilst others are parts of safety features. For autonomous driving, exactly the same information is also required.

Wheel speed sensors mounted in the wheel drum provide vital inputs to the anti-lock braking system by sensing the movement of the circumference of each tire in a passive or active setting.

Steering angle sensors are mounted on the steering shaft and measure the steering wheel angle, their outputs are interpreted as the intended direction of the vehicle, which is a key input to the electronic stability program (ESP).

Accelerometers measure the acceleration of the vehicle on the specified axis, multi-axis accelerometers are also in use. They are primarily used for inertial navigation in combination with yaw-rate sensors.

Yaw-rate sensors, often referred to as gyroscopes, measure the rotation of the vehicle along the vertical axis. Such a sensor provides an input to the single track vehicle model in conjunction with the ESP.

Accelerometers and yaw-rate sensors integrated in one cluster comprise the inertial measurement unit (IMU).

2.4 Optical as well as Sound- and Radio-Based Sensors

In automotive radar systems, a distinction is made between short range radar (SRR) and long range radar (LRR). The detection range of short range radar is from 0.2 to 50 meters with a detection angle of $\pm 35°$, whilst that of the long range radar is from 2 to 150 meters with a detection angle of $\pm 6°$. SRR is predominantly used in anti-collision and parking aid systems.

Radars are able to detect multiple objects as well as measure distance, relative speed and the angle to an object simultaneously. LRR is typically applied in adaptive cruise control and collision avoidance. Radar technology is affected by the weather and functionality cannot be guaranteed in extreme conditions. Overall information from radars can complement other location-related information sources.

Vision sensors are primarily used in vehicles to detect and possibly recognise its surroundings, e.g. other vehicles, obstacles, pedestrians and landmarks, which are potentially useful pieces of information for a localization algorithm.

Charge-coupled (CCD) and complementary metal oxide semiconductor (CMOS) devices are the main sensor technologies applied in digital cameras to generate an image of the surroundings, in fact both are semiconductor devices.

In CMOS devices, every pixel has its own charge-to-voltage conversion and digitalization, so their outputs are digital signals. Pixels that perform their own charge-to-voltage conversion decrease their uniformity and image quality as well as remove a useful area from light capture.

In CCD sensors, a pixel's charge signal is sent through a limited number of outputs to be converted into voltage and then transmitted out of the chip as an analogue signal to be processed and digitalized. This requires more time and energy when compared to CMOS sensors, however, results in a higher quality but less noisy image. As the CMOS manufacturing process is cheaper, recent developments have focused on overcoming the drawbacks of CMOS sensors.

Ultrasonic sensors transmit higher frequency sound waves and evaluate the echo received by the sensors. The sensors also measure the elapsed time between sending and receiving back the signal, then calculate the distance from the object. The types currently used in the automotive industry are able to measure within the range of 0.2 to 1.5 meters, with a horizontal angle of $\pm 60°$ and a vertical angle of $\pm 30°$, and are primarily used in parking aids. Nevertheless, the use of such sensors might provide useful inputs for a localization algorithm under given circumstances.

LiDAR (Light Detection and Ranging) measures the distance of an object by emitting laser light and detecting the returning light. The differences in return times and wavelengths then provide the basis for a 3D representation of the surroundings.

2.5 Databases and Maps

Maps stored in digital format differ from the classical map representations intended to be read by humans. Digital road maps are comprised of nodes and arcs connecting the nodes. Arcs are represented in a discrete form and every node and shape point on the arc has geospatial coordinates linked to them. They are often represented as planar models in applications currently on the market.

2.6 Dedicated short-range communications

For any vehicle to communicate with either the infrastructure (vehicle-to-infrastructure: V2I) or with another vehicle (vehicle-to-vehicle: V2V), it is assumed that a suitable wireless protocol is in place, which allows bidirectional information flow in real time when a vehicle is travelling at high speed and is able to simultaneously handle multiple vehicles. Based on these assumptions, only applications related to localization are considered here.

The main purpose of V2I communication is to support applications that target safety and mobility. Safety applications mainly consist of alerts and warnings, while mobility applications collect data from vehicles in order to capture the actual state of the traffic and provide such information to vehicles.

V2V applications determine the state of other nearby vehicles through the transmission of one or several messages. Overall, the location-related content of these messages alone might be insufficient for a vehicle to determine its own location, but can still provide useful complementary information to fusion algorithms.

3. Fusion Algorithms

The purpose of information fusion is to obtain more information from the sources than what is accessible from each individual source. This is achieved by combining sources which are complementary, moreover, the use of partially redundant sources reduces the ambiguity of the measured data which, overall, improves the performance of the system.

Fusion algorithms can be realized in either centralized or decentralized structures, as is shown in Fig. 2.

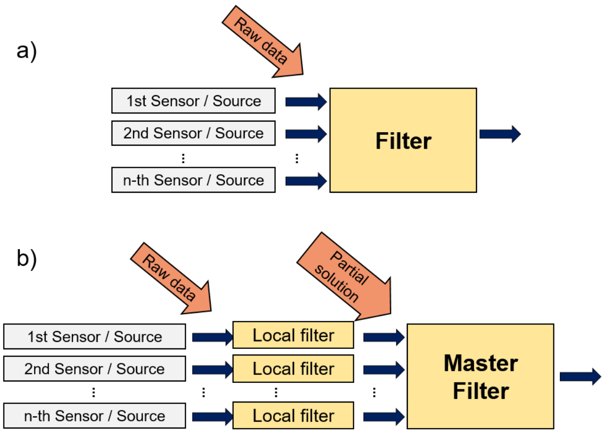


Figure 2: Filtering structures: a) centralized, b) decentralized

As the name suggests, in centralized structures, one filter performs the filtering of all signals which yields the benefit of minimal information loss as everything is directly available to the filter, while the amount of data to be processed in real time might imply an impractical degree of computational complexity. This is addressed by decentralized filters as every signal is filtered separately before being processed by a master filter. The computational load in general would be significantly less for one filtering unit, however, at the expense of partial information loss and reduced estimation accuracy.

In this chapter, selected filtering methods are presented: a conventional and widely used localization method, as well as a linear and a non-linear filtering method. Advanced filtering methods require prior knowledge of the system model and dynamics, the type of noise and their probability density functions as these are core elements to design a high-performance filter. The filtering algorithms presented in the following sections are used by the scientific community in various forms but often altered when compared to their originally published format (Kalman filter in [5], particle filter in [6]), in order to better suit the actual problem. In this paper, the thought process of [7] is followed.

3.1 Simple Algorithms: Dead Reckoning, Inertial Navigation and Map Matching

The conventional localization algorithm consists of two steps; the first is the GNSS which defines the coordinates, the second is to match the given coordinates to a map. This is shown in Fig. 3.

The process of calculating the position based on its previously known position, elapsed time and speed is referred to as dead reckoning [8]. Inertial navigation is a very similar concept where the position is calculated based on data from accelerometers and gyroscopes, also referred to as dead reckoning based on inertial sensors. In the following sections, no specific distinction is made between dead reckoning and inertial navigation.

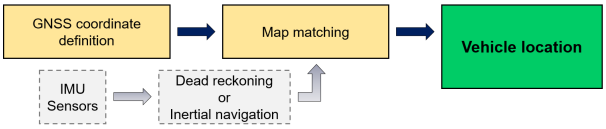

GNSS coordinate definition
Map matching
Vehicle location
IMU Sensors
Dead reckoning or Inertial navigation

Figure 3: Conventional localization algorithm

Traditional vehicle sensors and the inertial measurement unit (IMU) provide information about either the first or second order derivatives of the position of the vehicle together with the odometer which measures the distance travelled. All the data provided is relative to the starting position, therefore, none of the sensors provide information about its absolute position. In addition, all data will be incorporated into the coordinate system of the vehicle since all the sensors are mounted on the vehicle, therefore, coordinate transformations to the main coordinate system, which are used for the localization, are required. The measured values are integrated by taking into consideration an initial position and, over time, the errors will be accumulated as part of the integration. It is important to note that the extent to which the error can increase is infinite. Despite such a disadvantage, the popularity of the method lies in the fact that it does not rely on external sources of information and the update rate is determined by the system itself, which overall defines the complementary nature of inertial navigation to GNSS.

Map matching is the process of identifying on the map the coordinates given by the GNSS on the map. On digital road maps, the road network is represented in a discrete form as nodes and arcs which connect nodes, each of which has geospatial coordinate information linked to it. The purpose of the map matching algorithm is to match the GNSS coordinates to the road map. It is highly likely that the map will not contain the exact coordinates defined by the GNSS and inertial navigation, therefore, it has to be matched to one of the few possible ones. Map matching algorithms assign probabilities to each possible location based on a set of information including previous locations, speed and heading of the vehicle, subsequently the evaluation is concluded based on these probabilities. Such algorithms can provide useful inputs to assist a human driver, however, it is easy to realize that rarely rarely can they provide sufficiently reliable inputs for autonomous driving. This creates the need for more reliable algorithms, which are normally more complex and require more computational power.

3.2 Linear Filtering: The Kalman Filter

The Kalman filter is a useful engineering tool in many industries and control applications ranging from robotics, automotive, plant control, aircraft tracking and navigation. In general, they are relatively easy to design and code with an optimal degree of estimation accuracy for linear systems with Gaussian noise.

Let us describe a linear system with the following discrete state-space model equations:

$$
\begin{aligned}
x_k = A_{k-1}x_{k-1} + B_{k-1}u_{k-1} + w_{k-1} \\
y_k = C_k x_k + v_k
\end{aligned}
\tag{1}
$$

where k in subscript refers to states and measurements at each discrete time instant and $k-1$ in subscript to those at the previous time instant, x_k denotes the vector of the state variable, u_k stands for the input or control vector, y_k represents the output vector, A_k refers to the system matrix, B_k and C_k denote the input and output matrices, w_k and v_k stand for the process and measurement noise, respectively, which are white with Gaussian distribution and zero mean, and R_k and Q_k are known covariance matrices.

$$
\begin{aligned}
w_k &\sim (0, R_k) \\
v_k &\sim (0, Q_k) \\
E\left[w_k w_j^T\right] &= Q_k \delta_{k-j} \\
E\left[v_k v_j^T\right] &= R_k \delta_{k-j} \\
E\left[w v_j^T\right] &= 0
\end{aligned}
\tag{2}
$$

where δ_{k-j} is the Kronecker delta function ($\delta_{k-j} = 1$, if $k = j$ and $\delta_{k-j} = 0$, if $k \neq j$). The aim is to estimate the system state x_k by knowing the system dynamics and the noisy measurements y_k. The available information for the state estimation always depends on the actual problem at hand. If all measurements are up to date and accessible, including kth, then a posteriori estimation can be computed, which is denoted by $\hat{x}_k^+$. The meaning of the "$+$" sign in superscript means the estimation is an a posteriori estimation.

The best way to estimate the a posteriori estimation is by computing the expected value of x_k conditioned to all measurements up to now, including k as well.

$$
\hat{x}_k^+ = E\left[x_k | y_1, y_2, \ldots, y_k\right] \tag{3}
$$

If all measurements, apart from k, are accessible, then the a priori estimate can be computed, denoted by $\hat{x}_k^-$, where the "$-$" sign in superscript denotes the a priori estimate. The best way to estimate the a priori state estimate is if the expected value of x$_k$ conditioned to all measurements up to now, excluding k, is computed:

$$
\hat{x}_k^- = E\left[x_k | y_1, y_2, \ldots, y_{k-1}\right] \tag{4}
$$

It is important to note that $\hat{x}_k^-$ and $\hat{x}_k^+$ are estimates of the same quantity, before and after the actual measurement is obtained, respectively. Naturally, it is expected that $\hat{x}_k^+$ is a more accurate estimate as more information is available.

At the beginning of the estimation process, the first measurement is obtained at $k = 1$, therefore, the estimate of $\hat{x}_0^+$ ($k = 0$) is given by computing the expected value of x_0:

$$
\hat{x}_0^+ = E\left[x_0\right] \tag{5}
$$

Estimation of the error covariance is denoted by P_k, therefore, P_k^- represents the estimation of the error covariance of the a priori estimate $\hat{x}_k^-$ and P_k^+ stands for the estimation of the error covariance of the a posteriori estimation $\hat{x}_k^+$:

$$
\begin{aligned}
P_k^- &= E\left[\left(x_k - \hat{x}_k^-\right)\left(x_k - \hat{x}_k^-\right)^T\right] \\
P_k^+ &= E\left[\left(x_k - \hat{x}_k^+\right)\left(x_k - \hat{x}_k^+\right)^T\right]
\end{aligned}
\tag{6}
$$

The estimation process starts by computing $\hat{x}_0^+$, which is the best available estimate at this time instant for the value of $\hat{x}_0^+$. If $\hat{x}_0^+$ is known, $\hat{x}_1^-$ can be computed as follows:

$$
\hat{x}_1^- = A_0 \hat{x}_0^+ + B_0 u_0 \tag{7}
$$

then the general form to compute $\hat{x}_k^-$can be established:

$$
\hat{x}_k^- = A_{k-1}\hat{x}_{k-1}^+ + B_{k-1}u_{k-1} \tag{8}
$$

This is referred to as *time update* from time instants $(k-1)^+$ to k^-. No new measurement information is available between the two, therefore, the state estimation propagates from one time instant to the other, and all state estimations are based on knowledge of the system dynamics. The time update is often referred to as the *prediction step*.

The next stage is to compute P, the estimation of the error covariance. The process starts by computing P_0^+ which is the error covariance of $\hat{x}_0^+$. If the initial state is perfectly known, then $P_0^+ = 0$; if no information is available, then $P_0^+ = \infty I$. In general, the meaning of P_0^+ is the uncertainty regarding the initial estimation of x_0:

$$
P_0^+ = E\left[\left(x_0 - \hat{x}_0^+\right)\left(x_0 - \hat{x}_0^+\right)^T\right] \tag{9}
$$

If P_0^+ is known, then P_1^- can be computed as follows:

$$
P_1^- = A_0 P_0^+ A_0^T + Q_0 \tag{10}
$$

Based on the above, the generic form of the time update of P_k^- can be stated:

$$
P_k^- = A_{k-1} P_{k-1}^+ A_{k-1}^T + Q_{k-1} \tag{11}
$$

So far, the *time update* step has been presented, which is based on the system dynamics. The next step is the *measurement update*, where new information is obtained from the measurements. Using the logic from the method of recursive least squares, the availability of the measurement y_k changes the value of the constant x in the following way:

$$
\begin{aligned}
K_k &= P_{k-1} C_k^T \left(C_k P_{k-1} C_k^T + R_k\right)^{-1} = P_k C_k^T R_{k-1}^{-1} \\
\hat{x}_k &= \hat{x}_{k-1} + K_k (y_k - C_k \hat{x}_{k-1}) \\
P_k &= (I - K_k C_k) P_{k-1} (I - K_k C_k)^T + K_k R_k K_k^T = \\
&= \left(P_{k-1}^{-1} + C_k^T R_k^{-1} C_k\right)^{-1} = \\
&= (I - K_k C_k) P_{k-1}
\end{aligned}
\tag{12}
$$

where $\hat{x}_{k-1}$ denotes the estimation and P_{k-1} stands for the the estimation of the error covariance before processing measurement y_k, therefore, $\hat{x}_k$ and P_k refer to the same informaton but after y_k has been processed.

If the logic of $\hat{x}_{k-1} \to \hat{x}_k^-$ and $\hat{x}_k \to \hat{x}_k^+$ (a priori and a posteriori, respectively) is applied and the aforementioned equation reformulated, the a posteriori estimation is produced:

$$\begin{aligned} K_k &= P_k^- C_k^T \left(C_k P_k^- C_k^T + R_k\right)^{-1} = P_k^+ C_k^T R_k^{-1} \\ \hat{x}_k^+ &= \hat{x}_k^- + K_k (y_k - C_k \hat{x}_k^-) \\ P_k^+ &= \left(I - K_k C_k\right) P_k^- \left(I - K_k C_k\right)^T + K_k R_k K_k^T = \\ &= \left[\left(P_k^-\right)^{-1} + C_k^T R_k^{-1} C_k\right]^{-1} = \\ &= \left(I - K_k C_k\right) P_k^- \end{aligned} \tag{13}$$

These are the equations for the Kalman filter *measurement update* or *a posteriori* estimation. The matrix K_k is often referred to as the Kalman gain.

By summarising the Kalman filtering algorithm, after initiation, the a priori estimate for every time instant k is given by:

$$\begin{aligned} \hat{x}_k^- &= A_{k-1} \hat{x}_{k-1}^+ + B_{k-1} u_{k-1} \\ P_k^- &= A_{k-1} P_{k-1}^+ A_{k-1}^T + Q_{k-1} \end{aligned} \tag{14}$$

and the a posteriori estimation is given by:

$$\begin{aligned} K_k &= P_k C_k^T R_{k-1}^{-1} \\ \hat{x}_k^+ &= \hat{x}_k^- + K_k (y_k - C_k \hat{x}_k^-) \\ P_k^+ &= (I - K_k C_k) P_k^- \end{aligned} \tag{15}$$

The aforementioned Kalman filtering algorithm is the optimal state estimator for linear systems with Gaussian unimodal noise processes, however, most real-world systems are nonlinear and, in many cases, with multimodal non-Gaussian noise, include a probability density function. A number of variations of Kalman filters developed by the scientific community are trying to address the problem of nonlinearity. Most of them rely on the basic concept of Kalman filters using nonlinear adaptations, e.g. the extended Kalman filter which, at its core, is still a linear filter.

In general, versions of the nonlinear Kalman filter are considered to estimate accuracy well but are often poor compared with the theoretically optimal accuracy, with a real-time computational complexity in the order of d^3 where d denotes a dimension of the state vector [9].

3.3 Nonlinear Filtering: The Particle Filter

Given the concerns about the estimation accuracy of versions of Kalman filters, true nonlinear filters or estimators are needed. The particle filter is a statistics-based estimator where at every discrete time instant, a number of state vectors, referred to as particles, are assessed with regard to how likely they are to be the closest to the actual state. The mathematical formulation of the aforementioned idea is summarized in this section.

Let us describe a nonlinear system using the following equations:

$$\begin{aligned} x_{k+1} &= f_k\left(x_k, w_k\right) \\ \mathrm{y}_k &= h_k\left(x_k, v_k\right) \end{aligned} \tag{16}$$

where k denotes discrete time instants, x_k and y_k represent the state and measurement, respectively, and w_k and v_k stand for the noises of the system and measurement, respectively. The functions $f_k(\cdot)$ and $h_k(\cdot)$ are a time variant nonlinear system and a measurement function, respectively. The noises of the system and measurement are assumed to be white and independent from each other with known probability density functions.

The aim of the generic Bayes estimator is to approximate the conditional probability density function x_k based on measurements $y_1, y_2, \ldots, y_k$. This conditional probability density function is denoted as follows:

$$p\left(x_k | Y_k\right) = x_k \tag{17}$$

conditioned on measurements $y_1, y_2, \ldots, y_k$. The particle filter is the numeric implementation of the Bayes estimator, in the following section this will be described.

At the beginning of the estimation, it is assumed that the probability density function of $p(x_0)$ is known, then N number of state vectors based on the probability density function of $p(x_0)$ are randomly generated. These state vectors are the particles and are denoted by $x_{0,i}^+ \, (i = 1, \ldots, N)$. The value of N can be chosen arbitrarily, depending on the expected estimation accuracy and available computational capacity. At every $k = 1, 2, 3 \ldots$ discrete time instant, every particle is propagated to the next time instant using process equations $f_k(\cdot)$:

$$x_{k,i}^- = f_{k-1}\left(x_{k-1,i}^+, w_{k-1}^i\right) \tag{18}$$

where $(i = 1, \ldots, N)$ and every noise vector w_{k-1}^i is randomly generated based on the known probability density function of w_{k-1}. This is the *a priori* estimate of the particle filter.

Subsequently, at every time instant k, once the measurement result can be accessed, the relative conditional probability of each $x_{k,i}^-$ can be computed and $q_i = p_k\left(y_k \,\middle|\, x_{k,i}^-\right)$ evaluated if the nonlinear measurement equation and the probability density function of the measurement noise are known.

After the relative conditional probability of each particle has been evaluated, the relative probability of the actual state being equal to each of the particles is correct.

The relative probabilities q_i are then scaled to the interval $[0, 1]$ as follows:

$$q_i = \frac{q_i}{\sum_{j=1}^{N} q_j} \tag{19}$$

This ensures that the total probability is equal to one. The next stage is the resampling based on the computed and scaled probabilities. This means that a set of new $x^+_{k,i}$ particles is generated based on the relative probabilities q_i. This is the *a posteriori* estimation of the particle filter. The resampling is an important step with regard to the implementation due to the required computational capacity which needs to be considered carefully.

The distribution of the computed *a posteriori* $x^+_{k,i}$ particles is in accordance with the probability density function $p_k\left(x_k|y_k\right)$. Based on this, any kind of statistical evaluation can be carried out, for example, of the expected value, which can be considered as the statistical estimation of the actual state vector:

$$E\left(x_k|y_k\right) \approx \frac{1}{N}\sum_{i=1}^{N} x^+_{k,i} \tag{20}$$

A number of ways, the resampling algorithm in particular, are available to design and implement the steps of the filter. The number of particles required to achieve a given estimation accuracy increases in direct correlation with the dimension d of the state vector, this is linear for d of a particle filter using a complex resampling algorithm, but exponential for a plain resampling algorithm [10], while the real-time computational complexity is directly proportional to the number of particles.

4. Conclusion

Despite the fact that satellite-based systems are far from perfect, they are and most likely will continue to be the single most important information source of any localization algorithm combined with digital maps. The role of other information sources, on the one hand, is complementary in areas where GNSS has its weaknesses, but on the other hand they contribute to an increase in accuracy at the expense of computational complexity.

In a practical real-time application, the extra computational capacity and related costs are not necessarily proportional to each other. This seems to be the main drawback of using nonlinear filtering methods, while on the other hand autonomous vehicles are expected, in the long term, to fall into the category of high-volume low-cost products.

Hybrid approaches can be considered due to the fact that the equations for localization systems are only partially nonlinear or some of the subsystems can provide sufficiently accurate results using a linear approach. The filtering problem can then be divided into a linear and a nonlinear part, where the former, assuming Gaussian distributed noise, may be solved by using a simple Kalman filter and reducing the computational complexity and, therefore, the cost of the system. The proportion of linear filtering to nonlinear filtering within the full system is determined by the complexity of the system model chosen as the type of filtering is defined by the model, therefore, modeling and filtering cannot be separate elements in the design process.

Acknowledgements

The research was supported by EFOP-3.6.2-16-2017-00002 programme of the Hungarian National Government.

REFERENCES

[1] Eskandarian, A.: Handbook of Intelligent Vehicles, Springer Verlag, London, 2012, Chapter 50, p. 1278 DOI: 10.1007/978-0-85729-085-4

[2] Skog, I.; Handel, P.: In-car positioning and navigation technologies – a survey. *IEEE Trans. Intell. Transp. Syst.* 2009, **10**(1), 4–21 DOI: 10.1109/TITS.2008.2011712

[3] Riekert, P.; Schunck, T. E.: Zur Fahrmechanik des gummibereiften Kraftfahrzeugs. *Ing. Arch.* 1940, **11**, 210–224 DOI: 10.1007/BF02086921

[4] Pacejka, H. B.: Tire and Vehicle Dynamics, Butterworth-Heinemann, Oxford, 2012, 3rd Edition DOI: 10.1016/C2010-0-68548-8

[5] Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 1960, **82**(1), 35–45 DOI: 10.1115/1.3662552

[6] Gordon, N. J.; Salmond, D. J.; Smith, A. F. M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings*, **140**(2): 107–113 DOI: 10.1049/ip-f-2.1993.0015

[7] Simon, D.: Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches, Wiley-Interscience, 2006 DOI: 10.1002/0470045345

[8] Karlsson, R.; Gustafsson, F.: The Future of Automotive Localization Algorithms: Available, reliable, and scalable localization: Anywhere and anytime. *IEEE Signal Processing Magazine*, 2017, **34**(2), 60–69 DOI: 10.1109/MSP.2016.2637418

[9] Daum, F.: Nonlinear filters: beyond the Kalman filter. *IEEE Aerospace and Electronic Systems Magazine*, 2005, **20**(8), 57–69 DOI: 10.1109/MAES.2005.1499276

[10] Daum, F. E.; Huang, J.: The curse of dimensionality and particle filters. *Proceedings of IEEE Conference on Aerospace, Big Sky, MT, USA, 2003*, pp. 4-1979–4-1993. DOI: 10.1109/AERO.2003.1235126

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 51–59 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-08

MODELING OF AUTONOMOUS EV VECTOR-CONTROLLED POWER CONVERSION SYSTEM FOR BATTERY MANAGEMENT SYSTEM DESIGN

GÁBOR KOHLRUSZ*[1] AND DÉNES FODOR[1]

[1] Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

Abstract - Since the energy consumption of electric drivetrains can be optimized in an automatically controlled system, the driving ranges and efficiency of driverless electric cars can be enhanced. The analysis of a model of the electric power conversion system provides the opportunity to consider different driving circumstances, moreover, it is possible to evaluate the performance of a power conversion system when the vehicle is driven along different routes. The results provide detailed information on the transient operation of all the power modules as well as their components, and on the overall performance of the power conversion system. In this study, a permanent-magnet synchronous machine (PMSM) acts as the traction motor in the autonomously driven electric car. Besides the PMSM, the power electronics and battery are also modeled in an OrCAD PSpice circuit-simulation environment that serves as a model of an electric power conversion system for the simulation and testing of a battery management system algorithm. The battery management system and control algorithm are modeled in Simulink and can be tested together with the PSpice-modeled circuits utilizing an interconnected simulation environment. The input of the power conversion system model was a driving scenario that included uphill and downhill sections. The performance of the implemented battery management system algorithm was analyzed and evaluated.

Keywords: electric power conversion system, autonomous vehicle, battery management, interconnected simulation

1. Introduction

A driverless car is an automated system where the driving task is a self-acting function. As a result, it is possible to optimize the energy efficiency of the car in order to increase its driving range and reduce its energy consumption. The optimal operation can be achieved by applying well-designed control algorithms and battery management functions.

The electric power conversion system of an autonomous vehicle is a very complex system, especially when the electronic control unit is also considered to be part of the system. During the test procedure of such a complexly controlled electric power conversion system, several unforeseen or unexpected behaviors can occur which may lead to damage. In order to avoid damage, it is essential to identify unexpected system behaviors as early as possible.

In the early phase of development, the testing and design of these systems should be performed simultaneously to take advantage of the interconnected simulation environment. A mixed-signal simulation environment supports the analysis of controlled electric power conversion systems where digital and analogue domains can be connected and run simultaneously [1]. Simulations can be performed by taking into account practical aspects that introduce cycle times and delays into the model. In such a realistic simulation environment, the reliable design of a battery management system can be achieved.

2. Electric Power Conversion System Model

A system model, including a model of the electronic control unit, is a prerequisite to the simulation. The electric part of the system can be accurately modelled in a circuit design environment, while the digital environment and control algorithm should be implemented in an environment that is capable of running numerical computations.

The electronic components were modeled in OrCAD PSpice environment while the control algorithm was realized in MATLAB-based Simulink. The electric parts of the modeled power conversion system can be seen in Fig. 1. The system consists of a Nickel Manganese Cobalt Oxide (NMC)-type Li-ion battery, a three-phase inverter and a permanent-magnet synchronous machine as the traction motor. The modeled system is a low-power counterpart of a vehicle power conversion system assum-

*Correspondence: kohlrusz.gabor@mk.uni-pannon.hu

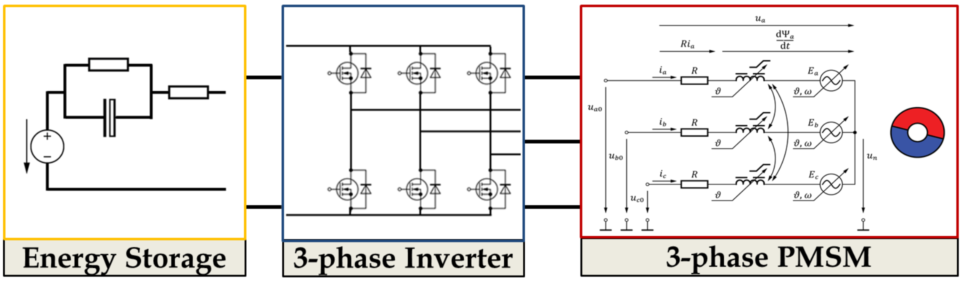


Figure 1: Structure of the Modeled Power Conversion System

ing that the system behavior is independent of the power rating.

2.1 Modeling of Energy Storage

The energy storage element of the modeled power conversion system is a Li-ion battery. The modeled battery cell consists of 4 individual NMC-type Sony VTC4A 2100 mAh battery cells connected in series. As Fig.2 shows, the model of the battery contains a State-of-Charge (SoC)-dependent voltage source and three passive components.

The model of the battery cell was developed and validated by using measurement data. A fully charged Sony VTC4A 2100 mAh battery cell was discharged slowly by setting the discharge current to a C/10 value until its voltage reached the recommended minimum voltage level. The time, battery voltage and discharge current were recorded and the SoC values were determined by time-integration of the current. Using the voltage and SoC time series, the $U(\mathrm{SoC})$ function of the battery cell could be obtained by curve fitting. The best fit was achieved by

$$\begin{aligned} U &= 0.227\log_{10}\left(0.1\left(\mathrm{SoC}+3.35\cdot10^{-5}\right)\right) \\ &+ 0.535\,\mathrm{SoC}^3+3.8926 \qquad (1) \end{aligned}$$

The fitted curve and the measured voltages are shown together in Fig. 3 as a function of the SoC.

In order to complete the battery model, it is also necessary to obtain the electrical parameters of the battery cell. The values of the passive components were provided by measurements which put the transient behavior of the battery cell into focus. In Table 1, the values of all three parameters can be seen.

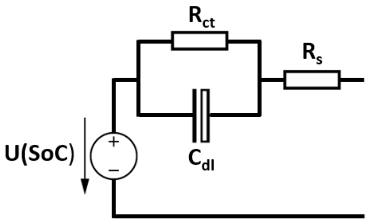


Figure 2: Equivalent Circuit Model of the Battery

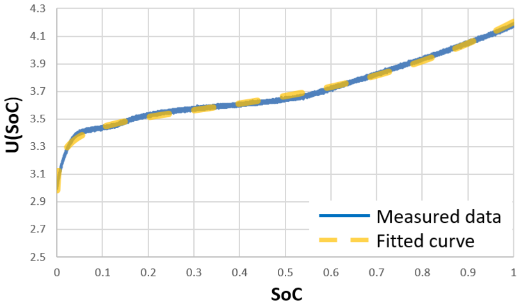


Figure 3: Approximation of the Voltage Response of the Battery

2.2 Modeling of the Three-phase Inverter

The energy conversion between the electric motor and the battery was provided by a three-phase full-bridge inverter with 2 Metal–Oxide–Semiconductor Field-Effect Transistor (MOSFET) semiconductors in each leg. The drive circuit was connected through a gate resistor to avoid an inrush current at the gate of the mosfets. The pulse-width modulated (PWM) gate-source voltages were provided by the digital controller which was implemented in the MATLAB-based Simulink environment.

A schematic diagram of the inverter can be seen in Fig. 4. As the figure shows, a shunt resistor was added to each phase wire for current sensing.

2.3 Modeling of the Three-phase PMSM

In the modeled system, the traction motor was a permanent-magnet synchronous motor. The implemented model is valid for constructions with surface-mounted magnets on the rotor and distributed winding in the wye-wound stator.

The electrical model was created by determining the voltages across the stator windings which consisted of

Table 1: Equivalent Circuit Model Parameters of the Battery

Circuit parameter	Notation	Value
Series Resistance	R_s	$0.1\,\Omega$
Charge Transfer Resistance	R_ct	$0.0052\,\Omega$
Double-layer Capacitance	C_dl	$0.9\,\mathrm{F}$

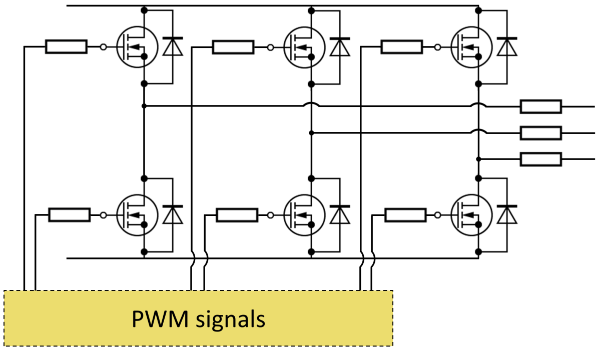


Figure 4: Three-phase Full-bridge Inverter

the resistive drop, the voltage across the inductor and the back electromotive force (back EMF). In the OrCAD environment, it was quite easy to create the electric part of the motor model due to the fact that the voltage across the inductor in each phase could be represented by defining time-dependent self- and mutual inductances as passive components of the circuit, while their analytical modeling would have been complicated.

The simulation of the motor requires the implementation of the mechanical submodel as well. Papers dealing with the model of the three-phase PMSM also included the electromagnetic torque [2, 3], however, it was challenging to determine the expanded form of an expression which could be easily implemented in an OrCAD environment. In Ref. [4], a suitable form is presented for implementation in an OrCAD environment, hence this model was used in this work to obtain the mechanical model.

In the simulation, a MAXON EC-4pole 252463 type motor was considered. The parameters of the motor can be seen in Table 2.

3. Modeling of Electronic Control Unit

The Electronic Control Unit was modeled in a MATLAB-based Simulink environment which is suitable for quick algorithm and was equipped with an interface to OrCAD PSpice circuit simulation software.

A realistic electronic control unit could be modeled using a trigger-activated Simulink function which run the control algorithm when a trigger signal is received. The trigger signals were generated to fit the operating frequency of the digitally controlled discrete-time system. This solution ensured that the control algorithm only ran once during each sampling period, consequently, the Pulse Width Modulation (PWM) signals of each MOSFET were generated only once during each sampling period. This solution is able to accurately simulate the operation of a microcontroller or even an electronic control unit.

The discrete-time control algorithm was implemented in the trigger-activated block. The structure of the algorithm can vary depending on how the difference equation is determined.

Table 2: PMSM Model Parameters

MOTOR parameter	**Notation**	**Value**
Phase resistance	R	0.49 Ω
Leakage inductance	L_{sl}	33 μH
Mean value of the mutual inductances	L_{so}	92 μH
Amplitude of inductance fluctuation	L_x	10 μH
Amplitude of permanent magnet-created flux linkage	Ψ_{PM}	22.4 mWb
Friction coefficient	B	5.25×10^{-5} Nms
Rotor inertia	J	220 gcm^2
Number of pole pairs	z_{P}	2

3.1 Discrete-Time Control Algorithm

The frequently used Proportional Integral (PI) algorithm was implemented to control the power conversion system. The structure of the control system is shown in Fig. 5. As the figure shows, a speed-cascaded torque control loop and a magnetic field control loop were created. This strategy is the widespread field-oriented control scheme of three-phase alternating current (AC) motors.

In the case of a PI controller, the only dynamic component of the controller is the integrator since the proportional component simply multiplies the error signal. Consequently, the proportional component can be implemented in the same form as in a continuous-time algorithm, however, the discrete-time form of the integrator has to be determined by using a Z-transform. The transformation can be conducted using different formulae to approximate the continuous-time system.

An easy way to determine the discrete-time form of the integrator is to use the backward Euler method. Using this method, the value of the integral at time j is

$$Y_j = Y_{j-1} + e_j T T_{\mathrm{I}} \tag{2}$$

where Y_j denotes the result of the integral at time j, Y_{j-1} represents the result of the integral at time $j-1$, e_j stands for the error signal at time j, T is the sampling time and T_{I} denotes the integral gain.

A drawback of this method is that a continuous-time system could be unstable in its discrete-time form using the backward Euler method [5].

The bilinear transform (also known as Tustin's method) could be a more reasonable technique to obtain the discrete-time representation of a system since it preserves stability. The value of an integral at time j is determined by this method as follows:

$$Y_j = \frac{Y_{j-1} + \left(e_j + e_{j-1}\right) T T_{\mathrm{I}}}{2} \tag{3}$$

where Y_j denotes the result of the integral at time j, Y_{j-1} stands for the result of the integral at time $j-1$, e_j represents the error signal at time j, e_{j-1} is the error signal

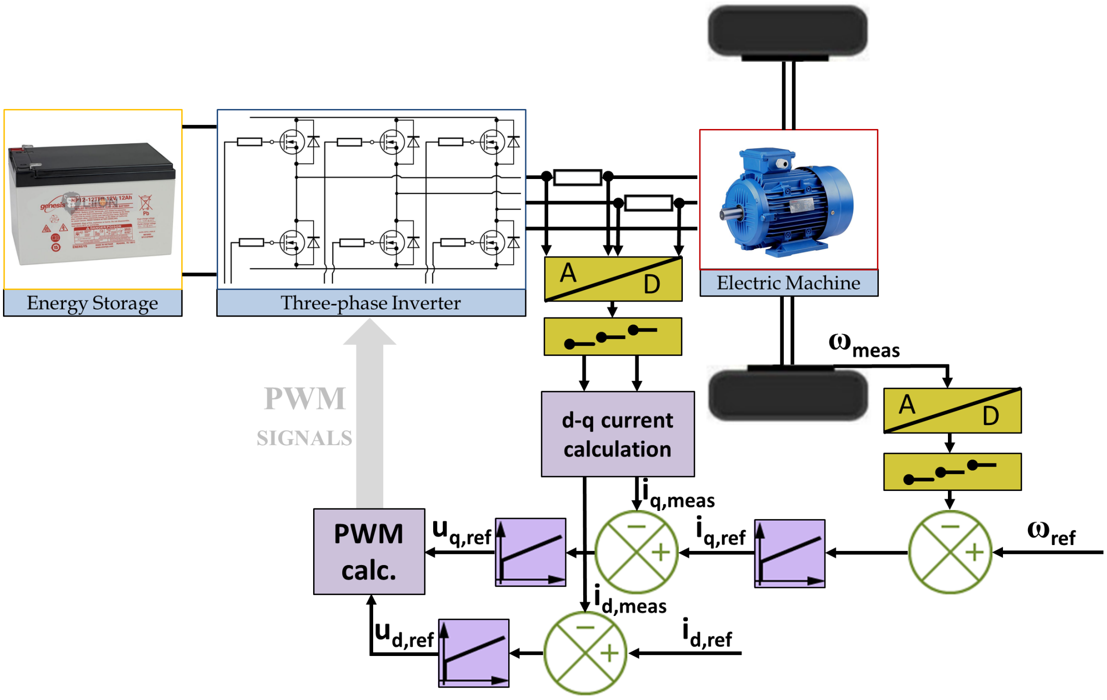


Figure 5: Vector-Controlled Power Conversion System of an Electric Vehicle

at time $j-1$, T denotes the sampling time and T_I stands for the integral gain.

The realization of a controller that applies the backward Euler method is easier and this form generally provides a stable operation, thus in the implemented model of a controller, the integrator was approximated by Eq. 2.

3.2 Design of the Controller Parameters

The implemented controllers were PI controllers in parallel form, that is the integral and proportional terms were independent of each other.

One of the simplest controller design procedures is the pole-zero cancellation but this method provides a frequency bandwidth for the controller that is identical with that of the controlled system resulting in poor performance occasionally. This problem is particularly severe when a system with a large time constant is controlled by a slow integrator with the same time constant as the system [6].

A better controller performance can be achieved by another analytical tuning method where the closed-loop poles are chosen to obtain an arbitrarily determined closed-loop Transfer function. The closed-loop is shaped by utilizing some predefined parameters such as the desired damping factor and bandwidth of frequencies.

The parameters of the PI controller can be computed as

$$K_\mathrm{c} = \frac{2\xi\omega_0\tau - 1}{K} \tag{4}$$

where K_c denotes the proportional gain of the controller, ω_0 stands for the natural frequency of the desired closed-loop system, ξ represents the desired damping factor of the second-order closed-loop system, τ is the time constant of the controlled first-order subsystem, and K denotes the gain of the controlled first-order subsystem.

The value of the integral gain is determined as

$$T_\mathrm{I} = \frac{\omega_0^2\tau}{K} \tag{5}$$

where ω_0 denotes the natural frequency of the desired closed-loop system, τ stands for the time constant of the controlled first-order subsystem, and K represents the gain of the controlled first-order subsystem.

The damping ratio was chosen to ensure an overshoot of approximately 1%, therefore, for the computations, $\xi = 0.85$ was applied. The desired natural frequency of the closed-loop system (ω_0) was obtained by determining the desired settling time of the controlled variable by considering a desired error band:

$$\omega_0 = -\frac{\ln p}{\xi t_\mathrm{b}} \tag{6}$$

where p denotes the difference between the reference signal and the value of the controlled variable as a percentage of the reference value, ξ stands for the desired damping factor of the second-order closed-loop system, and t_b represents the period of time until the controlled variable is expected to reach the desired error band [7–9].

3.3 Implementation of the Smith predictor

When a discrete-time unit controls a system, a time delay is always present. During each time period, some measurements serve as the basis of control signal generation.

However, the calculated control signals are only realized as controller outputs during the following time period, which leads to a delay of one period in duration.

Compensation for this one-period delay can be solved by implementing a Smith predictor. The predictor is an algorithm which modifies the actual measurement signal by extrapolation based on former measurements. The implementation of a Smith predictor requires the difference equation of the controlled variable as a function of the controller output to be obtained. Based on the difference equation, the measurement signal was compensated by predicting the measurement value of the following time period:

1. At time T_1, measurement signal $\hat{u}_1$ is considered for the computation of controller output y_2 which will only take effect at time T_2.
2. From y_2, it is possible to compute the predicted measurement value $\hat{u}_2$ that can be used to obtain $\hat{u}_1 - \hat{u}_0$
3. $\hat{u}_1 = u_1 + (\hat{u}_1 - \hat{u}_0)$ is the corrected measurement signal, which is used to compute y_2 as has already been stated in Step 1:

where u_1 denotes the measured value of the controlled variable at time T_1 modified by the correction term, y_2 stands for the controller output at time T_2, $\hat{u}_2$ represents the predicted value of the measurement at time T_2, and $\hat{u}_1$ and $\hat{u}_1$ are the computed values of the controlled variable at times T_1 and T_0, respectively.

In Step 2, the difference equation of the controlled subsystem is used for the computation of the predicted $\hat{u_2}$ at time T_2. In the case of the current controllers, the prediction algorithm is

$$\hat{\imath}_2 = (T_2 - T_1)\,\frac{R}{L}\left(\frac{U_2}{R} - \hat{\imath}_1\right) + \hat{\imath}_1 \tag{7}$$

where the current $\hat{\imath}_2$ is the predicted value of the measurement signal at time T_2, which was the basis of the calculation of the correction term $\hat{\imath}_1 - \hat{\imath}_0$. The measured current i_1 could be corrected as

$$\hat{\imath}_1 = i_1 + (\hat{\imath}_1 - \hat{\imath}_0)\,. \tag{8}$$

The predictor was the same in both the i_d and i_q current controllers, only the parameters were different as L_d was substituted for Eq. 7 in case of the i_d control loop and L_q was substituted for Eq. 7 in case of the i_q control loop. The Smith predictor in the cascade control loop requires the difference equation of the angular velocity as a function of the current i_q:

$$\hat{\omega}_2 = (T_2 - T_1)\,\frac{1.5 z_{\mathrm{P}} \Psi_{\mathrm{PM}} i_{q,1} - B\hat{\omega}_1 - M_{\mathrm{T}}}{J} + \hat{\omega}_1 \tag{9}$$

The measured speed ω_1 could also be corrected BY calculating the values $\hat{\omega}_1$ and $\hat{\omega}_0$ from $\hat{\omega}_2$:

$$\hat{\omega}_1 = \omega_1 + (\hat{\omega}_1 - \hat{\omega}_0) \tag{10}$$

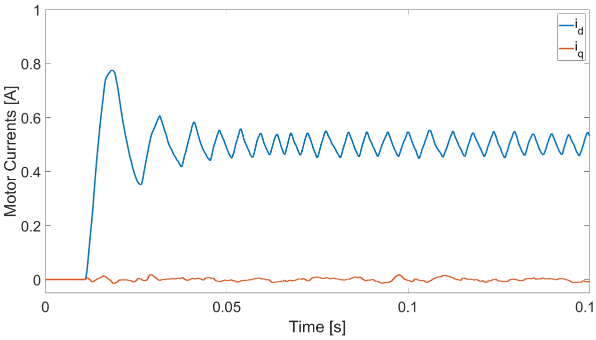


Figure 6: Testing of the i_d and i_q current controllers

4. Testing of the Controlled Power Conversion System Model

The testing was performed with an operating frequency of 25 kHz in the control loop. The current controllers were designed to provide a settling time of $t_{\mathrm{b}} = 2$ ms with $p = 0.02$ that equated to an error of 2% which resulted in the desired frequency bandwidth of $\omega_0 = 2301$ rad/s. By taking into consideration the frequency bandwidth, the calculated controller parameters were

$$K_{\mathrm{c},i_d} = 0.12 \tag{11}$$

$$T_{\mathrm{I},i_d} = 824 \tag{12}$$

for the d-axis current controller and

$$K_{\mathrm{c},i_q} = 0.238 \tag{13}$$

$$T_{\mathrm{I},i_q} = 985 \tag{14}$$

for the q-axis current controller.

The testing of the controllers was performed without the implementation of the cascaded speed control loop. A reference current was set up for both current loops, d and q. The results can be seen in Figs. 6 and 7.

It is shown in Fig. 6 that the current i_d overshot the reference value of $0.5\,\mathrm{A}$ and started to oscillate around the setpoint with an amplitude of $50\,\mathrm{mA}$, which is 10% of the reference value. The current i_q tracked its setpoint of $0\,\mathrm{A}$ with an inappreciable error. The results show that the integral gain of the current i_d should be decreased, therefore, the controllers were tested with redesigned i_d

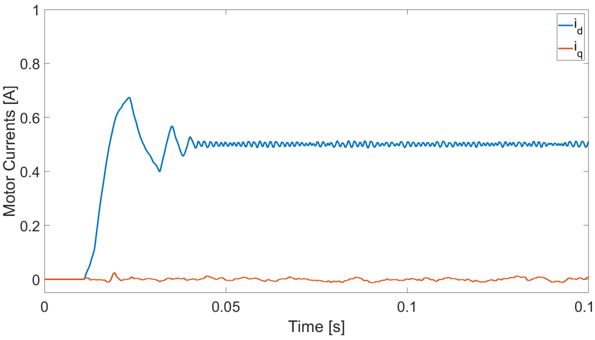
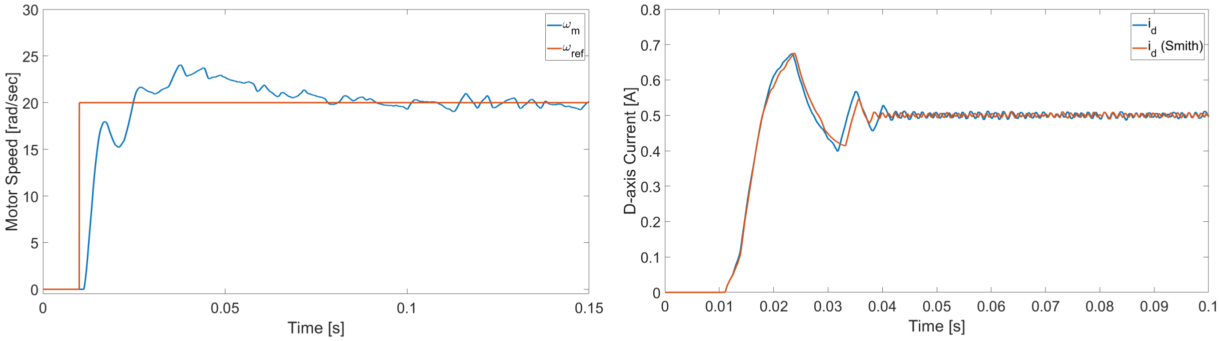


Figure 7: Results of the redesigned current controllers



(a) Speed of the Electric motor

(b) Performance of the d-axis current controller with and without the Smith predictor

Figure 8: Simulation results of the controlled system using the Smith predictor in the current controllers

controller while the i_q controller parameters remain unchanged.

$$K_{\mathrm{c},i_d} = 0.12 \tag{15}$$

$$T_{\mathrm{I},i_d} = 275 \tag{16}$$

for the current controller d.

In Fig. 7, the results for the d-axis controller testing are presented. The performance of the redesigned d-axis current controller i_d was enhanced. The controlled variable reached the setpoint and stabilized after approximately $2\,\mathrm{ms}$.

The angular velocity ω was controlled by a cascaded PI controller in the q-axis current loop. The controller parameters were obtained by determining the desired frequency bandwidth which provided a settling time of $t_\mathrm{b} = 0.05\,\mathrm{s}$ and $p = 0.05$, which equated to an error of 5%:

$$K_{\mathrm{c},\omega} = 0.038 \tag{17}$$

$$T_{\mathrm{I},\omega} = 1.627 \tag{18}$$

Fig. 8a shows the controlled motor-speed signal which has a reference value of $20\,\mathrm{rad/s}$. It can be seen that the speed signal overshot slightly and stabilized after $0.065\,\mathrm{s}$ which was acceptable since the designed settling time was defined as $t_\mathrm{b} = 0.05\,\mathrm{s}$.

4.1 Testing of the Smith predictor

The predictor was only implemented for the current controllers since the speed tracking performance of the system was satisfactory without the application of a predictor. The implemented predictor was tested under the same conditions as the current controllers, that is the current references i_d and i_q were set at $0.5\,\mathrm{A}$ and $0\,\mathrm{A}$, respectively. Fig. 8b shows the controller performance with and without the Smith predictor.

In Fig. 8b, it can be seen that the controller performed better when the Smith predictor modified the error in the input signal of the controller by correcting the measured current signal. The controlled current i_d stabilized faster and in a steady state it could be observed that the PWM-related high-frequency oscillation of the current occurred with a reduced amplitude compared to when the Smith predictor was not applied. The behavior of the current signal i_q did not change significantly since its reference value was set at $0\,\mathrm{A}$ and no dynamic changes in the reference signal needed to be handled.

5. Battery Management System Design

The controlled power conversion system was validated by tests. The created model of the complex system is suitable for simulations that are intended to analyze the operation of the system. By analyzing the operation of the power conversion system, the design aspects of a battery management system can be summarized. Simulation of the power conversion system can be a suitable procedure to validate the operation of the designed battery management system under realistic conditions.

5.1 Implementation of a Battery Management System

When a battery management strategy is implemented, it influences both the hardware and software components of the supervised system. Battery management solutions actuate in order to ensure efficient system operation as well as keep the system safe. If the management algorithm is used to ensure the system operates safely, all the avoided situations should be summarized during the design phase. Moreover, for each undesirable situation, a hardware solution has to be designed, which can be activated when necessary.

In this study, a battery overcharge protection system was analyzed as a battery management system. For this purpose, the power electronics had to be extended by additional power switches that were capable of changing operation mode when overcharging occurred (see Fig. 9). The implementation of the battery management system also included augmentation of the software.

In Fig. 9, a power semiconductor (M_1) can be seen that ensures an alternative path for the current when the battery cannot be charged. In this case, the recuperated energy is dissipated into a resistor. When switch M_1 is

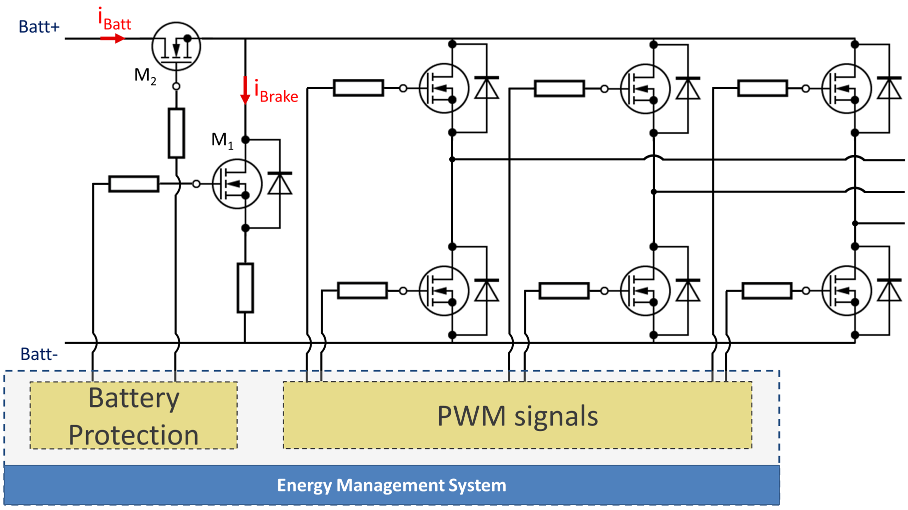


Figure 9: The analyzed battery management system

turned on, another switch (M_2) becomes responsible for decoupling the battery, otherwise the battery would be discharged through the resistor. In certain situations, generation of the PWM signal is also affected by the battery management algorithm. In this case, the PWM signals are generated in the same way whether the battery protection circuit operates or not.

5.2 Simulation of a Driving Cycle of a Vehicle

The simulation of a driving cycle of a vehicle provides the opportunity to observe the efficiency of the battery management strategy of the power conversion system and its interactions with the components of the system. The results serve as a basis for implementing different battery management strategies or modifying the implemented one.

A driving cycle can be simulated by defining different load torques of the traction motor. When an uphill or downhill stretch of road should be represented, the motor is loaded with a positive or negative load torque, respectively. If the gradients of the uphill and downhill sections of road are unchanged, the load torque remains constant.

A simulation of 5.5 s in duration was performed. The simulated driving cycle consisted of uphill, downhill and flat sections (see Fig. 10). In the picture, the load torques of the PMSM can also be seen below the corresponding sections of road. During the driving cycle, the speed of the vehicle was expected to be constant, hence the angular velocity was assumed to also be constant at $20\,\mathrm{rad/s}$. At the beginning of the simulation, it was assumed that the battery was fully charged.

The results of the simulation can be seen in Figs. 11 and 12. In Fig.11, the speed of the traction motor can be seen throughout the driving cycle. It can be observed that the changes in load torque influenced the speed of the motor, however, the controller promptly compensated for the transient overshoots.

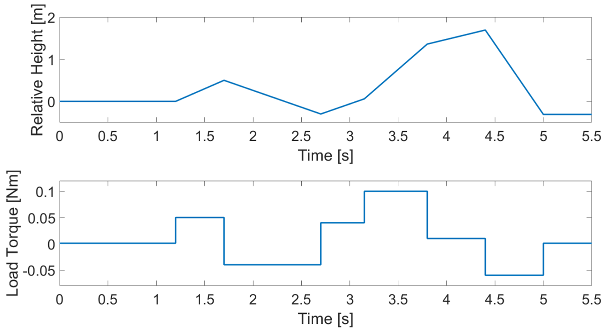


Figure 10: Simulated vehicle path and related load torques

The currents of the power conversion system are presented in Fig. 12. The braking current and the current of the battery are also shown in Fig. 9. In Fig. 12, it can be noted that the q-axis current controller achieved precise and fast reference tracking when the system was perturbed by changes in load torque. If the reference current i_q was negative, the energy was recuperated, that is the traction motor was in generator operation mode. Since the battery was fully charged, energy must have been dissipated into the braking resistor. In the simulation, two cases when the battery management SYSTEM activated the braking circuit can be seen in the figure after $T_1 = 1.7\,\mathrm{s}$ and $T_2 = 4.4\,\mathrm{s}$. When the current flowed through the resistor the battery current was zero due to decoupling performed by switch M_2. In the figure, a current spike can be easily observed when the battery was connected to the circuit after reference current i_q became positive.

6. Conclusions

This paper presents models of the power conversion system of an autonomous electric vehicle as well as the electronic control unit. An interconnected simulation environment enables an integrated model to be simulated. This kind of simulation is useful for the comparative performance analysis of different control algorithms as well

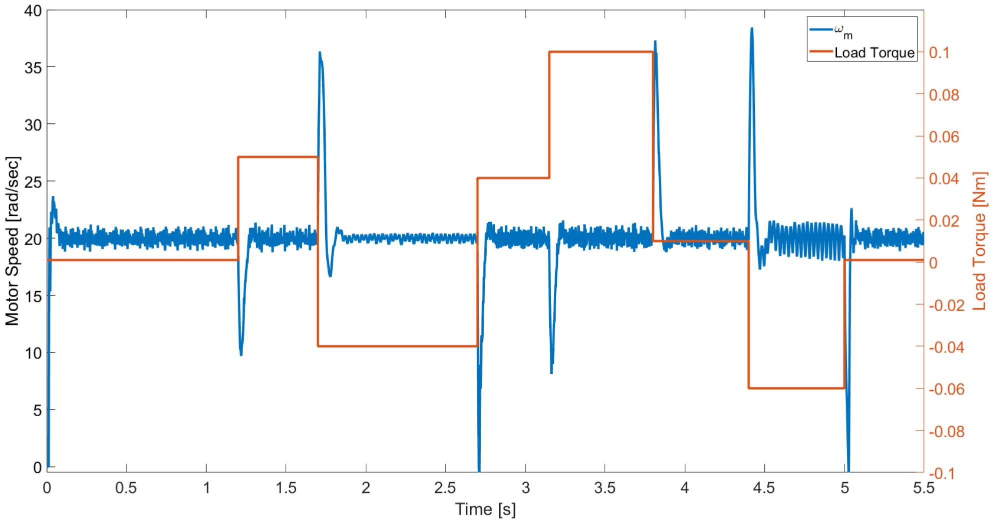


Figure 11: Simulated motor speed during the driving cycle of a vehicle

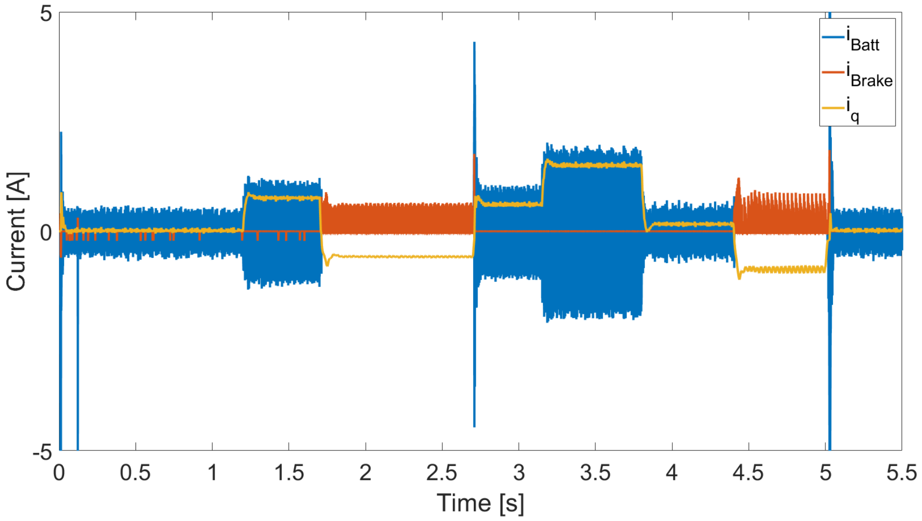


Figure 12: The simulated currents of the power conversion system with BMS

as system level analysis of interactions between different system modules and components.

A power conversion system model is very complex and it is hard to predict interactions between components of the system following dynamic changes at system inputs. Furthermore, the validation of the controllers is only plausible if the simulation model includes all relevant components of the system and all significant factors such as sampling of the measured data or time delays in the control loop. The simulation is very useful if the operation of the hardware level system is the focus of observations and analyses.

However, the simulation of the top-level management system of an autonomous electric vehicle is not feasible due to lengthy simulation times. In order to analyze the performance of different controllers, the simulation must be run with a sampling time of 2 μs or less, while the effectiveness of a higher-level management system can be measured after a few hours of driving. In cases when the analysis of the efficiency of the management system takes into consideration the performance of the digital controller, it is essential to identify a modeling procedure which provides a model of the system that is suitable for long-time simulation.

Acknowledgements

The project has been supported by the European Union, co-financed by the European Social Fund. (EFOP- 3.6.2-16-2017-00002).

REFERENCES

[1] Kohlrusz, G.; Csomós, B.; Enisz, K.; Fodor, D.: Electric energy converter development and diagnostics in mixed-signal simulation environment, *ACTA Imeko*, 2018, **7**(1), 20–26 DOI: 10.21014/acta_imeko.v7i1.512

[2] Hang, J.; Zhang, J.; Cheng, M.; Huang, J.: Online Interturn Fault Diagnosis of Permanent Magnet

Synchronous Machine Using Zero-Sequence Components, *IEEE Transactions on Power Electronics*, 2015, **30**(12), 6731–6741 DOI: 10.1109/tpel.2015.2388493

[3] Qi, Y.; Bostanci, E.; Gurusamy, V.; Akin, B.: A Comprehensive Analysis of Short-Circuit Current Behavior in PMSM Interturn Short-Circuit Faults, *IEEE Transactions on Power Electronics*, 2018, **33**(12), 10784–10793 DOI: 10.1109/tpel.2018.2809668

[4] Kohlrusz, G.; Szalay, I.; Fodor, D.: OrCAD PSpice Implementation of a Realistic Three-Phase PMSM Model for Diagnostic Purposes, San Diego, CA, USA, 2019, 372–376 DOI: 10.1109/IEMDC.2019.8785371

[5] LeVeque, R. J.: Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA), 2007 ISBN: 978-0-898716-29-0

[6] Åström, K. J.; Hägglund, T.: PID controllers : theory, design, and tuning (2nd edition) (International Society for Measurement and Control, Research Triangle Park, N.C., USA), 1995 ISBN: 1556175167 9781556175169

[7] Kuo, B. C.; Golnaraghi, F.: Automatic Control Systems (John Wiley and Sons Inc., New York, NY, USA), 2002 ISBN: 0470048964

[8] Katsuhiko, O.: Modern Control Engineering, (5th Edition) (Pearson), 2009 ISBN: 0136156738

[9] Guzmán, J. L.; Moreno, J. C.; Berenguel, M.; Moscoso, J.: Inverse pole placement method for PI control in the tracking problem, in *IFAC-PapersOnLine*, **51**(4), 406–411 DOI: 10.1016/j.ifacol.2018.06.128

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 61–65 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-09

NONLINEAR MAGNETIC PROPERTIES OF MAGNETIC FLUIDS FOR AUTOMOTIVE APPLICATIONS

BARNABÁS HORVÁTH*[1] AND ISTVÁN SZALAI[2]

[1]Institute of Physics and Mechatronics, University of Pannonia, Egyetem utca 10, Veszprém H-8200, HUNGARY
[2]Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk utca 18/A, Zalaegerszeg H-8900, HUNGARY

The external magnetic field required to activate a magnetic fluid in an industrial application is sufficiently large that magnetization is no longer a linear function of the external field strength, i.e. magnetic fluids exhibit nonlinear characteristics. The aim of our research was to develop a measuring system which is capable of determining the nonlinear AC susceptibility of magnetic fluids at discrete frequencies and in the presence of a high-intensity driving magnetic field. The measurement of susceptibility is based on the determination of the change in frequency of a low-intensity field, which is generated by an LC oscillator. The application of sinusoidal excitation to the material results in a variation in the susceptibility that modulates the frequency of the measured low-intensity field and in the appearance of higher-order harmonics of the driving field. The higher-order components of the nonlinear AC susceptibility are extracted from the measured response by Fourier analysis. By applying the measuring system, the nonlinear susceptibility of water-based ferrofluids (Ferrotec's EMG 700) and its dependence on the magnetic field strength were investigated.

Keywords: AC susceptometry, nonlinear susceptibility, ferrofluid

1. Introduction

Regarding automotive applications, one of the most important characteristics of magnetic fluids is their behavior in magnetic fields. The majority of these applications are for seals [1], vibration damping [2] and torque transmission [3]. In most cases, the magnitude of an external magnetic field required to activate a fluid is on the scale where the **M** magnetization is no longer a linear function of the **H** field strength. This means that magnetic fluids show nonlinear characteristics, and the $\chi = \partial M/\partial H$ susceptibility depends on the magnetic field strength. In the case of magnetic fluids, the nonlinearity is the result of two effects: normal saturation (alignment of the magnetic dipole moments) and formation of structures (particle chains). The latter influences the susceptibility as the magnetic field shifts the equilibrium between the structures with different dipole moments (single particles and particle chains of different sizes).

In a weak time-varying $H_\mathrm{e}(t) = H_{\mathrm{e}0}\sin(\omega t)$ magnetic field (where t denotes the time and $\omega = 2\pi f$ represents the angular frequency), the magnetization also changes periodically but lags behind the magnetic field because reorientation of the magnetic dipole moments is not instantaneous. In this case, the dynamic magnetic susceptibility can be defined as a complex quantity $\chi^*(\omega) = \chi'(\omega) - i\chi''(\omega)$. The real part, χ', is related to the reversible magnetization process and it is in-phase with the alternating field. Within the low-frequency limit, $(f \to 0)$ χ' approaches the initial gradient of the steady-state magnetization (initial DC susceptibility). If $f \to \infty$, the reorientation of the magnetic dipole moments cannot follow the alternating field and χ' approaches zero. The imaginary, out-of-phase component, χ'', is proportional to power losses due to energy absorption from the field and peaks at a characteristic frequency, f_c.

However, if the amplitude of the alternating field is sufficiently large, higher-order harmonics appear in the magnetization due to the nonlinear characteristics. In this case, the real part of the alternating current (AC) magnetic susceptibility is:

$$\chi' = \chi_0 + \chi_{2\omega}\cos(2\omega t) + \chi_{4\omega}\cos(4\omega t) + \ldots, \qquad (1)$$

where χ_0 denotes the base component, and $\chi_{2\omega}$ and $\chi_{4\omega}$ represent the amplitude of the second- and fourth-order harmonics, respectively [4]. A similar equation holds for the imaginary component, χ''. In this work, only the harmonics of the real part are considered, because the nonlinear susceptibility at frequencies much higher than the characteristic frequency of the magnetic fluid is investigated, where the imaginary part is very close to zero. By taking magnetic fluids into consideration, the amplitudes of sixth- and higher-order harmonics are so small

*Correspondence: bhorvath@almos.uni-pannon.hu

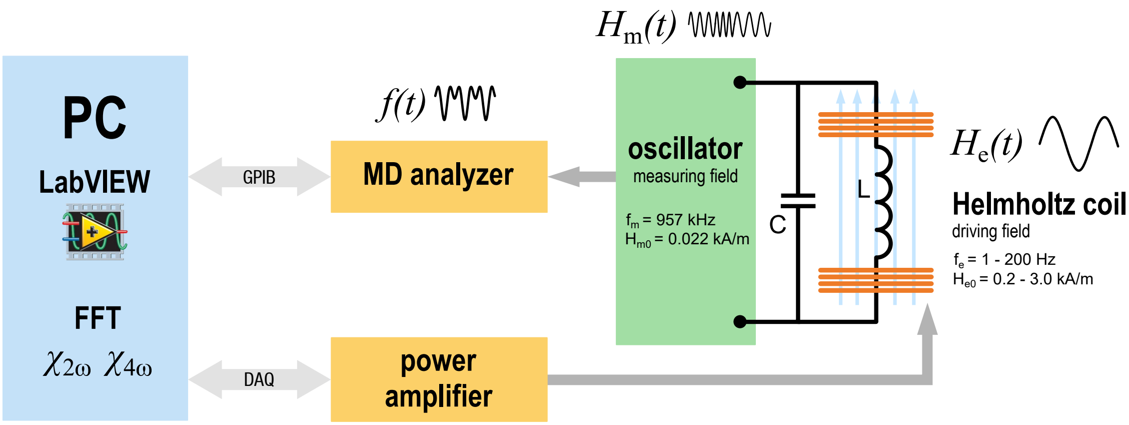


Figure 1: A block diagram of the nonlinear magnetic measuring system.

that they can be neglected. In alternating magnetic fields, only the even-order harmonics appear, because the magnetization of magnetic fluids exhibits inversion symmetry with respect to a change in the direction of **H**. Should a symmetry-breaking DC bias field be superimposed on the AC magnetic field, both the odd- and even-order harmonics appear. As the nonlinearity is caused partly by structure formation, the components of the nonlinear susceptibility and their ratios are sensitive to changes in the structure of magnetic fluids.

Numerous studies have discussed the theoretical description of the nonlinear susceptibility of magnetic fluids [4–7], but experimental results are rather scarce. Nevertheless, nonlinear magnetic susceptibility measurements are quite useful for the study of magnetic fluids and provide additional information besides linear AC susceptometry. If the linear response to small-amplitude fields is measured, the relaxation processes of the magnetic dipoles can be studied. However, by applying the nonlinear method, it is possible to obtain information about the structural evolution of an activated magnetic fluid.

The aim of our work was to develop a measuring system which is capable of determining the nonlinear AC susceptibility of magnetic fluids at discrete frequencies in the presence of external magnetic fields. The dependence of the 2nd- and 4th-order nonlinear AC susceptibility of ferrofluids on the magnetic field strength was investigated by the developed susceptometer.

2. Experimental

2.1 *Nonlinear measurement method and setup*

The measurement method of susceptibility is based on our nonlinear dielectric system [8], which is adapted to magnetic measurements. With this method, the real part of complex AC susceptibility is determined from the change in frequency of a low-intensity measuring field (H_m). An air core solenoid (L) was filled with the sample of fluid, and the inductance determined the frequency of the sinusoidal measuring field generated by an LC oscillator (Fig. 1). Therefore, any change in the susceptibility of the sample caused the resonance frequency f_m to shift. The susceptibility of the sample was measured at this base frequency. Inside the solenoid, the amplitude of the measuring field is so small ($H_{\mathrm{m}0} = 0.022$ kA/m) that within this region, the magnetization curve of the fluid can be regarded as linear, thus no significant structural change can be expected.

The oscillator which generated the measuring field was a Colpitts-type parallel LC circuit. The active feedback element was a double triode vacuum tube (ECC88). The capacitor bank C was composed of high-quality silver mica capacitors with a low temperature coefficient. The measuring coil L was connected in parallel with the capacitor block, and by changing the coil, the base frequency of the oscillator could be varied. With different inductors, the measuring frequency was set to discrete values, namely 153 kHz, 590 kHz and 957 kHz. The dimensions of the measuring coils were identical: 25 mm in length with an inner diameter of 7.1 mm. They were composed of enameled copper wires of different sizes and wound on coil formers made of plexiglass. The resonance frequency of the LC oscillator was measured by a Hewlett Packard 53310A Modulation Domain (MD) Analyzer. The relationship between the resonance frequency and susceptibility of the sample was determined by calibration using different materials of known susceptibility.

A high-intensity driving field (H_e) was generated by a pair of Helmholtz coils which were placed around the measuring coil. The axis of the Helmholtz coil and, therefore, the direction of the field, were parallel to the axis of the solenoid. The upper limit on the amplitude of the driving magnetic field with the current setup was $H_{\mathrm{e}0} = 6.4$ kA/m, which was two orders of magnitude greater than the amplitude of the measuring magnetic field. The maximum of $H_{\mathrm{e}0}$ depended on the frequency of the field: as the frequency increased, the maximum of $H_{\mathrm{e}0}$ decreased. The uniformity of the magnetic field strength in the volume of the sample was better than 1 %. The Helmholtz coil was driven by a high-current function generator, which consisted of a Labworks PA-138 linear power amplifier and a signal source. The input signal of the amplifier was provided by a multifunction data acquisition (DAQ) card (National Instruments PCI-6052E). Any arbitrary waveform could be generated by the signal generator, and the waveform offset by a DC value.

Under the influence of the high-intensity driving field, the susceptibility of the sample changes, therefore, the frequency of the measuring field was modulated. If the driving field is sinusoidal, then the time-domain susceptibility response will contain the higher-order harmonics of the field. The 2nd- and 4th-order components of the nonlinear AC susceptibility were extracted from the measured response by Fourier analysis. The Fast Fourier Transform (FFT) algorithm was implemented using custom-developed LabVIEW software, which provided control and data acquisition functions of the measuring system.

2.2 Supplemental measurements

During the nonlinear measurements, the change in susceptibility relative to the zero-field susceptibility was measured, so the real part of the AC susceptibility of the fluid at the base frequency had to be determined by another method. For this purpose, a spectrum was obtained within the frequency range of 200 Hz - 1 MHz. The initial DC susceptibility of the ferrofluid was determined from the DC magnetization curve. The AC susceptibility in zero field was measured by an inductive method where the inductance of a solenoid filled with the sample was measured by an impedance analyzer (Agilent 4284A). By determining the impedances of the empty air core solenoid and when it is filled with the sample, χ' was calculated. The solenoid used for these measurements was the same as the measuring coil of the nonlinear setup.

2.3 Material

By applying the nonlinear magnetic measuring system, the nonlinear properties of Ferrotec's EMG 700 ferrofluid were investigated. This material is water-based and contains magnetite particles with a nominal diameter of ~ 10 nm. The volume concentration of the particles is 5.8 %(v/v). The fluid is stabilized by an anionic surfactant. For the measurements, glass-tube sample holders with an inner diameter of 3.1 mm were filled with the ferrofluid. The length of the tube extended beyond the length of the measuring coil at both ends. The volume of the sample was 0.49 cm^3.

3. Results and Discussion

According to the DC magnetization curve, the initial DC susceptibility of the EMG 700 ferrofluid was $\chi_{\mathrm{DC}} = 12.57$. This is the limiting value of the real part of the complex susceptibility if f approaches zero. The AC susceptibility spectrum of the ferrofluid is shown in Fig. 2. Relaxation of χ' was observed within the lower frequency region. The experimental frequency-dependent susceptibility data was fitted by the Cole-Cole equation (solid line in Fig. 2). This relaxation model is suitable for describing the response of dipoles to an alternating field should the relaxation not be ideal (e.g. if instead of a single relaxation time, a distribution of relaxation times exists).

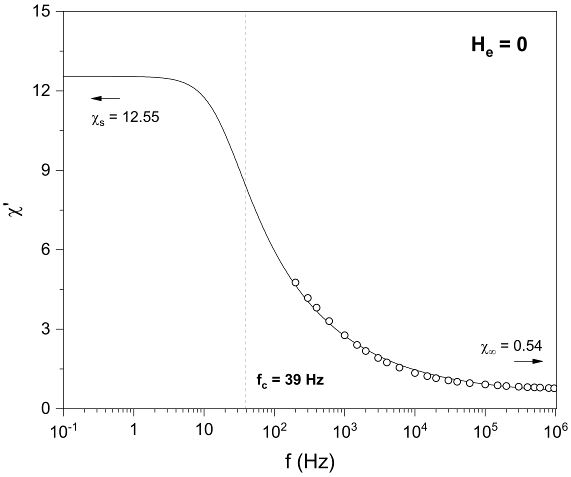


Figure 2: Relaxation of the real part of AC susceptibility in the absence of a driving magnetic field (hollow symbols) by applying the Cole-Cole equation (solid line) together with the estimated parameters of the relaxation.

The fitting process yielded the static ($f \rightarrow 0$) $\chi_{\mathrm{s}} = 12.55 \pm 0.07$ and infinite frequency ($f \rightarrow \infty$) $\chi_\infty = 0.54 \pm 0.02$, susceptibilities and the so-called central characteristic time of the relaxation (τ_0). The reciprocal of the characteristic time yielded the characteristic frequency $f_{\mathrm{c}} = 1/(2\pi\tau_0)$, which is 39 ± 5 Hz for the ferrofluid EMG 700. The spectrum shows that relaxation occurs at much lower frequencies than the base frequency of the nonlinear susceptibility measurements ($f_{\mathrm{c}} = 39$ Hz vs. $f_{\mathrm{m}} = 957$ kHz). At the high-frequency end ($\sim$1 MHz) of the spectrum, χ' decreased to ~ 0.8 and changed slightly as the frequency increased in this region. At such a large distance from f_{c}, the imaginary part of the complex susceptibility was close to zero, thus it is justified to consider only the real part (see Eq. 1).

To investigate the nonlinear AC susceptibility, sinusoidal excitation was applied. The frequency of the driving magnetic field was changed from 1 Hz to 200 Hz with an amplitude as high as $H_{\mathrm{e}0} = 3$ kA/m (depending on f_{e}). The nonlinear AC susceptibility was measured at 957 kHz in all cases. At $H_{\mathrm{e}0} = 0$, the real part of the AC susceptibility of the ferrofluid at the measurement frequency of $f_{\mathrm{m}} = 957$ kHz was $\chi' = 0.78$. Fig. 3 shows a typical susceptibility response of the ferrofluid EMG 700 at an excitation frequency of $f_{\mathrm{e}} = 1$ Hz and using different magnetic field strengths. It can be seen that the response contained higher-order harmonics. The magnetic field always caused a decrease in the susceptibility (because of the saturation) regardless of its direction. This inversion symmetry was expressed by the fact that the susceptibility response during the first half period

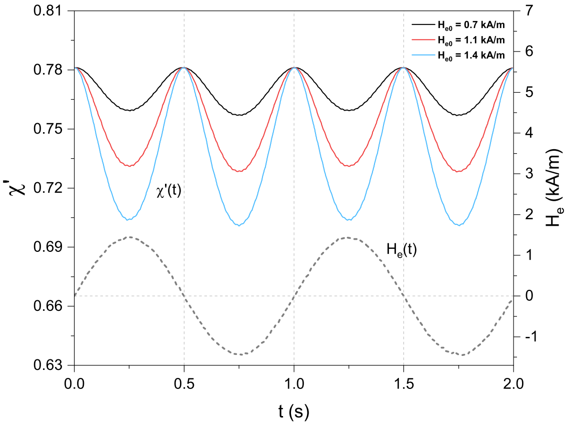


Figure 3: The susceptibility response of the ferrofluid EMG 700 at various driving field amplitudes (f_e = 1 Hz).

of $H_e(t)$ was the same as during the second half even if the direction of the field was reversed. Therefore, the maxima of the susceptibility responses were always 0.78, which corresponded to the point when $H_e(t)$ intersected the value zero.

The dependence of the 2nd- and 4th-order harmonics ($\chi_{2\omega}$ and $\chi_{4\omega}$ in Eq. 1) on the magnetic field strength extracted from the measured response is shown in Figs. 4a and 4b. At a given driving field strength, $\chi_{2\omega}$ was one order of magnitude larger than $\chi_{4\omega}$. The amplitude of both components increased as the magnetic field strength rose (at a given driving field frequency). This occurred because at higher magnetic field strengths, the nonlinearity of the magnetization curve increased as the magnetization approached the saturation level.

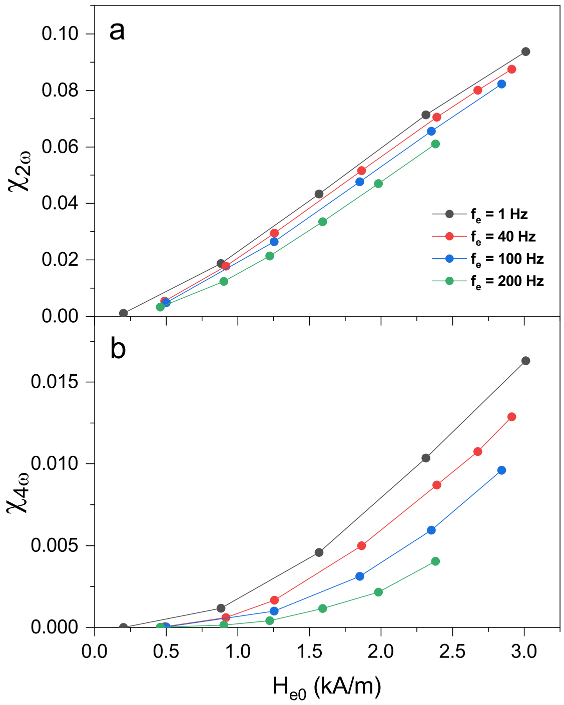


Figure 4: The dependence of the components $\chi_{2\omega}$ (a) and $\chi_{4\omega}$ (b) of the nonlinear susceptibility of the ferrofluid EMG 700 on the amplitude of the driving field at different driving field frequencies.

4. Conclusions

A nonlinear magnetic measuring system was developed to determine the nonlinear AC susceptibility of magnetic fluids. The system was applied to measure the components of the nonlinear susceptibility of the water-based ferrofluid EMG 700 up to the 4th order. It has been shown that by using the aforementioned system, it is possible to determine the dependence of the nonlinear susceptibility on the magnetic field strength. If the excitation frequency is varied, the relaxation of the nonlinear susceptibility can be also studied. Our future aim is to investigate this relaxation and other types of ferrofluids (based on organic carrier and polydisperse fluids), moreover, to study the influence of the magnetite concentration on the nonlinear magnetic properties.

Acknowledgments

This project was supported by the European Union and co-financed by the European Social Fund. EFOP-3.6.2-16-2017-00002.

REFERENCES

[1] Ravaud, R.; Lemarquand, G.: Mechanical Properties of a Ferrofluid Seal: Three-Dimensional Analytical Study based on the Coulombian Model, *Prog. Electromagn. Res. B*, 2009, **13**, 385–407, DOI: 10.2528/PIERB09020601

[2] Sassi, S.; Cherif, K.; Mezghani, L.; Thomas, M.; Kotrane, A.: An innovative magnetorheological damper for automotive suspension: from design to experimental characterization, *Smart Mater. Struct.*, 2005, **14**(4), 811–822, DOI: 10.1088/0964-1726/14/4/041

[3] Rizzo, R.: An innovative multi-gap clutch based on magneto-rheological fluids and electrodynamic effects: magnetic design and experimental characterization, *Smart Mater. Struct.*, 2017, **26**(1), 015007, DOI: 10.1088/0964-1726/26/1/015007

[4] Wang, G.; Huang, J.P.: Nonlinear magnetic susceptibility of ferrofluids, *Chem. Phys. Lett.*, 2006, **421**(4), 544–548, DOI: 10.1016/j.cplett.2006.02.010

[5] Huke, B.; Lücke, M.: Magnetic properties of colloidal suspensions of interacting magnetic particles, *Rep. Prog. Phys.*, 2004, **67**(10), 1731–1768, DOI: 10.1088/0034-4885/67/10/R01

[6] Szalai, I.; Nagy, S.; Dietrich, S.: Linear and nonlinear magnetic properties of ferrofluids, *Phys. Rev. E*, 2015, **92**(4), 042314, DOI: 10.1103/PhysRevE.92.042314

[7] Ivanov, A.O.; Kuznetsova, O.B.: Magnetic properties of dense ferrofluids, *J. Magn. Magn. Mater.*, 2002, **252**, 135–137, DOI: 10.1016/S0304-8853(02)00639-X

[8] Horváth, B.; Szalai, I.: Structure of electrorheological fluids: A dielectric study of chain formation, *Phys. Rev. E*, 2012, **86**(6), 061403, DOI: 10.1103/PhysRevE.86.061403

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 67–70 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-10

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
HJIC

VIBRATION GENERATOR DEVICE BASED ON INDUSTRIAL VIBRATORS

PÉTER DECSI *[1] AND ISTVÁN SZALAI[1]

[1]Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk utca 18/A, Zalaegerszeg, 8900, HUNGARY

A low-cost vibration generator device based on industrial vibrators was designed. The control software was implemented in LabVIEW Environment. The device is able to generate an oscillating force of 8 kN and an amplitude of up to 4 mm at a frequency of 50 Hz to model low-amplitude, high-frequency vehicle vibrations. A National Instruments myRIO device was responsible for data acquisition, with which a signal of a piezoelectric accelerometer was detected. The test results show that the device is able to generate a sinusoidal harmonic acceleration.

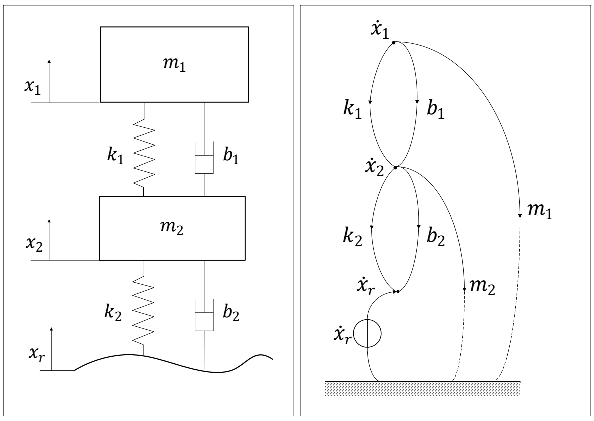
Keywords: industrial vibrator, vibration generator, test equipment

1. Introduction

The suspension of a car acts as a connector between the chassis and the road. It provides a comfortable ride for passengers while maintaining maneuverability. Nowadays, comfort is a prerequisite so car manufacturers have started developing new technologies for isolating vibrations [1, 2]. The vibration can be dampened or converted into electrical energy [3]. Electro- [4] and magnetorheological dampers [5, 6] are filled with a special fluid. Small ferromagnetic particles are dispersed in a carrier fluid, usually silicone oil. The diameters of the particles fall within the micro- and nanometre ranges. When subjected to an electric or magnetic field, the particles form chains. These chains provide an elevated level of resistance against shear stress, therefore, the apparent viscosity of the fluid rises [7, 8].

Using adjustable shock absorbers, the natural frequency of the suspension can be altered. The damping coefficient can be varied depending on the road conditions or the driver's preference. The reaction time of these fluids is very short (approximately $10\,\mathrm{ms}$), so the technology can be used in systems where short reaction times are required [9].

Suspension-testing equipment is crucial during the development phase. The acquisition cost of appliances for this task is very high. Our goal was to develop a low-cost piece of equipment for testing shock absorbers that is capable of generating high frequencies (within an acceptable range for vehicles) and vibrations of low amplitude.



(a) structure model (b) structure graph

Figure 1: Quarter-car model

2. Modeling the vehicle

Several options of modeling the suspension of vehicles are available; quarter-, half- and full-car models can be used to describe the system, which is chosen depends on the aim of the study. The quarter-car model was chosen for this study due to its simplicity. Industrial vibrators are able to generate vibrations of constant amplitude. To change the generated force and amplitude, it is necessary to stop the device and mechanically set the eccentricity, therefore, the roll, pitch and yaw of the vehicle is hard to model with such a device. Fig. 1 shows a quarter-car model. The system has two degrees of freedom (DOF), m_1 denotes the mass of the chassis, m_2 represents the mass of the wheel, which is usually referred to as the unsprung mass. The related displacements are labelled as x_1

*Correspondence: decsi.peter@mk.uni-pannon.hu and x_2, spring stiffnesses as k and viscous damping coefficients as b. The wheel and chassis are connected through a spring k_1 and a damper b_1. The wheel and the road are connected through a spring and damper (elastic tire with damping loss) k_2, b_2. Based on the structure graph, the equilibrium equations can be written in the form:

$$\begin{aligned} m_1\ddot{x}_1 &= -b_1(\dot{x}_1 - \dot{x}_2) - k_1(x_1 - x_2) \\ m_2\ddot{x}_2 &= b_1(\dot{x}_1 - \dot{x}_2) + k_1(x_1 - x_2) + \\ &\quad + b_2(\dot{x}_r - \dot{x}_2) + k_2(x_r - x_2). \end{aligned} \tag{1}$$

Based on Fig. 1 and Eq. 1, it is clear that the system has a cross-variable source, namely speed.

3. Modeling the proposed system

This research presents an equivalent model based on industrial vibrators. Industrial vibrators are traditional asynchronous motors with a dual shaft on which eccentric masses are mounted.

$$f_c = mr\omega^2 \tag{2}$$

where m denotes the eccentric mass, r represents the eccentricity, which is the distance between the axis center point and the center of gravity of the eccentric mass, and ω stands for the rotational speed. Due to the eccentricity, a centrifugal force is generated because of the rotation (Eq. 2). If two motors of opposite rotational directions are mounted together, lateral forces cancel each other out, therefore, a one-axis oscillation is created. A system with two degrees of freedom was designed using industrial vibrators. The model of the system is depicted in Fig. 2: The differential equations of this system are:

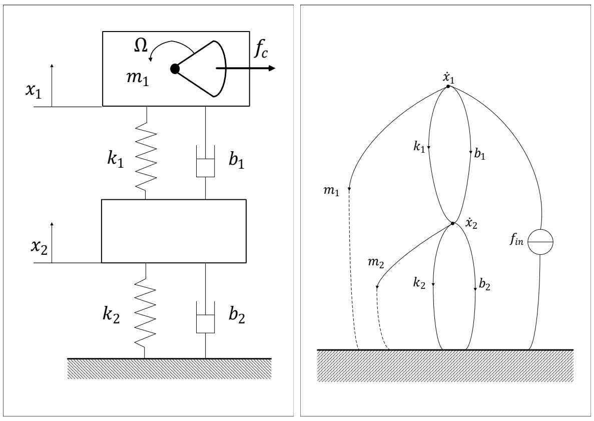
$$\begin{aligned} m_1\ddot{x}_1 &+ k_1(x_1 - x_2) + \\ &+ b_1(\dot{x}_1 - \dot{x}_2) + f_{in} = 0 \\ m_2\ddot{x}_2 &- k_1(x_1 - x_2) - \\ &- b_1(\dot{x}_1 - \dot{x}_2) + k_2x_2 + \dot{x}_2b_2 = 0 \end{aligned} \tag{3}$$

$$\frac{X_1}{X_r} = \frac{s^2b_1b_2 + s(k_2b_1 + b_2k_1) + k_1k_2}{s^4m_1m_2 + s^3(m_1(b_1 + b_2) + m_2b_1) + s^2(m_1(k_1 + k_2) + m_2k_1 + b_1b_2) + s(k_1b_2 + k_2b_1) + k_1k_2} \tag{4}$$

$$\frac{X_1}{F} = \frac{s^2m_2 + s(b_1 + b_2) + k_1 + k_2}{s^4m_1m_2 + s^3(m_1(b_1 + b_2) + m_2b_1) + s^2(m_1(k_1 + k_2) + m_2k_1 + b_1b_2) + s(k_1b_2 + k_2b_1) + k_1k_2} \tag{5}$$

According to the impedance network of the quarter-car model shown in Fig. 3, the transfer function can be written in the form Eq. 4. The impedance network of the proposed system is shown in Fig. 4 and contains the transfer function of Eq. 5:

It can be seen that the transfer function X_1/X_r of the quarter-car model (Eq. 4) is similar to that of the transfer



(a) structure model *(b)* structure graph

Figure 2: Vibrator model

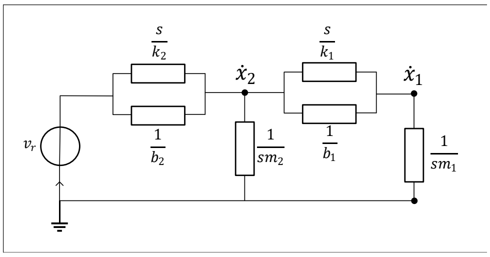


Figure 3: Impedance network of the quarter-car model

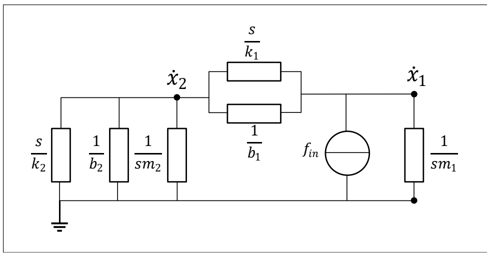


Figure 4: Impedance network of the vibrator model

function X_1/F of the proposed system (Eq. 5). The characteristic polynomials, i.e. the denominators, are identical as the two systems differ only in terms of the excitation source. The numerators are of the same degree, therefore, the characteristics of the system response are similar, that is to say the dynamics of the two systems are not significantly different. The breakpoints in terms of the frequency response are expected to be shifted.

4. Implementation

After deriving the differential equations, a Simulink model was created to choose the optimal combinations of parameters (mass, spring stiffness and damping coefficients). According to the simulations, the connection between the chassis and the ground should be as stiff as possible in order to focus the force on the shock absorbers. It should not be too rigid, otherwise the force acting on the ground would be excessive and affect the building too significantly. The moving mass, where the motors are mounted, should be small to achieve the maximum possible exciting force, therefore, more force can be used to accelerate the payload.

During the design process, models with two or three degrees of freedom were examined. The motion became chaotic with three independent masses, that is 3 DOFs, so the spring stiffness available was insufficient to produce a stable and predictable shape of sinusoidal motion. It was concluded that a system with 2 DOFs generates a stable sinusoidal oscillation with multiple frequency components in terms of the shape of motion.

The device, shown in Fig. 5, consists of two three-phase asynchronous two-pole industrial vibrators each with a nominal performance of $300\,\mathrm{W}$. The centrifugal force can be incrementally set up to $4070\,\mathrm{N}$ at a frequency of $50\,\mathrm{Hz}$ on one motor. The motors are operated with frequency converters, thus the frequency can be set. If the motors operate in opposite rotational directions and a suitable arrangement is applied, the lateral forces cancel each other out, thereby creating a one-axis oscillation. The frequency converters are controlled with analogue signals generated by a National Instruments myRIO device.

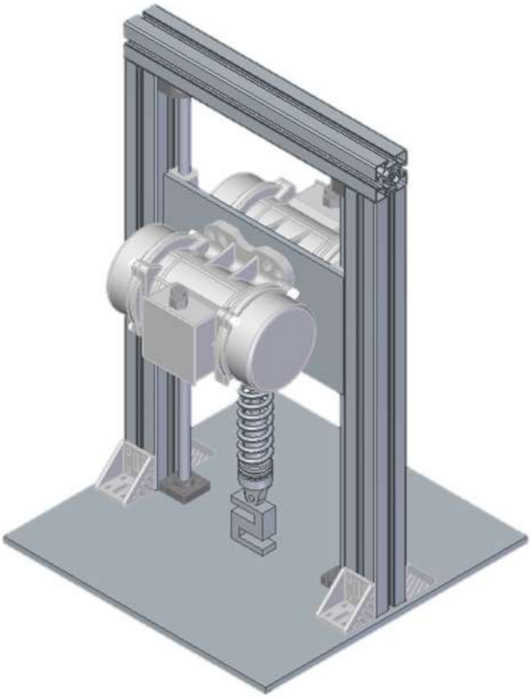


Figure 5: The implemented vibration generator

A piezoelectric accelerometer was mounted on the sheet holding the vibrating motors. A high-pass filter with a cutoff frequency of $1\,\mathrm{Hz}$ was installed to cancel out the bias voltage of the accelerometer. Data was acquired by the Field-Programmable Gate Array (FPGA) module of the myRIO device on analogue channels. The AC signal was coupled to a high-pass filter with a cutoff frequency of $1\,\mathrm{Hz}$. The sensitivity of the accelerometer was $100\,\mathrm{mV/g}$ and the analogue input of the myRIO device was $\pm 10\,\mathrm{V}$. An instrumentation amplifier with an amplification of 7.08 was built to utilize the full range of the A/D converter. The FPGA module takes a sample according to the previously set sampling frequency, which can be set up to $100\,\mathrm{kHz}$. The data is stored temporarily in the First In First Out (FIFO) memory on the myRIO device.

A second Virtual Instrument (VI), which displays and stores the data in Technical Data Management Streaming (.TDMS) file format, was run on the controlling PC. The VI read out data from the FIFO memory in batches. These batches were stored in the TDMS file, creating a reliable data acquisition. Meanwhile, following the application of a Fast Fourier Transform algorithm, a frequency spectrum was displayed on the controlling PC.

5. Test results

A test measurement was taken after the implementation. For test purposes, the frequency converters were set at $16\,\mathrm{Hz}$. Fig. 6 shows a long-term test and Fig. 7 shows the measurement of a short-term acceleration. It can be seen that the acceleration is approximately sinusoidal and consists of two main frequency components. Fig. 6 shows that the acceleration was stable over an extended period of time with several protrusions. Fig. 8 shows that the set

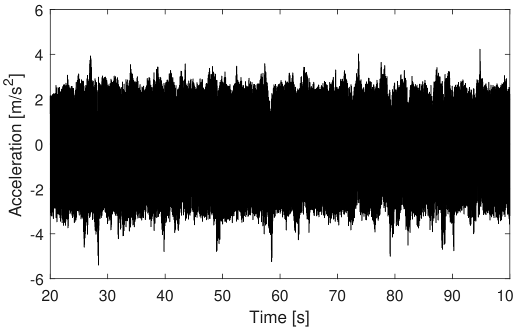


Figure 6: Test measurement, long duration

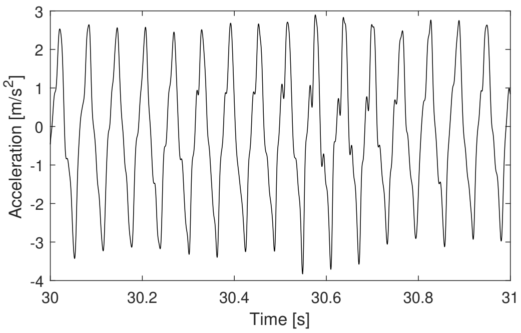


Figure 7: Test measurement, short duration

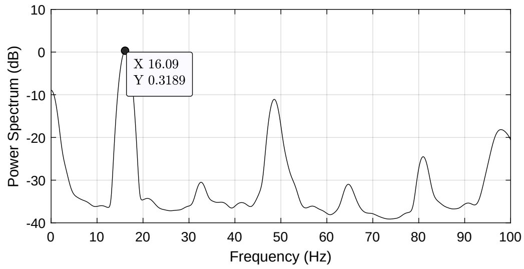


Figure 8: Frequency spectrum of the test measurement

frequency was present in the spectrum alongside multiple frequency components.

6. Conclusion

According to our experience and the aforementioned results, it can be concluded that a cost-effective vibration generator was designed and produced. The device is able to generate an amplitude of vibration equal to 4 mm with a frequency of up to 50 Hz. The total centrifugal force was as high as 8 kN, of which 7500 N could be focused on the examined sample. The frequency can be set between 1 Hz and 50 Hz, however, as shown in Eq. 2, the centrifugal force depends on the rotational speed.

The amplitude cannot be set, and when the natural frequency of the system, namely 1.3 Hz, is exceeded, the amplitude becomes constant. At this frequency, the centrifugal force is low so resonance can be avoided. The excitation force can be set between 0 and 100% by modifying the eccentricity of the motors. To set the eccentricity, the motors must be powered down and the side covers removed.

The vibration generators that are typically used (electrodynamic and hydraulic) are able to create higher amplitudes and higher vibrational frequencies, but these devices belong to a different cost category.

Acknowledgement

The project has been supported by the European Union, co-financed by the European Social Fund. EFOP-3.6.2-16-2017-00002

REFERENCES

[1] Qin, Y.; Tang, X.; Jia, T.; Duan, Z.; Zhang, J.; Li, Y.; Zheng, L.: Noise and vibration suppression in hybrid electric vehicles: State of the art and challenges, *Renew. Sustain. Energy Rev.*, 2020, **124**, DOI: 10.1016/j.rser.2020.109782

[2] Ning, D.; Sun, S.; Du, H.; Li, W.; Li, W.: Control of a multiple-DOF vehicle seat suspension with roll and vertical vibration, *J. Sound Vib.*, 2018, **435**, 170–191, DOI: 10.1016/j.jsv.2018.08.005

[3] Zhang, Z.; Xiang, H.; Shi, Z.; Zhan, J.: Experimental investigation on piezoelectric energy harvesting from vehicle-bridge coupling vibration, *Energy Convers. Manag.*, 2018, **163**, 169–179, DOI: 10.1016/j.enconman.2018.02.054

[4] Holzmann, K.; Kemmetmüller, W.; Kugi, A.; Stork, M.: Design, mathematical modeling and control of an assymetrical electrorheological damper, *IFAC Proceedings Volumes*, 2006, **39**(16), 372–377, DOI: 10.3182/20060912-3-DE-2911.00066

[5] Graczykowski, C.; Pawłowski, P.: Exact physical model of magnetorheological damper, *Appl. Math. Model.*, 2017, **47**, 400–424, DOI: 10.1016/j.apm.2017.02.035

[6] Yao, G.Z.; Yap, F.F.; Chen, G.; Li, W.H.; Yeo, S.H.: MR damper and its application for semi-active control of vehicle suspension system, *Mechatronics*, 2002, **12**(7), 963–973, DOI: 10.1016/S0957-4158(01)00032-0

[7] Carlson, J.D.: What Makes a Good MR Fluid?, *J. Intel. Mat. Syst. Str.*, 2002, **13**(7-8), 431–435, DOI: 10.1106/104538902028221

[8] Rankin, P.J.; Ginder, J.M.; Klingenberg, D.J.: Electro- and magneto-rheology, *Curr. Opin. Colloid Interface Sci.*, 1998, **3**(4), 373–381, DOI: 10.1016/S1359-0294(98)80052-6

[9] Olabi, A.; Grunwald, A.: Design and application of magneto-rheological fluid, *Mater. Des.*, 2007, **28**(10), 2658–2664, DOI: 10.1016/j.matdes.2006.10.009

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 71–75 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-11

DESIGN AND CONSTRUCTION OF A VIBRATION DATA ACQUISITION SYSTEM FOR ROAD VEHICLES

Roland Nagy*[1] and István Szalai[1]

[1] Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk utca 18/A, Zalaegerszeg, 8900, HUNGARY

This article describes the design of a vibration data acquisition system which can be mounted on the undercarriage of a vehicle to acquire information about the quality of and defects in road surfaces. It is important to be able to deduce the condition of a road section from its data. For practical reasons, a microcontroller-based control unit was used and a separate power supply created. Bumps in the road were detected by a piezoelectric accelerometer. Once the system was completed, different measurements were made and the results analyzed. According to the results, it can be stated that the whole system worked well since they are identical to reality. The bumps in the road were clearly visible on the diagrams. It was concluded that the completed vibration data acquisition system is more than capable of detecting bumps in roads. The advantage of the system is that it can be easily mounted on any car which does not need to be driven at low speeds.

Keywords: road vehicle, vibration, data acquisition, microcontroller, accelerometer

1. Introduction

Road vehicles are continuously subjected to vibrations which have very important and serious consequences. These vibrations should be dampened as they have a detrimental effect on both mechanical properties and vehicle occupants. Vibrations can significantly reduce the lifetime of mechanical and electrical components. The effects of degradation are greatly increased if the frequency of the vibration is identical to that of a component. These resonances cause health issues for passengers in the car, e.g. back pain is a frequent complaint. Vibrational frequencies of between 4 and 8 Hz are the most dangerous for the human body [1].

Such vibrations can be caused by poor road surfaces and the engine of a car. The former reason is more significant, so this issue will be tackled. Due to the important role of vibrations that can occur in road vehicles, it is important to know the condition and characteristics of a road network.

In Hungary, the measurement of road quality is based almost exclusively on visual observations and manual data recording. Inspectors observe the roads and record any defects in a database. Even though hardly any automated survey equipment is used, these vehicles have the advantage of faster assessment and more objective evaluation. The purchase of one of these automated measurement systems could cost hundreds of millions of Hungarian forints [2].

Our goal is to design and build a vibration data acquisition system that can be used on road vehicles. The system should easily be mountable on any vehicle and measurements recorded at any speed. The system should be simple, but the measurement results must contain information about the quality of the road section [3].

2. Experimental

2.1 Tools of implementation

First, a design analysis was carried out and the structure of the system designed accordingly. A compact and light design is important, moreover, the housing should be sufficiently massive. Reliability and a constant sampling rate are also crucial. A block diagram of the system is shown in Fig. 1.

The system is based on an ATMega328P Microcontroller. The device is optimal for the control task as it supports several communication standards to connect the units to each other as well as possesses an inner A/D converter, PWM channels and several digital pins. The main microcontroller is responsible for managing communication between the individual units and handles the input data. A 16 MHz external oscillator IC is necessary to generate the clock signal and the supply voltage has to be stabilized at 5 V. As the microcontroller possesses a watchdog timer, a reset button is unnecessary. Along

*Correspondence: nagy.roland@mk.uni-pannon.hu

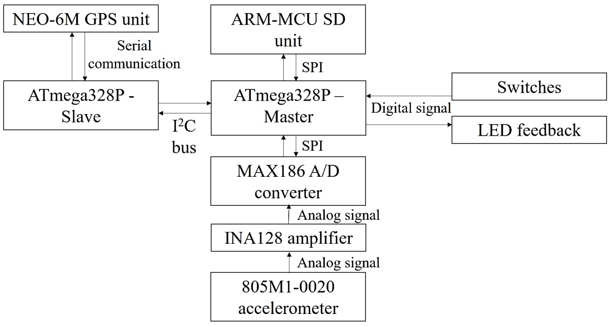


Figure 1: The structure of the system

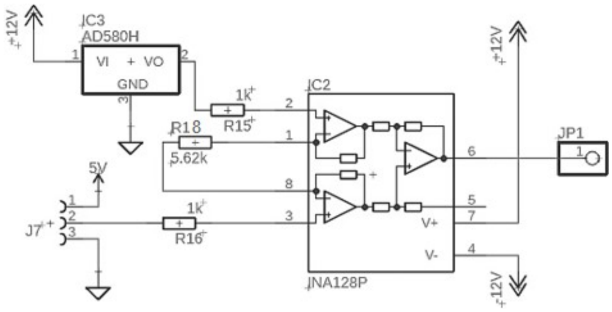


Figure 3: A wiring diagram of the amplifier section

with it, due to stable and robust operation, a pull-up resistor was placed in connection with the microcontroller's reset pin. In addition, an FTDI connector is placed near the microcontroller which enables it to be programmed in the programming language C++. The wiring diagram of the ATMega328P Microcontroller is shown in Fig. 2.

Vibrations caused by road defects were detected by a 805M1-0020 piezoelectric accelerometer. The sensitivity of the sensor is 100 mV/g with an offset voltage of 2.5 V. Previous studies suggest that accelerations of up to 2 g are expected, and because the analog-to-digital converter has a measuring range of $0 - 5$ V, a voltage gain of approximately 16 dB was necessary to ensure the analog signal fill the measuring range. For this purpose, an INA128P instrumentation amplifier was used, where the voltage gain was determined by the external resistor R_{G} between pins 1 and 8. The measuring amplifier amplified the potential difference between the two inputs. In this case, due to the 2.5 V offset voltage of the acceleration sensor, a voltage reference IC that provided 2.5 V was used. This voltage reference was connected to the inverter input of the amplifier, which caused the amplification to change linearly under different accelerations. An AD580 voltage reference IC was used. The output of the instrumentation amplifier is referred to as the output reference terminal. With 2.5 V connected to this pin, the offset voltage of the analog signal was in the middle of the measuring range. A wiring diagram of the amplifying section is shown in Fig. 3.

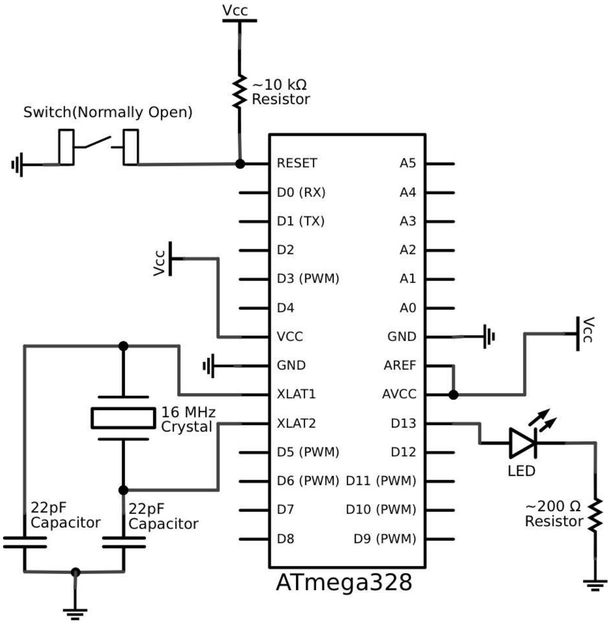


Figure 2: A wiring diagram of the ATMega328P Microcontroller

Following the amplifying section, the analog signal was digitized. Due to the increase in resolution, the voltage resolution was $5000/4096$, therefore, a change in voltage of 1.22 mV was detected instead of the previous 5 mV. The MAX186 is a successive-approximation converter with 8-channel single-ended or 4-channel differential inputs. The single-ended input mode was used, so the input signal voltages are referred to as AGND. It should be noted that the device worked in unipolar mode, therefore, an analog input signal of between 0 V and VREF could be converted. VREF was 5 V in our application [4].

Firstly, in terms of programming the MAX186 process, the parameters were set with the first 8 input bits. These bits were expected to shift synchronously with the clock signal. Once set up, the bits of digitized signal could be read out. The read-out process consisted of a 12-times repetitive 'For cycle', which reads out the digitized value bit by bit. It can be seen that the microcontroller obtains the value of the most significant bit (MSB) first. All the input bits have to shift left and must connect to each other with a Logical OR connection to yield the original measurement of the accelerometer. Eventually, the conversion process ends with the A/D converter chip select pin level changing to logic HIGH. A wiring diagram of the MAX186 is shown in Fig. 4.

In addition to the accelerations, the GPS coordinates were also collected to help identify road humps and potholes that were detected by the accelerometer. For this task, a NEO-6M GPS module was used that provides momentary velocity and altitude data in addition to longitudinal and latitudinal coordinates. To control this GPS module, another ATMega Microcontroller was installed which was necessary as a debugging process associated with the GPS communication was present. Its purpose was to verify the adequacy of the data and the connection with the satellites. As a result, the running time and sampling rate of the program cycle became quite slow.

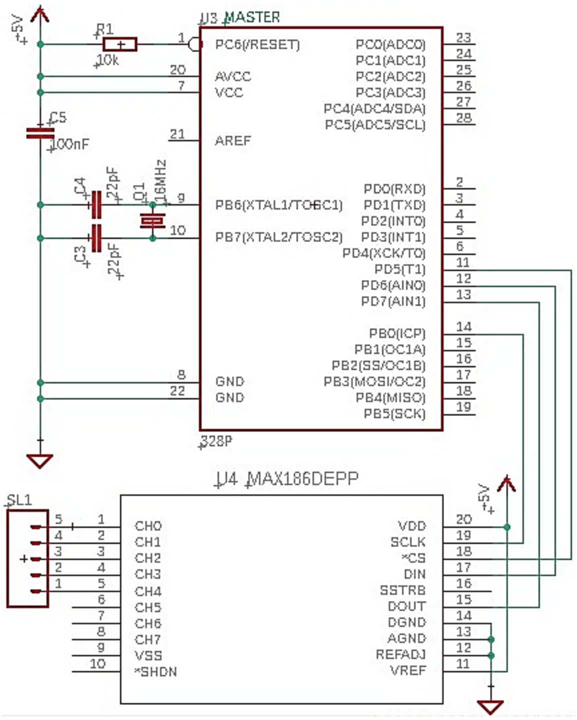


Figure 4: A wiring diagram of the Analog-to-digital converter

Following the installation of a second microcontroller, the program could be run synchronously, which a single ATMega328P Microcontroller would have been unable to do. As a result, the main microcontroller could be defined as the Master and the controller that controls the GPS as the Slave. The GPS unit communicated with the Slave controller by using a standard universal asynchronous receiver-transmitter (UART).

The two microcontrollers were connected by a BUS system, using the I^{2}C protocol which could only handle 8 bits of data in one cycle that corresponded to a number between 0 and 255 in the decimal system. As the GPS coordinates had to be transferred, which were 16 bits long in the binary system, these needed to be converted into 3 bit arrays that each contained 8 bits of data. The C++ program of the Slave controller handled these numbers in the binary system. First, it executed the Right Shift by a 16-bit command, which was followed by the bitwise AND command, masked by 0xff, the hexadecimal number FF. In this way, the first array was obtained, followed by the second array with a Right Shift using an 8-bit command and finally the last array by masking only. No conversion was needed for data concerning speed and elevation, because these values were only 8 bits long. The Master controller decoded the input arrays to the original 16-bit-long data. The data from the first input array had to Left Shift by 8 bits then perform a bitwise OR operation using the 8 bits of the following array to yield the original 16-bit-long data. Therefore, the GPS data from the Slave controller immediately became available, resulting in a much faster sampling frequency.

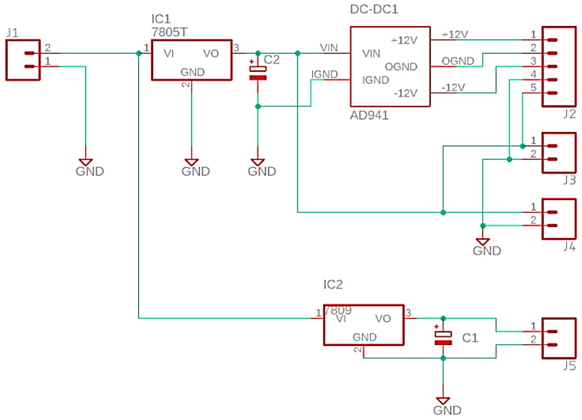


Figure 5: A wiring diagram of the power supply unit

Throughout the measurement process, the acceleration and GPS data were collected on an SD card using an ARM-MCU SD card module slot socket reader, which communicated with the Master controller by applying a standard SPI interface.

The Master controller's C++ program specified the sampling frequency of the accelerometer, i.e. 60 Hz, but is only suitable up to 30 Hz according to Shannon's law.

The power supply of the system was provided by its own integrated power supply. A maximum of 35 V could be connected to the power input connection of the power supply, thus, the system could be connected to the car's cigarette lighter. A NEE 78H05ASC voltage regulator converted 5 V from the input voltage. As ±12 V needed to be supplied to the instrumentation amplifier, an AD941 DC-DC converter was used to convert 5 V into ±12 V. Furthermore, 5 V was available from the output of the power supply unit to supply the microcontrollers, other units and sensors. A wiring diagram of the power supply unit is shown in Fig. 5.

2.2 Construction of the system

After designing the system and circuit diagrams, a printed circuit board was planned. Given the complexity of the panels, a double-sided PCB was designed. All of the integrated circuits (ICs) were used with the housing of the Dual Inline Package (DIP). The design rules were set for a CNC machine as the PCBs were handled using an engraving needle. The width of the wiring track was 0.9 mm at 5 V and 1.3 mm at 12 V. The clearance between the wiring tracks and solder pads was 1 mm and 0.4 mm, respectively. By taking into account insulation against electrical noise, larger copper surfaces were left on both sides of the PCB and placed on ground potential. The PCB was designed using Eagle CAD software.

After the design process, the CAM programs were generated from CAD files. PCB panels were made by an ISEL ICP 4030 3-axes CNC-milling machine. All parts of the system were placed in an IP-protection box. Two switches were placed on one side of the box, enabling the

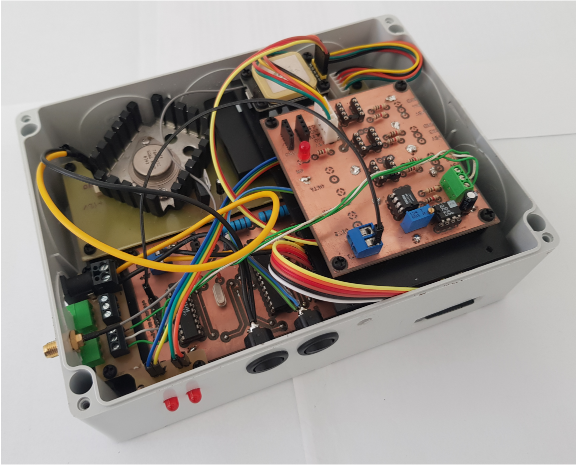


Figure 6: The completed data acquisition system

system to be turned on and the data acquisition process started. The complete system is shown in Fig. 6.

To check the output signal waveform in the relevant frequency bandwidth, a frequency response function of the amplifier unit was applied. The input signal was in the form of a sine wave within the frequency bandwidth of 10 Hz to 3 MHz with an amplitude of 40 mV. The data obtained is shown in Fig. 7. The natural frequency was approximately 700 kHz, so the frequency response function of the system performed well in the relevant frequency bandwidth.

Furthermore, amplification of the instrumentation amplifier was checked. The frequency of the input signal was 5 Hz with an amplitude of 400 mVpp and an offset voltage of 2.5 V. The amplification factor was set at 6.5 V/V. The data obtained is shown in Fig. 8. By plotting the data, the difference between the amplified and original signals can be seen. The waveform of the amplified signal was unchanged and by reading the amplitudes it can be seen that the voltage gain (G) was preset to 6.5 V/V.

3. Results and Analysis

Following inspections, the sensor was placed in a car and fixed to the rear control arm to avoid the presence of any damping object between the wheel and sensor [5]. After installing the data acquisition system, several measurements were carried out by vehicle-to-record data on different road surfaces. The results of the measurements are as follows:

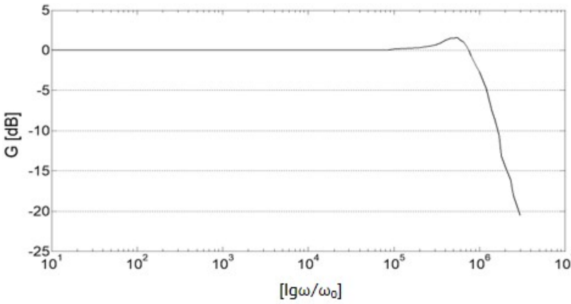


Figure 7: The frequency response function of the amplifier unit, where $\omega_0 = 1$ Hz

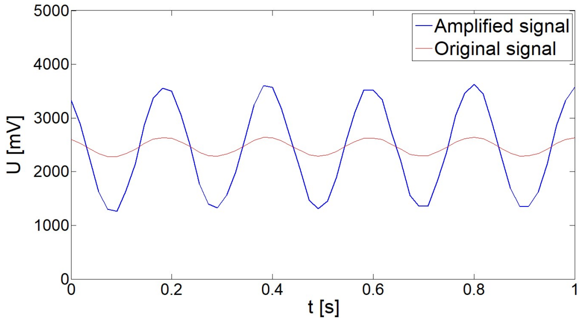


Figure 8: The amplified and original output signals at 5 Hz

The first measurement was taken in the vicinity of Zalaegerszeg on a 5.5 km stretch of road. The data was saved on an SD card as a TXT file and plotted. The GPS coordinates helped to analyze the data, through which the exact location of each acceleration value was determined. The voltage gain was set at 16 dB (6.5 V/V) [6] and the data from the accelerometer was visible on the diagram without amplification. From the first route, whilst driving in a populated area at the beginning of the measurement period, it can be concluded that the quality of the road surface was generally good with only a few significant potholes and cracks. Later, while leaving the town, the poorer quality of the road became immediately obvious given the increase in amplitudes. The road surface consisted of flange grooves and creases. According to the diagram, road defects continued to increase since the highest accelerations occurred towards the end of the measurement period. The acceleration data from the first measurement period is shown in Fig. 9.

The conditions of the second set of measurements were the same as those of the first. The initial section of this measurement period was identical to that of the first set of measurements (Fig. 10). The quality of the road surface was initially quite good as only accelerations of 1 g were recorded, but after leaving the populated area, the quality of the road surface also deteriorated. At the end of the route, several wearing courses, bonding layers and spotting errors were identified on the road surface.

The third stretch of road was 3 km long. Compared to the other measurements, the quality of the road surface here, which was covered in potholes, patches and surface sinks, was the worst. Consequently, in Fig. 11, almost all the accelerations reached the limit of 1 g, moreover, for a moment, accelerations of 2 g were also recorded. The right side of the diagram shows that when arriving in a populated area, the amplitudes reduced significantly as the quality of the road substantially improved. Only minor cracks were visible on the surface.

A small degree of asymmetry can be observed in the graphs, which is caused by the car's shock absorbers. As

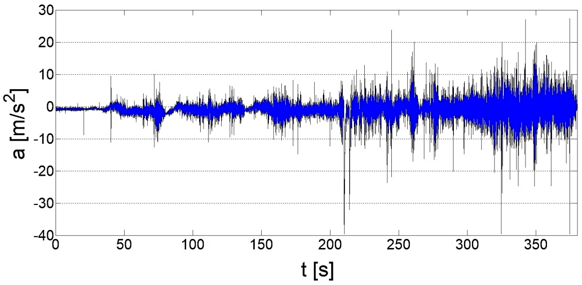


Figure 9: The acceleration of the first set of measurements plotted as a function of time

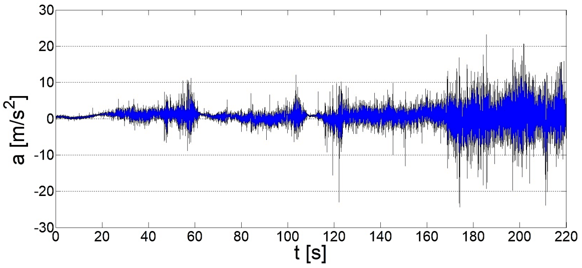


Figure 10: The acceleration of the second set of measurements plotted as a function of time

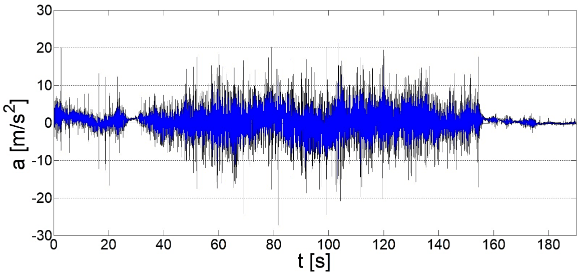


Figure 11: The acceleration of the third set of measurements plotted as a function of time

the damping force of the shock absorbers is less in the negative direction, higher accelerations might occur than in the positive direction. This is also supported by the characteristics of the shock absorbers.

4. Conclusion

The aim of our research was to design, build and program a vibration data acquisition system suitable for measuring the quality of a motor vehicle. This task involved numerous requirements: the measuring system had to function reliably, withstand the car's load and be easily installed, moreover, the data had to be produced and recorded as required.

In the light of the results, the system worked well; it detected road defects as expected. After measurements were taken, the data was illustrated in charts and the results were identical. The poorest road surface could be detected by the largest amplitudes and was identified according to the GPS coordinates when each measurement was made. The advantages of the system are that it can be easily mounted on any model of car and does not require driving at low speeds.

Acknowledgements

This research was supported by the European Union and co-financed by the European Social Fund under the project EFOP-3.6.2-16-2017-00002.

REFERENCES

[1] Souissi, H.; Hamaoui, A.: Effect of Human Exposure to Whole-Body Vibration in Transport, *Neuroergonomics*, 2018, 229 DOI: 10.1016/b978-0-12-811926-6.00049-x

[2] Szántó, M.: Közúti adatbázisok valós idejű frissítése közösségi megoldás alkalmazásával, *Útügyi Lapok,* 2017, **5**(1), 13-19 http://utugyilapok.hu/wp-content/uploads/2017/06/ul_5_c2_SZM.pdf

[3] Brown, J. C.; Robertson, A. J.; Serpento, S. T.: Motor Vehicle Structures (Butterworth-Heinemann) 2001, 4–84 ISBN: 978-0-7506-5134-9

[4] William, B.: Programmable Logic Controllers, (Elsevier Books*)* 2015 ISBN 01-28-0-2929-3

[5] Bidgoli, M. A.; Golroo, A.; Nadjar, H. S.; Rashidabad, A. G.; Ganji, M. R.: Road roughness measurement using a cost-effective sensor-based monitoring system, *Automat. Constr.*, 2019, **104**, 140–152 DOI: 10.1016/j.autcon.2019.04.007

[6] Kennedy, J.; Oakley, C.; Sumon, S.; Parry, I.; Wilkinson, E.; Brown, J.: Impact of road humps on vehicles and their occupants, *TRL Report*, 2004, **614**, 3–16 https://trl.co.uk/sites/default/files/TRL614%281%29.pdf

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 77–80 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-12

TORQUE TRANSMISSION TIME-CONSTANT EXAMINATION OF A DISK-TYPE MAGNETORHEOLOGICAL CLUTCH

SÁNDOR MESTER*[1] AND ISTVÁN SZALAI[1]

[1]Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk utca 18/A, Zalaegerszeg, 8900, HUNGARY

In this paper, the torque transmission time-constant of a simple disk-type magnetorheological (MR) clutch is investigated. By using MR fluid, controlled torque transmission can be easily implemented, facilitating the widespread use of similar systems. In order to describe the dynamical properties of the system, time constants were measured at different speeds and magnetic inductions. The time constants were derived by fitting an exponential function to the data.

Keywords: magnetorheological fluid, clutch, torque transmission

1. Introduction

The implementation of intelligent materials has significantly increased not only in everyday life, but in the industrial environment as well. Magnetorheological fluids are suspensions that consist of at least two phases, whose apparent viscosity changes quickly (within $20 - 30$ ms) in the presence of an external magnetic field. The suspension usually consists of small ferromagnetic particles of approximately 10 μm in diameter dispersed in oil. Usually surface active agents are also present in the fluid to prevent the particles from being attracted to each other. In the presence of an external magnetic field, the particles gain an induced dipole moment and form pairs, followed by chain-like structures. If the fluid is exposed to shearing, these chain-like structures oppose each other leading to an increase in apparent viscosity which can be used in shock absorbers, torque transmission, superfinishing of optical lenses and civil engineering [1–4]. In addition to the rapid change in apparent viscosity, its low power requirement makes MR fluid ideal for use in semiactive shock absorbers and variable torque-transmission clutches.

The most basic forms of magnetorheological clutches are the disk and cylindrical types. The differences between the two types are the transmittable torque and the volume of the idling waste. A stream of particles is present in disk-type clutches due to the centrifugal force. In the cylindrical type, the effect is less significant [5, 6]. This study focuses on a disk-type clutch with a well-defined mathematical model [7].

2. Experimental

The model of our magnetorheological clutch is shown in Fig. 1. The clutch consisted of the following main components: 1. DC servomotor with a gearbox; 2. toothed belt drive with $1 : 1$ gear ratio; 3. input or drive shaft; 4. input shaft bearings; 5. lower disk; 6. upper disk; 7. upper or driven shaft; 8. upper shaft bearings; 9. torque sensor; and 10. toothed belt drive with $4 : 1$ gear ratio. The electromagnetic coil, which consisted of approximately 1000 turns, was located around the disks with the coils encircling them, providing magnetic field lines parallel to the shafts. The MR fluid was loaded into the lower disk. The diameter of the disks was 116 mm and the gap between them was 1 mm.

The input and output signals were read and recorded by an National Instruments (NI) USB-6281 data acquisition device with a sampling rate of 1 kHz.

The complete magnetorheological clutch is shown in Fig. 2. For future measurements, a 3-phase AC induction motor with a frequency inverter was installed to serve as an artificial load.

2.1 Samples and Measurements

A Hall effect sensor, located at the top of the upper disk, measured the magnetic field. An encoder was located at the bottom of the drive shaft of the clutch to measure the input speed. The control and measurement tasks were carried out by a NI USB-6281 data acquisition device controlled by LabVIEW software. The input speed and output torque were recorded. The MR fluid used was LORD MRF-122EG fluid diluted by 50 m/m% with silicone oil.

*Correspondence: hollosi.janos@sze.hu

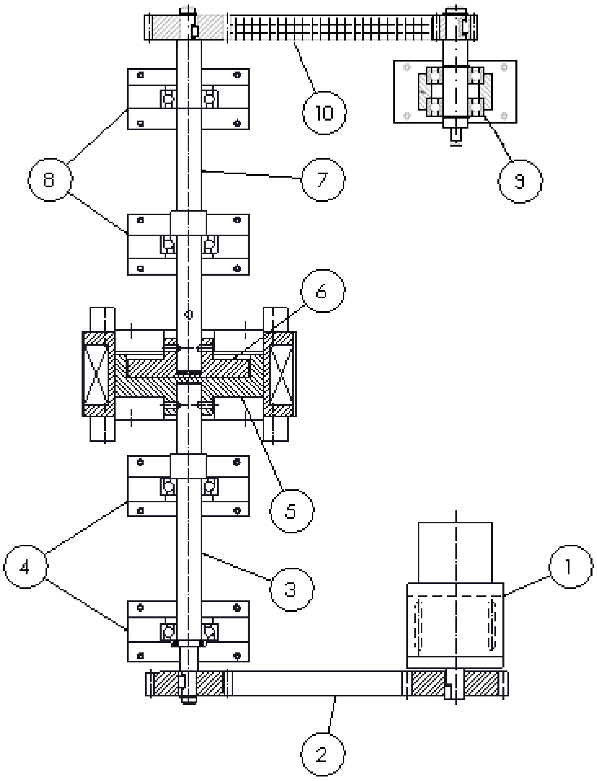


Figure 1: Model of a disk-type magnetorheological clutch

The output side of the clutch was fixed at the torque sensor with the intention of measuring the torque transmission time-constants, so the output speeds at both the sensor and driven shaft were zero. The measurement cycles, using a preset magnetic field, consisted of a 3-second-long mixing phase in the absence of a magnetic field, followed by a 6-second-long measurement period with the magnetic field turned on, and finally another 3-second-long mixing phase, all at a constant speed. After the cycle, a new speed was set and a new cycle started. The speeds measured were 11.5, 17.5, 23.5, 29.5, 35.5, 40.5, and 46.5 rpm. The magnetic induction was controlled by currents of 1.8, 2.5, 3.3, and 4.4 A applied to the coil, thereafter the corresponding magnetic inductions were 114, 144, 173, and 202 mT.

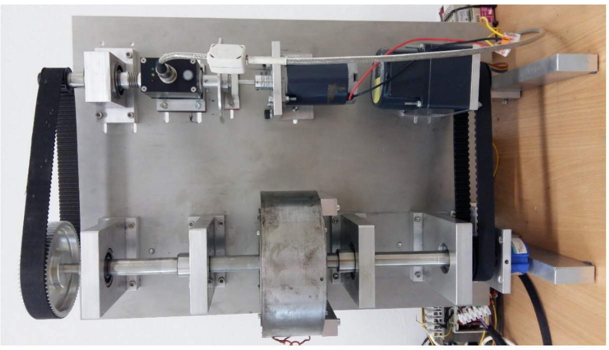


Figure 2: Complete magnetorheological clutch

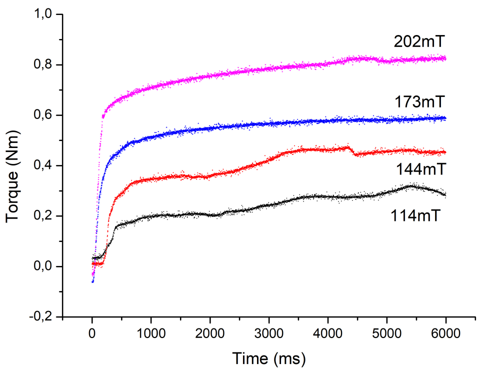


Figure 3: Transmission torque from when the magnetic field was turned on at a speed of $n = 11.38$ rpm and using different magnetic inductions

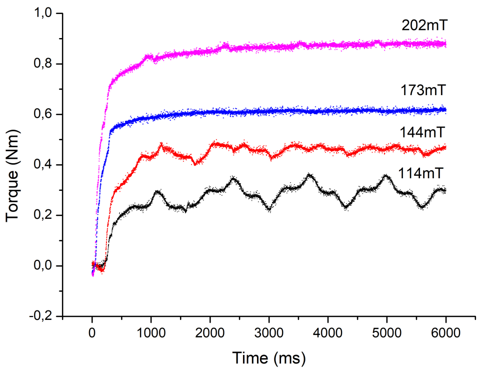


Figure 4: Transmission torque from when the magnetic field was turned on at a speed of $n = 46.5$ rpm and using different magnetic inductions

3. Results and Discussion

3.1 Experiments

Using the aforementioned method, a series of 12 measurements were carried out at each speed and in each magnetic field by taking into consideration the uncertainty of the mechanical parts (toothed belts, bearings, etc.).

Examples of the measurements can be seen in Fig. 3 ($n = 11.38$ rpm) and Fig. 4 ($n = 46.50$ rpm), as well as a periodicity due to the inaccuracies of the mechanical parts. The maximum output torque was $M = 0.85$ Nm.

3.2 Calculation of time constants

Time constants were determined by fitting an exponential function to the measured data:

$$M(t) = A\left(1 - e^{-\frac{t}{\tau}}\right) \tag{1}$$

where M denotes the transmission torque, t represents the time, and τ stands for the time constant. The

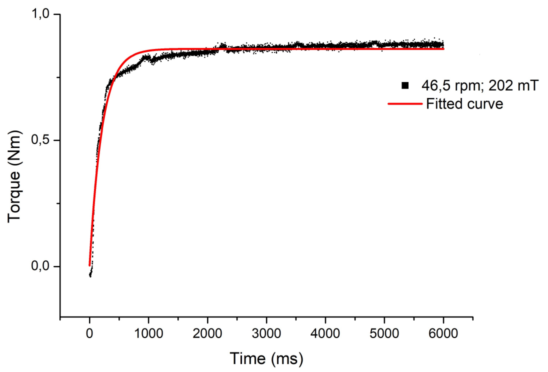


Figure 5: CFitted curve of measurements at a speed of $n = 46.5$ rpm and a magnetic induction of $B = 202$ mT

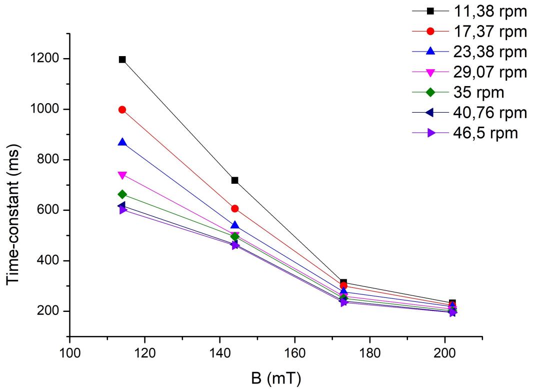


Figure 6: Time constants as a function of magnetic induction at different speeds

fitting was carried out according to the measurements. An example of the fitting is shown in Fig. 5.

The fitted curve correlates well with the measurement data. The results of the fitting were averaged to obtain the time constants for the applied magnetic inductions and rotational speeds. The time constants are shown in Fig. 6 as a function of magnetic induction. The measured time constants took into account the whole system, including the effects of the MR fluid as well as mechanical parts.

By increasing the magnetic field strength, the time constants decreased. In the weakest magnetic field and at the lowest speed, the time constant was extremely high, greater than 1 second to be exact, due to the characteristics of the fluid.

The time constants as a function of the speeds are shown in Fig. 7. Increasing the speed also resulted in a reduction in the time constant that was less significant than in stronger magnetic fields. In strong magnetic fields, the influence of the speed on the time constants was insignificant, therefore, its effect was almost negligible.

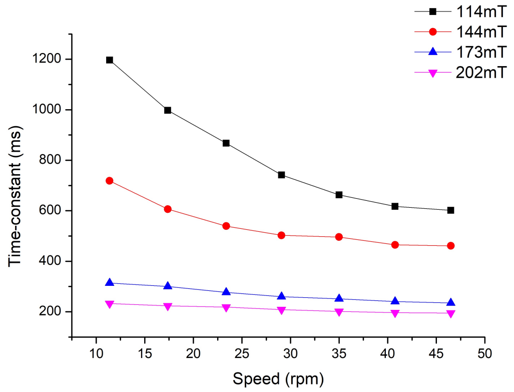


Figure 7: Time constants as a function of speed using different magnetic inductions

4. Conclusion

Based on the measurement results, the torque transmission time-constants for our magnetorheological clutch using the MR fluid were obtained. The time constants differed greatly from the response times of a magnetorheological fluid due to the inertia of the mechanical components. In the future, the influence of the mechanical parts on the time constants will be measured.

Acknowledgements

This research was supported by the European Union and co-financed by the European Social Fund under the project EFOP-3.6.2-16-2017-00002.

REFERENCES

[1] Rankin, P. J.; Ginder, J. M.; Klingenberg, D. J.: Electro- and magneto-rheology. *Current Opinion in Colloid & Interface Science*, 1998, **3**(4), 373–381 DOI: 10.1016/s1359-0294(98)80052-6

[2] Olabi, A. G. ; Grunwald, A.: Design and application of magneto-rheological fluid, *Mater. Des.*, 2007, **28**(10), 2658–2664 DOI: 10.1016/j.matdes.2006.10.009

[3] Mazlan, S. A. ; Ekreem, N. B.; Olabi, A. G.: An investigation of the behaviour of magnetorheological fluids in compression mode, *J. Mater. Process. Technol.*, 2008, **201**(1–3), 780–785 DOI: 10.1016/j.jmatprotec.2007.11.257

[4] Lampert, S. G. E. ; van Ostayen, R. A. J.: Experimental results on a hydrostatic bearing lubricated with a magnetorheological fluid, *Curr. Appl. Phys.*, 2019, **19**(12), 1441–1448 DOI: 10.1016/j.cap.2019.09.004

[5] Törőcsik, D.: Some Design Issues of Multi-Plate Magnetorheological Clutches, *Hungarian J. Ind. Chem.*, 2011, **39**(1), 41–44 https://mk.uni-pannon.hu/hjic/index.php/hjic/article/view/380

[6] Rabinow, J.: The Magnetic Fluid Clutch, *Trans. Am. Inst. Electr. Eng.*, 1948, **67**(2), 1308–1315 DOI: 10.1109/t-aiee.1948.5059821

[7] Decsi, P.; Mester, S.; Szalai, I.: Tárcsás magnetoreológiai tengelykapcsoló modellezése, a rendszer időbeli viselkedésének vizsgálata, OGÉT XXVI. Nemzetközi Gépészei Konferencia 2018, 91–94

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 81–85 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-13

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
HJIC

OPTIMIZING THE PLANNING AND MANUFACTURING PROCESSES OF ELECTROMAGNETIC ENERGY HARVESTING EQUIPMENT

LÁSZLÓ MÓRICZ*[1] AND ISTVÁN SZALAI[1]

[1]Institute of Mechatronics Engineering and Research, Faculty of Engineering, University of Pannonia,
Gasparich Márk u. 18/A, Zalaegerszeg, 8900, HUNGARY

The main aim of this paper is to create an energy harvesting system, which can convert vibrational energy into electrical energy efficiently. Our research was carried out in the field of electromagnetic energy conversion using the principles of linear generator construction for both low and high frequency vibrations. Energy can be recovered efficiently. During the measurements, how the induced voltage is dependent on the impulsive frequency and the amplitude of impulses was investigated.

Keywords: energy harvesting, induced voltage, vibration, linear generator, energy

1. Introduction

Many forms of energy sources exist (vibrational, thermal, wind) in the environment which can be converted into electrical energy with a good degree of efficiency. The harvesting of this energy from the environment has the potential to reduce the rate of depletion of non-renewable energy sources [1] and can be converted by using electromagnetic [2,3], electrostatic [4,5] and piezoelectric [6,7] energy conversion processes.

Our research was conducted in the field of electromagnetic energy conversion for both low and high frequency vibrations. Numerous energy harvesting mechanisms are based on the damped driven harmonic oscillator (DDHO) [8]. The essence of the process is to create relative displacement between a permanent magnet and a coil [9]. Electric power (energy) is induced in the coil due to changes in magnetic flux. To achieve the relative displacement, the magnet and leading house must come into physical contact which can be achieved mechanically or magnetically [10].

Each mechanical system has a mechanical damping factor. If the damping of the system is too low, the device exhibits no resistance to harmonic motion. However, if the value becomes too high, the resistance of the device to motion increases dramatically, thus no relative displacement of the device occurs. Both the damping force and relative displacement are essential to convert energy efficiently into the system [11].

One of the most difficult tasks of the design process is to define the appropriate degree of damping that maximizes the extractable efficiency. An important aspect of the design process is the tuning of the natural frequency of the structure. If the impulsive frequency deviates from the resonant frequency, a loss of power can be detected.

One possibility is that the bandwidth of operation is enhanced which results in the value of the "Quality (Q) factor" decreasing and diminishes the amount of extractable energy [5]. To achieve a good degree of efficiency of the system, the harvesting of very low frequency vibrations must be taken into account.

Regarding energy harvesting systems for low frequency applications, the possibilities of frequency upconversion are introduced and achieved in different ways. Ashraf et al. [11] optimized the mechanical design of the system by applying the Finite Element Method to broaden the low frequency range.

Haroun et al. [9] tried to keep the natural frequency of their system, namely CEH, low. They concluded that if the spring is not fixed to the moving frame (FIEH), then the natural frequency of the system is lower than that of the fixed spring system (CEH).

2. Design process and evolution of the structure

There are two types of generator-based energy harvesting systems:

- System 1: based on linear movement
- System 2: based on rotational movement

The linear generator converts the mechanical movement directly into electrical energy. Several basic construction solutions can achieve this, e.g. the linear motors can

*Correspondence: moricz.laszlo@mk.uni-pannon.hu

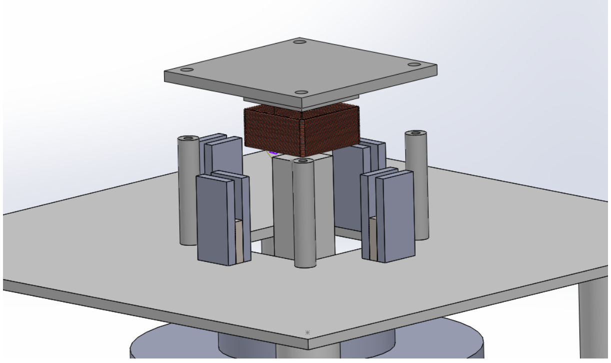


Figure 1: Mechanical structure of the EH system

be straightened versions of permanent magnet motors. The structure chosen is presented in Fig. 1. The energy harvesting model was made using SOLIDWORKS 2016 software. The assembled system is shown in Fig. 2.

The structure consists of two main parts; the stationary part possesses a coil holder and the moving part was produced from a square section slip. Four horseshoe neodymium magnets were mounted on the moving part. The horseshoe magnets consisted of two iron plates and a square neodymium magnet.

The thickness of the two iron plates was equal to that of the square neodymium magnet. It is important that the iron plate contains less alloys. The best solution from the options available was to use an iron core of a transformer.

To determine the optimum layout of the magnets, the direction of the current vectors ($\mathbf{E}$) must be identical. As the direction of movement of the structure was definite ($\mathbf{v}$), according to the right-hand rule the direction of the magnetic induction vectors ($\mathbf{B}$) must point to the center as shown in Fig. 3.

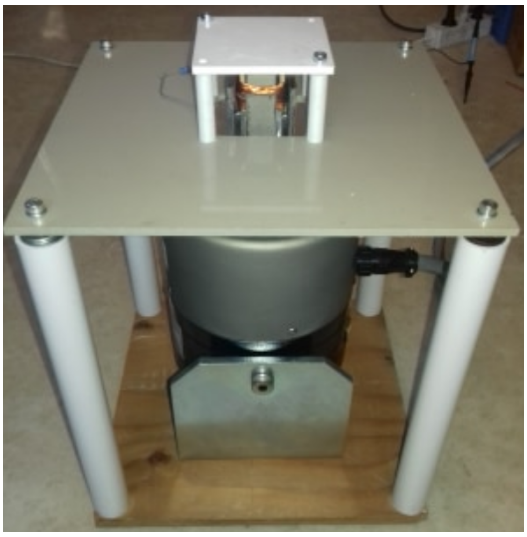


Figure 2: The assembled system

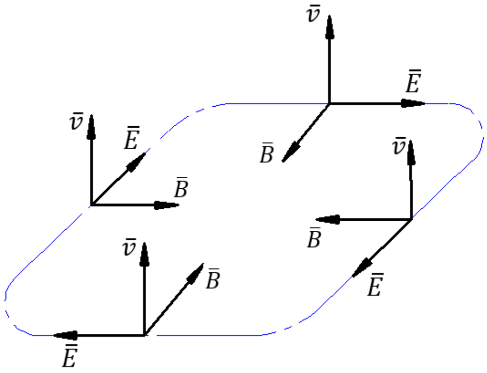


Figure 3: Optimum layout of the magnets

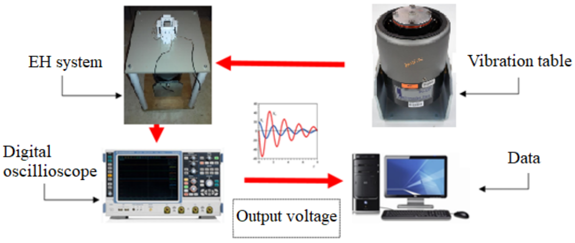


Figure 4: Schematic structure of the loop test

3. Structure of the loop test

The equipment for the loop test was provided by the Institute of Mechatronics Engineering and Research of the University of Pannonia in Zalaegerszeg. The schematic structure of the loop test is shown in Fig. 4.

Energy harvesting was executed by a type of Labworks ET-139 electrodynamic shaker. The induced voltage was displayed by an Agilent DSO5054A digital oscilloscope. The examined parameters were changed by a function generator, which was connected to a Labworks PA-138 amplifier on a vibration table as illustrated in Fig. 5.

4. Results and Analysis

4.1 Based on experiments

Throughout the experiments, the following attributes were examined:

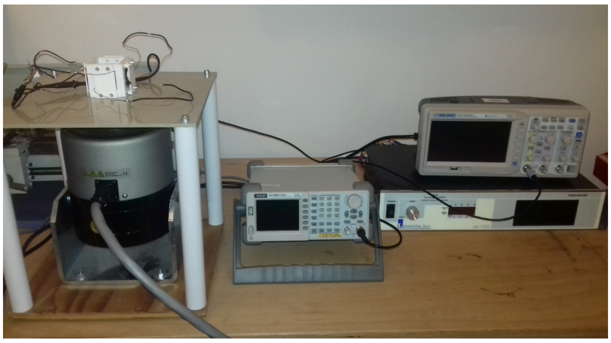


Figure 5: The set-up of the loop test

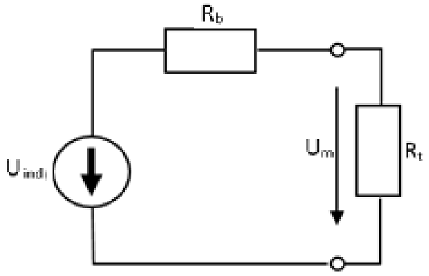


Figure 6: Energy-harvesting circuit diagram

- Maximum induced voltage without load
- Power without load
- Load on the power
- The impact of the number of coils on the induced voltage and power

The examined energy-harvesting circuit diagram is shown in Fig. 6. The structure consists of an internal resistance R_b and an external resistance R_t (load).

$$P_\mathrm{total} = \frac{U_\mathrm{ind}^2}{R_\mathrm{total}} = \frac{U_\mathrm{ind}^2}{R_\mathrm{b} + R_\mathrm{t}} \tag{1}$$

$$U_\mathrm{m} = U_\mathrm{ind} \frac{R_\mathrm{t}}{R_\mathrm{t} + R_\mathrm{b}} \tag{2}$$

As a result of the impulsive frequency and amplitude of impulses, electrical energy was induced. The induced voltage was equal to the measured voltage in the absence of external resistance. Measured and induced voltages differed when the system was subjected to an external resistance. The relationship between them is described in Eq. 2. The maximum power can be determined from Eq. 1.

4.2 Results

Initially, the device was tested with 100 turns of the coil. The internal resistance of the coil was 3.1 Ω. The impulsive frequency was set between 1 and 20 Hz and the amplitude of impulses between 2.5 and 15 mm. During the experiment, a decrease in the induced voltage was observed above 20 Hz. Thus, the investigated bandwidth was maximized at 20 Hz, whereas the trend was still visible in terms of the change in the curves, so 20 measurement points were examined during the experiments.

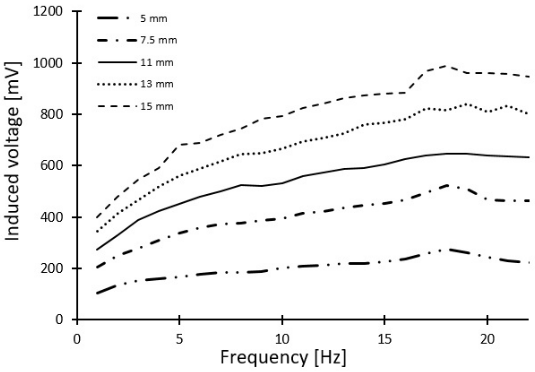


Figure 7: Induced voltage by applying 100 turns of the coil in the absence of external resistance

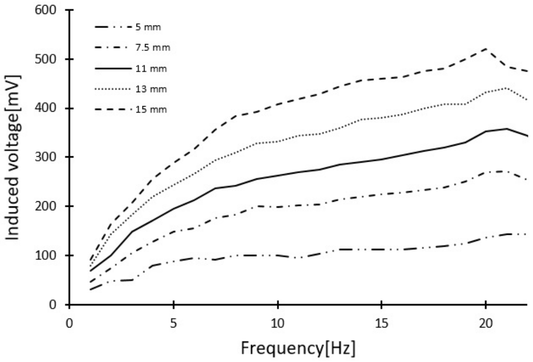


Figure 8: Induced voltage by applying 100 turns of the coil in the presence of an external resistance

The results are summarized in Fig. 7. The maximum induced voltage and power were 986 mV and 322 mW, respectively.

During the experiments below, an internal resistance equal to the external resistance was applied to the structure. The applied external resistance was 3.4 Ω. The results are summarized in Fig. 8.

The maximum voltage measured was 520 mV. Given the values of the external and internal resistances, the induced voltage was 994 mV based on Eq. 2. The maximum power was calculated to be 152 mW from Eq. 1.

The impact of the external resistance on the power During the experiment, a constant excitation amplitude of 15 mm was applied, while the impact of the resistance on the power was examined. The resistances applied were 1, 3.4, 10, 22, 47 and 74 Ω. The relationship between the changes in resistance and power are summarized in Fig. 9.

As is shown in Fig. 9, an exponential decrease in power was observed as the resistance increased. Based on previous studies, an external resistance that is smaller

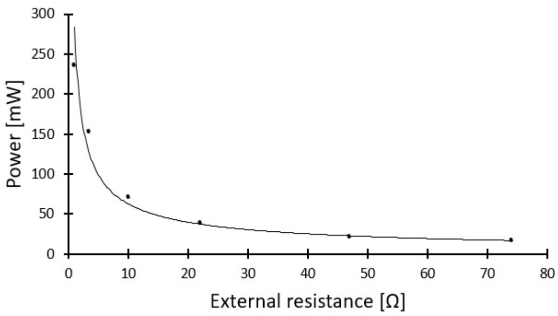


Figure 9: The relationship between the resistance and power

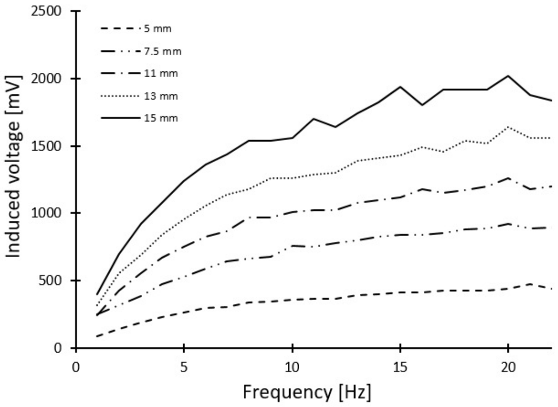


Figure 10: Induced voltage by applying 240 turns of the coil in the absence of an external resistance

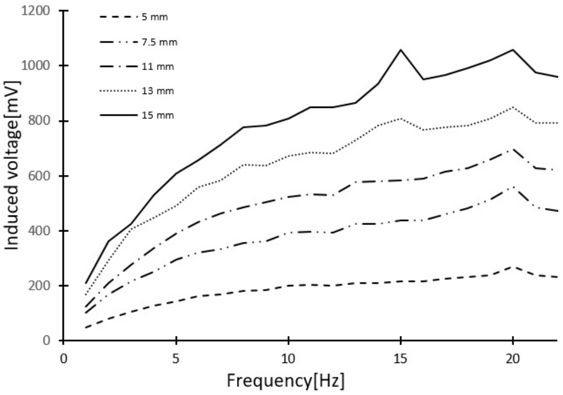


Figure 11: Induced voltage by applying 240 turns of the coil in the presence of an external resistance

than the internal resistance is impractical. Ideally, the external resistance would be equal to the internal resistance of the coil.

Next, the number of turns of the coil was increased from 100 to 240. The other aforementioned variables remained unchanged. The results are summarized in Fig. 10.

As shown in Fig. 11, the maximum induced voltage without a load and the maximum power were 2020 mV and 559 mW, respectively. Following the aforementioned procedures, the loaded system was analyzed.

The external resistance applied was 8 Ω. The maximum voltage measured was 1060 mV. By taking into account the values of the external and internal resistances, the induced voltage was 2020 mV based on Eq. 2. Based on Eq. 1, the maximum power calculated was 268 mW. Both the induced voltage and power of the system were doubled by increasing the number of turns of the coil by 60 %, the induced voltage increased from 994 mV to 2020 mV and the maximum power rose from 152 mW to 268 mW to be exact.

5. Discussion

The aim of the research was based on the principles of linear generator construction and manufacturing. At this stage of the process, it was important that the structure was free of mechanical damping. During the experiment, the structure was examined by means of changing the load resistance and number of turns of the coil in addition to the specified amplitude and frequency. An exponential decrease in the efficiency was observed as the resistance increased. Ideally, the external resistance would be equal to the internal resistance of the coil. The induced voltage and the power of the system were doubled by increasing the number of turns of the coil by 60 %. As a result, by increasing the number of turns of the coil by 60 %, the efficiency of the system also increased by approximately 57 %. However, a deeper understanding of the relationship between the efficiency of the structure and variables requires further investigation. After doubling the number of turns of the coil, the maximum power generated was 1 W. One advantage of this system in particular is that the neodymium magnets are cheap to produce. Applying a series connection to this system results in a sufficient degree of efficiency to operate the electronic devices in cars.

6. Conclusion

In the aforementioned experiments, the maximum induced voltage and power achieved by applying 240 turns of the coil were 2020 mV and 559 mW, respectively. During the experiments in the presence of a load resistance, the best value of the power was calculated when the external resistance was equal to the internal resistance of the coil. The efficiency of this energy harvesting system can be further enhanced by increasing the number of turns of the coil and the strength of the neodymium magnet.

Symbols

U_{ind}	induced voltage
U_{m}	measured voltage
P_{total}	power
R_{b}	internal resistance
R_{t}	external resistance

Acknowledgements

The project was supported by the European Union and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00002.

REFERENCES

[1] Elmes, J.; Gaydarzhiev, V.; Mensah, A.; Rustom, K.; Shen, J.; Batarseh, I.: Maximum Energy Harvesting Control for Oscillating Energy Harvesting Systems, 2007 IEEE Power Electronics Specialists Conference, 2007 DOI: 10.1109/pesc.2007.4342461

[2] von Büren, T.; Tröster, G.: Design and optimization of a linear vibration-driven electromagnetic micro-power generator, *Sensor Actuat. A-Phys.*, 2007, **135**(2), 765–775 DOI: 10.1016/j.sna.2006.08.009

[3] Beeby, S.P.; Torah, R.N.; Tudor, M.J.; Glynne-Jones, P.; O'Donnell, T.; Saha, C.R.; Roy, S.: Micro electromagnetic generator for vibration energy harvesting, *J. Micromech. Microeng.*, 2007, **117**(7), 1257–1265 DOI: 10.1088/0960-1317/17/7/007

[4] Mitcheson, P.D.; Green, T.C.: Maximum effectiveness of electrostatic energy harvesters when coupled to interface circuits, *IEEE T. Circuits-I*, 2012,**59**(12), 3098–3111 DOI: 10.1109/tcsi.2012.2206432

[5] Kiziroglou, M.E.; He, C.; Yeatman, E.M.: Electrostatic energy harvester with external proof mass, *Proceedings of PowerMEMS*, 2007, 117–120

[6] Marzencki, M.; Basrour, S.; Charlot, B.; Spirkovich, S.; Clin, M.: AMEMS piezoelectric vibration energy harvesting device, *Proceedings of PowerMEMS*, 2005, 45–48

[7] Isarakorn, D.; Briand, D.; Janphuang, P.; Sambri, A.; Gariglio, S.; Tricone, J. M.; Guy, F.; Reiner, J. W.; Ahn, C.H.; de Rooij, N. F.: Energy harvesting MEMS device based on an epitaxial PZT thin film: fabrication and characterization, *Technical Digest of PowerMEMS*, 2010, 203–206

[8] Niu, P.; Chapman, P.: Design and Performance of Linear Biomechanical Energy Conversion Devices, (Power Electronics Specialists Conference, 2006. PESC '06. 37th IEEE), 2006, 1–6 DOI: 10.1109/PESC.2006.1711996

[9] Haroun, A.; Yamada,I.; Warisawa, S.: Study of electromagnetic vibration energy harvesting with free/impact motion for low frequency operation, *J. Sound Vib.*, 2015, **349**, 389–402 DOI: 10.1016/j.jsv.2015.03.048

[10] Móricz, L.; Szalai, I.: Mágneses lebegtetés elvén működő vibrációs energiaátalakító tervezése és építése, (OGÉT 2019 XXVII. Nemzetközi Gépészeti Konferencia, Nagyvárad, Románia), 2019, 352–355

[11] Ashraf, K.; Md Khir, M.H.; Dennis, J.O.; Baharudin, Z.: Improved energy harvesting from low frequency vibrations by resonance amplification at multiple frequencies, *Sensor Actuat. A-Phys.*, 2013, **195**, 123–132 DOI: 10.1016/j.sna.2013.03.026

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 87–93 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-14

EXAMINATION OF FUEL CONSUMPTION FACTORS, BASICS OF PRECISION AND ON-BOARD DIAGNOSTIC MEASUREMENTS

TIBOR BUSZNYÁK*[1] AND ISTVÁN LAKATOS[1]

[1]Department of Road and Rail Vehicles, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY

In this paper, different factors of fuel consumption are examined. Driveload equitation is used as a basis and the parts that handle energy consumption in particular are analyzed. For the purposes of visibility, it was implemented using MATLAB. In statistical works, fuel consumption data require that the energy consumption of vehicles be analyzed correctly. Variables which affect fuel consumption during a given drive are defined. Research is analyzed in the second part of the paper where vehicle diagnostics are combined with global positioning. Examinations are necessary to create on-board diagnostics-based positioning.

Keywords: GPS, OBD, correlation, drive, assistance

1. Introduction

Nowadays, innovation is a key. Economical, safety-centred or traffic optimization tasks are increasingly regulated. These criteria require developers to actuate and consequently upgrade their conceptions. New technologies are rapidly emerging so industries have to keep up to date. Drive options, including alternative drive solutions, are continuously being updated, the number of driver-assistance features is ever-increasing towards a possible fully autonomous level [1].

The role of development focusing on Smart City concepts and sustainable traffic is becoming more important. Critical aspects of it are efficient energy use (the central question of the present paper), range of online communication systems, autonomous transport systems and conceptions of autonomous vehicles. Reliable operation requires cooperation between different participants, e.g. the information technology, urban development and automotive industries. These aspects are interrelated, therefore, a more efficient Intelligent Transportation System (ITS) could be realized [2–4].

Information technologies between different units of traffic are elementary in terms of automated traffic. The stability of dataflow is unavoidable. Communication channels play a key role in everyday life as information is accessed from the Internet.

As information content defines the quality of data, the demands of traffic quality have recently been increasing.

The number of automobiles in Hungary has almost doubled over the past twenty years. Safety issues and accidents are increasingly commonplace. Besides accidents, traffic jams have also become more frequent.

As a result, driving has become harder. Rush-hour traffic that slowly inches forward, searching for a parking space or simply parking itself put drivers to the test under crowded, metropolitan conditions. The need to avoid similar situations has led to the emergence of driver-assistance systems.

The quality of data transmissions as well as trouble logger- and indicator systems, which evaluate inputs from sensors or on-board diagnostics, are closely connected to vehicle information. The aforementioned technologies help driver-assistance systems to function. Due to information technology and automatization, it is possible to create a vehicle network. One of these networks is the vehicle-to-everything (V2X) communication platform where vehicles communicate with each other along with the infrastructure provider to share information about the locations of traffic jams and avoid congestion. Vehicle communication and driver-assistance systems help to improve road traffic safety and make more accurate predictions [5–7]. An important task is to define databases based on the optimization of traffic.

Several methods, e.g. based on vehicles or infrastructure, are available in order to build a database.

If the vehicle investigated predominantly drives in well-maintained, intelligent infrastructure, then the number and complexity of built-in vehicle systems can be reduced.

In this case, information is supplied to the vehicle by an uninterrupted connection with external systems. This could also be true of the drive of a vehicle on predefined routes, e.g. buses. It is easier to build infrastructure for

*Correspondence: busznyaktibor@gmail.com

public transport vehicles because their routes are predefined. On the other hand, a vehicle can be defined as a separate unit. Without infrastructure, vehicles rely on built-in sensors and can drive anywhere, external infrastructure is unnecessary.

How could the complexity of a given vehicle's sensor system be reduced? Would it be possible to use built-in on-board diagnostics for positioning tasks.

Basic ideas originate from simple experiences. If people drive uphill in cruise control, the amount of data concerning fuel consumption that appears on the dashboard increases. The core of this research is the possible connection between elevation and fuel consumption:

1. Can a connection between elevation data from global positioning and fuel consumption data from on-board diagnostics at a constant or various speeds be identified?
2. Is it possible to create a topographic elevation model from fuel consumption data?
3. If it is possible, then the fuel consumption can be predicted from road conditions.
4. By integrating on-board diagnostics into conventional or intelligent transportation systems using the presented relations, a vehicle can be located.

Connections between data from global positioning systems and fuel consumption are sought. It is necessary to define important variables that affect the fuel consumption of a vehicle. The relevant equations and propulsion power requirements are analyzed.

2. Experiment

2.1 Propulsion power requirements and fuel consumption – defining variables

Internal combustion engines function by burning fuel which is blended with air in line with energy requirements. Propulsion power is necessary for a vehicle to move but its movement is restricted by various internal and external driving resistances.

External driving resistances

Rolling resistance is

$$F_\mathrm{g} = \mu m g \tag{1}$$

The rolling force resists motion when tires are rotating on a given surface. Internal and external factors are included in the equation.

The external factor is the rolling resistance coefficient which depends on contacting surfaces. The internal factor is the deformation of the tires which is dependent on the load of the vehicle. A loss in power results. Power against rolling resistance is

$$P_\mathrm{g} = F_\mathrm{g} v \tag{2}$$

Aerodynamic drag is

$$F_\mathrm{l} = c_\mathrm{w} \rho A v^2/2 \tag{3}$$

Drag acts in the opposite direction to which the vehicle is moving. It plays a major role in terms of vehicle dynamics and efficiency.

At higher speeds, it is more significant because drag increases with the square of the velocity. Power against drag is

$$P_\mathrm{l} = F_\mathrm{l} v \tag{4}$$

Climbing resistance is

$$F_\mathrm{e} = m g \sin(\alpha) \tag{5}$$

Climbing resistance depends on the elevation of the route, mass of the vehicle and road gradient. Power against climbing resistance is

$$P_\mathrm{e} = F_\mathrm{e} v \tag{6}$$

Internal driving resistances

Acceleration resistance is

$$F_\mathrm{gy} = (1+\theta) m a, \tag{7}$$

where θ is a coefficient of rotating components (Table 1). Energy is required to accelerate. The acceleration resistance can be calculated from the masses of the rotating components and vehicle. Power against acceleration resistance is

$$P_\mathrm{gy} = F_\mathrm{gy} v \tag{8}$$

Other internal resistances, e.g. transmission resistance, are

$$P_\mathrm{eff} = (1-\eta) P_\mathrm{h} \tag{9}$$

Another internal resistance arises when the transmission system moves and depends on the efficiency of its parts, moreover, it is used to calculate power.

This internal resistance is constant and includes the efficiency of the differential (0.93), efficiency of the clutch (0.99), efficiency of the drive shaft (0.99), efficiency of the gearbox (0.97) and efficiency of the bearings (0.98):

$$\eta = \eta_\mathrm{tk}\, \eta_\mathrm{diff}\, \eta_\mathrm{kt}\, \eta_\mathrm{cs}\, \eta_\mathrm{ny} \tag{10}$$

Finally, energy produced by the combustion of fuel is translated into the energy requirements of given resistances. At constant velocities, the acceleration resistance is zero and transmission resistance constant as well as calculable, as is shown in Table 2. Thus, the traction force or driveload equitation can be written in the following well-known form:

$$F_\mathrm{v} = F_\mathrm{e} + F_\mathrm{g} + F_\mathrm{l} \tag{11}$$

$$P_\mathrm{v} = F_\mathrm{v} v \tag{12}$$

Table 1: Values of θ

Gear [ith]	θ
1	0.4
2	0.3
3	0.2
4	0.1
5	0.08

Table 2: Defined variables

Known values	A, m, g, μ, c_w, ρ
Variables	v, α

2.2 MATLAB implementation

The analysis of traction force components was conducted in the MATLAB development environment to try and define how variations in velocity and road gradient can explain power requirements. An analysis was conducted based on theoretical elements and data were defined by given measurements.

Vehicle: Ford B-Max (2014)

- empty mass (m) = 1275 kg;
- maximum power (P_{max}, P_{eff}) = 74 kW;
- drag coefficient (c_{w}) = 0.32;
- frontal area (A) = 2.8 m^2;
- rolling coefficient (μ) = 0.007

Velocity codomain:

- v = [0, 140 km/h]

Road gradient codomain:

- α = [0, 30 °]

Figs. 1-3 show the effects of different resistances. The rolling resistance diagram (Fig. 1) exhibits a linear trend. The power demand increases as the velocity and road gradient increase. The climbing resistance diagram (Fig. 2) also exhibits a linear trend. According to real data, it is necessary to define a power limit, in this case 74 kW, which is the maximum power of the vehicle.

Analysis above this limit in not required since the engine is incapable of providing more power. On the contrary, the vehicle would decelerate or remain stationary beyond this limit.

The air resistance diagram (Fig. 3) exhibits a square trend between the velocity and power demand of the vehicle.

The power demands of external resistances are presented in Fig. 4. Important values were compiled in Tables 3–5.

In the first part of this chapter, constant, discrete velocities were assumed. The next step is the parameterization of acceleration. For this task, values of theta are required (Table 1).

Acceleration codomain

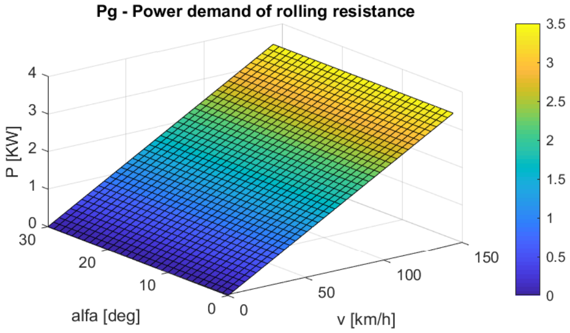


Figure 1: Diagram of the power demand of rolling resistance as a function of velocity and road gradient

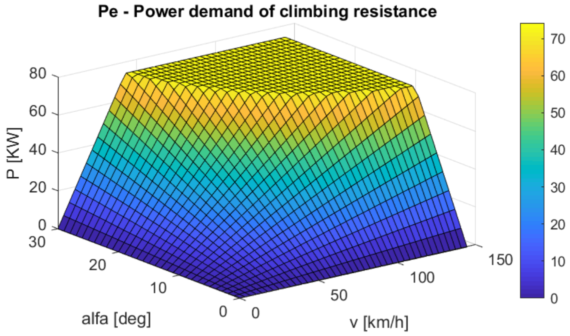


Figure 2: Diagram of the power demand of climbing resistance as a function of velocity and road gradient

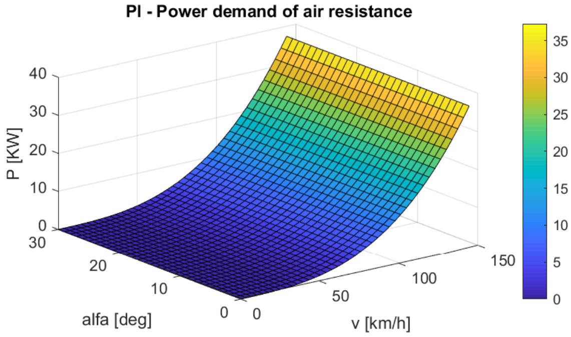


Figure 3: Diagram of the power demand of air resistance as a function of velocity and road gradient

- a = [0, 5 m/s^2]

Gravitational acceleration [G] is a dimensionless, unofficial and descriptive measure. G codomain can be derived from a codomain.

The effects of acceleration are shown in Fig. 5. It is visible that at predefined shifts, diagram flow refracts and represents real cases. Important values are compiled in Tables 6–8.

3. Results and Analyses

At high velocities and on steep road gradients, the power demand is also higher. The declaration of variables is necessary as a result of precise planning to follow on-board diagnostics (OBD) measurements, especially routes. Two independent measurement systems, OBD and GPS, are

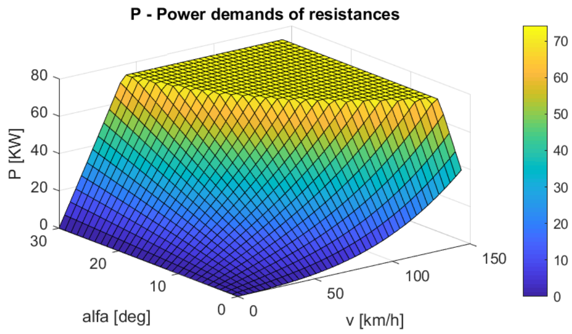


Figure 4: Diagram of the power demand of external resistances as a function of velocity and road gradient

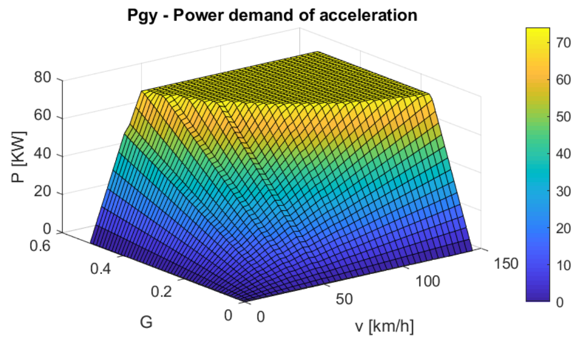


Figure 5: Diagram of the power demand of acceleration as a function of velocity and G

comparable to connect the concept [8]. Precision positioning is widely used and consists of numerous important boundary conditions. This paper examines the OBD side of the concept, details of precise GPS and GNSS measurements are presented in previous papers of ours. A statistical analysis of the fuel consumption database is given from the equation of motion.

For this database, work was used, that is the product of the force and displacement in the direction of the force.

Lifting work is

$$W_{\text{em}} = F_{\text{em}}\Delta s = mg\Delta h \tag{13}$$

Lifting work is the work that is done by lifting an object over a given period of time. It is proportional to its mass and change in height.

Friction (or rolling) work is

$$W_{\text{s}} = \mu mg\Delta s \tag{14}$$

Table 3: Notations of v and α variables

$v(\downarrow)$	low velocities
$v(\leftarrow)$	medium velocities
$v(\uparrow)$	high velocities
$\alpha(\downarrow)$	shallow road gradients
$\alpha(\leftarrow)$	medium road gradients
$\alpha(\uparrow)$	steep road gradients

Table 4: Values for $P[v, \alpha]$ calculated in the MATLAB environment

$v(\downarrow)$ = 3.6 [km/h]	α $(\downarrow)$ = 0°	P= 0.3064 [kW]
	$\alpha(\leftarrow)$ = 5°	P = 1.178 [kW]
	$\alpha(\uparrow)$ = 30°	P = 6.342 [kW]
$v(\leftarrow)$ = 50 [km/h]	α $(\downarrow)$ = 0°	P= 2.824 [kW]
	$\alpha(\leftarrow))$ = 5°	P = 18.09 [kW]
	$\alpha(\uparrow)$ = 30°	P = 74 [kW]
$v(\uparrow)$ = 140 [km/h]	α $(\downarrow)$ = 0°	P = 40.78 [kW]
	α_{max} (140) = 4°	P = 74 [kW]
	α $(\uparrow)$ = α $(\leftarrow)$ = α_{max}	

Table 5: $P[v, \alpha]$ matrix

$P[v,\alpha]$	$v(\downarrow)$	$v(\leftarrow)$	$v(\uparrow)$
$\alpha(\downarrow)$	$P(\downarrow)$	$P(\downarrow)$	$P(\leftarrow)$
$\alpha(\leftarrow)$	$P(\downarrow)$	$P(\leftarrow)$	$P(\uparrow)$
$\alpha(\uparrow)$	$P(\leftarrow)$	$P(\uparrow)$	$P(\uparrow)$

Table 6: Notation of v and G variables

$v(\downarrow)$	low velocities
$v(\leftarrow)$	medium velocities
$v(\uparrow)$	high velocities
$G(\downarrow)$	low accelerations
$G(\leftarrow)$	medium accelerations
$G(\uparrow)$	high accelerations

Table 7: Values for $P[v, G]$ calculated in the MATLAB environment

$v(\downarrow)$=3.6 [km/h]	$G(\downarrow)$=0.01	P=0.1785 [kW]
	$G(\leftarrow)$=0.1	P=1.185 [kW]
	$G(\uparrow)$ = 0.3	P = 5.93 [kW]
$v(\leftarrow)$=50 [km/h]	$G(\downarrow)$=0.01	P=2.142 [kW]
	$G(\leftarrow)$=0.1	P=21.42 [kW]
	$G(\uparrow)$=0.3	P=64.26 [kW]
$v(\uparrow)$=140 [km/h]	$G(\downarrow)$=0.01	P=5.508 [kW]
	G_{max}(140)=0.1424	P=74 [kW]
	$G(\uparrow)$=$G(\leftarrow)$=G_{max}	

Table 8: $P[v, G]$ matrix

$P[v,G]$	$v(\downarrow)$	$v(\leftarrow)$	$v(\uparrow)$
$G(\downarrow)$	$P(\downarrow)$	$P(\downarrow)$	$P(\leftarrow)$
$G(\leftarrow)$	$P(\downarrow)$	$P(\leftarrow)$	$P(\uparrow)$
$G(\uparrow)$	$P(\leftarrow)$	$P(\uparrow)$	$P(\uparrow)$

Table 9: Proportionalities over the period of time

W_{em}	v, Δh
W_{s}	v
W_{gy}	v^2
W_{k}	v^2

Table 10: Determination of coefficients

Constant velocity [km/h]	Coefficient of determination [R^2]
30	0.9549
40	0.9160
50	0.8370

Friction work is proportional to its mass and displacement.

Acceleration work is

$$W_{\mathrm{gy}} = \frac{1}{2} m \Delta v^2 \tag{15}$$

Acceleration work is proportional to its displacement, mass and the square of its velocity.

Work done by air resistance is

$$W_{\mathrm{k}} = \frac{1}{2} A c_{\mathrm{w}} \rho v^2 \Delta s \tag{16}$$

Work done by air resistance is proportional to its displacement, drag coefficient (c_{w}), frontal area (A), density (ρ) and square of its velocity. For the purpose of statistical analysis, proportionalities were compiled in Table 9.

Now the statistical analysis can be conducted. In the first step, a single variable analysis is carried out. Previously, a given route was measured, thus GPS and OBD databases were available. A connection between elevation and fuel consumption data was sought.

Table 10 shows that the concept is highly usable at low velocities, but when the range of velocities increases, the coefficient of determination becomes less efficient.

Multivariate analysis provides a solution to this problem. In this case, experienced variables, as summarized in Table 9, were used. A route comprised of different road gradients and velocities was examined.

A visual check is recommended to summarize the regression model, with which it is possible to forecast correlations according to different predictors (R^2).

Range of velocity = [20, 70 km/h]

1. Examination with $\Delta(v^2)$
   - $R^2 = 57.2\%$
   - where $\Delta(v^2)$ = variation in the square of the velocity.
2. Examination with $\Delta(v^2)$ and Δh
   - $R^2 = 86.4\%$
   - where Δh = change in height.
3. Examination with $\Delta(v^2)$, Δh and v
   - $R^2 = 87.1\%$
   - where v = actual velocity.

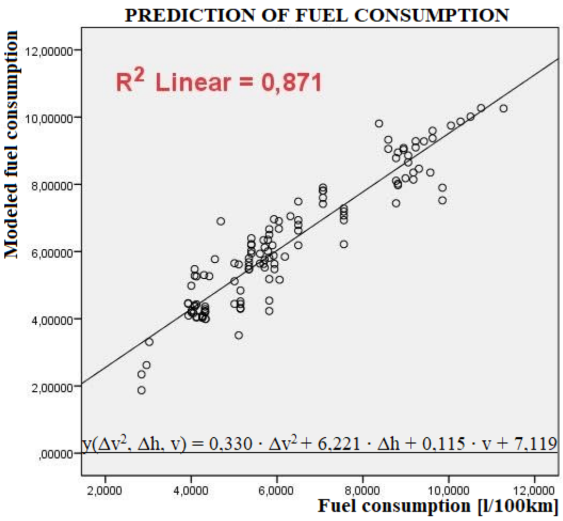


Figure 6: Results of the multivariate analysis

Fig. 6 rerspresents the regression equation with the coefficient of determination.

The regression equation can be rewritten in the following form:

$$C_{\mathrm{on}} = A\Delta(v^2) + B\Delta h + Cv + D \tag{17}$$

C_{on} is an abbreviation of fuel consumption and appears constant. It reflects other possible predictors that have not been examined, for example, losses of the internal combustion engine.

3.1 OBD-based positioning

A drawback of precision positioning devices on the market are their prices, but the OBD connectors are basic, standardized accessories of vehicles. The presented structure, when a connection is made between the positioning and on-board diagnostics, can be used for driver assistance tasks [9].

A MATLAB implementation of OBD-based positioning has been proposed that is connected to the aims of this paper and will be presented shortly.

Dataflow and the stability of the system with regard to a precision positioning measurement are crucial. Its boundary conditions are the following:

- Connection to 5 GNSS satellites simultaneously;
- Dataflow stability in terms of the satellites and the base;
- Online connection with the base, from where the correction of data originates.

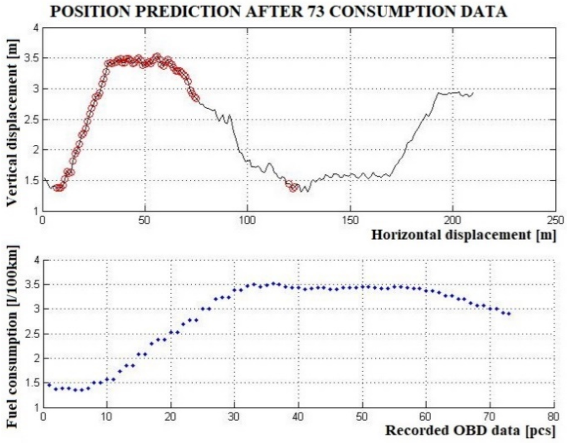


Figure 7: Operation of the MATLAB algorithm for OBD-based positioning

While weighing up the risks of two independent measurement methods, it is clear that the precision positioning technique is riskier.

A significant safety risk can be reduced if it can be substituted for other alternatives. An alternative to the elevation database of the routes and OBD data, which is accessible to every vehicle, may exist.

To summarize our MATLAB implementation, continuously incoming OBD data are compared to a reference database which consists of a map with coordinates.

By searching for the minima of the squared differences of the two databases, an algorithm was derived that is capable of defining position based on changing trends.

Fig. 7 presents the operation of the developed algorithm at a constant velocity of 30 km/h. Few incorrect OBD data points were obtained, for example, at a horizontal displacement of 125 m. As the database of fuel consumption is continuously expanding, the significance of this imprecision is decreasing.

4. Conclusion

In this article, the driveload equitation was examined and special care taken with regard to its power demands. In the MATLAB environment, characteristics of different resistances were shown. Moreover, the fixing of dependent variables was the main exercise in this research besides understanding the basic connections between on-board diagnostics and precision positioning. OBD-based positioning is a possible method to determine the actual position of a vehicle without constantly being connected to GPS or GNSS. It could be useful as part of V2X or other intelligent transportation systems. In order to extend the concept to electric vehicles, the velocity and road gradient are the main variables, the power demands of both are comparable, and optimal charging points on a given route can be calculated. This could form the basis for a future paper.

Symbols

μ	rolling resistance coefficient
m	mass of moving object
g	gravitational acceleration
v	velocity
c_w	drag coefficient
ρ	density
A	front surface
α	road gradient
θ	coefficient of rotating object
a	acceleration
G	gravitational constant
η	efficiency
η_tk	efficiency of gearbox
η_diff	efficiency of the differential
η_kt	efficiency of cardan-shaft
η_cs	efficiency of bearings
η_ny	efficiency of the clutch
h	altitude
s	displacement

Acknowledgements

This research was carried out as part of the EFOP-3.6.2-16-2017-00016 project within the framework of the New Széchenyi Plan. The completion of this project was funded by the European Union and co-financed by the European Social Fund.

REFERENCES

[1] Takács, Á.; Rudas, I.; Bösl, D.; Haidegger, T.: Highly Automated Vehicles and Self-Driving Cars, *IEEE Robotics & Automation Magazine*, 2018, **25**(4), 106–112 DOI: 10.1109/MRA.2018.2874301

[2] Derbel, O.; Peter, T.; Zebiri, H.; Mourllion, B.; Basset, M.: Modified intelligent driver model for driver safety and traffic stability improvement, *IFAC Proceedings Volumes,* 2013, **46**(21), 744–749 DOI: 10.3182/20130904-4-JP-2042.00132

[3] Iordanopoulos, P.; Mitsakis, E.; Chalkiadakis, C.: Prerequisites for Further Deploying ITS Systems: The Case of Greece, *Periodica Polytechnica Transportation Engineering*, 2018, **46**(2), 108–115 DOI: 10.3311/PPtr.11174

[4] Lim, C.; Kim, K.; Maglio, P. P.: Smart cities with big data: Reference models, challenges, and considerations, *Cities*, 2018, **82**, 86–99 DOI: 10.1016/j.cities.2018.04.011

[5] Omae, M.; Fujioka, T.; Hashimoto, N.; Shimizu, H.: The application of RTK-GPS and Steer-by-wire technology to the automatic driving of vehicles and an evaluation of driver behavior, *IATSS Research*, 2006, **30**(2), 29–38 DOI: 10.1016/S0386-1112(14)60167-9

[6] Péter, T.; Bokor, J.: Modeling road traffic networks for control, *Annual International Conference on Network Technology & Communications: NTC 2010*, 2010, Paper 21, 18–22 ISBN: 978-981-08-7654-8

[7] Péter, T.; Bokor, J.: New road traffic networks models for control, *GSTF International Journal on Computing*, 2011, **1**(2), 227–232 DOI: 10.5176/2010-2283_1.2.65

[8] Sun, Q.; Xia, J.; Foster, J.; Falkmer, T.; Lee, H.: Pursuing Precise Vehicle Movement Trajectory in Urban Residential Area Using Multi-GNSS RTK Tracking, *Transportation Research Procedia,* 2017, **25**, 2356–2372 DOI: 10.1016/j.trpro.2017.05.255

[9] Busznyák, T.; Pálfi, G.; Lakatos, I.: On-Board Diagnostic-based Positioning as an Additional Information Source of Driver Assistant Systems, *Acta Polytechnica Hungarica*, 2019, 16(5), 217–234 ISSN: 1785-8860

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 95–107 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-15

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
HJIC

BROWNIAN DYNAMICS SIMULATION OF CHAIN FORMATION IN ELECTRORHEOLOGICAL FLUIDS

DÁVID FERTIG*[1], DEZSŐ BODA[1], AND ISTVÁN SZALAI[2,3]

[1]Department of Physical Chemistry, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY
[2]Institute of Physics and Mechatronics, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY
[3]Institute of Mechatronics Engineering and Research, University of Pannonia, Gasparich Márk u. 18/A, Zalaegerszeg, 8900, HUNGARY

Brownian dynamics (BD) simulations based on a novel Langevin integrator algorithm are used to simulate the dynamics of chain formation in electrorheological (ER) fluids that are non-conducting solid particles suspended in a liquid that has a dielectric constant different from that of the ER particles. An external electric field induces polarization charge distributions on the spheres' surfaces that can be modeled as point dipoles in the centers of the spheres. The interaction of these aligned dipoles leads to formation of chains and other aggregates in the ER fluid. In this work, we introduce our methodology and report results for various quantities characterizing the structure of the ER system as obtained with BD simulations. These quantities include the potential energy, diffusion constant, average chain length, chain length distributions, and pair correlation functions. Their behavior as a function of time is presented as the electric field is switched on. The properties of the ER fluid change considerably making this system a potential basic material of many applications.

Keywords: electrorheological fluids, chain formation, Brownian dynamics

1. Introduction

Electrorheological (ER) fluids are [1] suspensions of fine non-conducting solid particles in an electrically insulating liquid. If the particles, imagined as closely spherical, have a dielectric constant that is different from that of the solvent, the arising dielectric boundaries respond to an applied electric field. This dielectric response is the polarization of the spheres resulting in a polarization charge distribution whose dominant component in the multipole expansion is the dipole moment.

The interactions of these dipoles then lead to a structural change in the ER fluid known as the ER response. This structural change is basically a formation of chains and other forms of clusters as the polarized spheres are linked together into head-to-tail positions. This structural phase transition is reversible and relatively fast.

This structural change results in a dramatic change in the physical properties of the ER fluid of which the most important is viscosity. This externally controllable, fast and reversible change in viscosity makes ER fluids a kind of a smart material, a central component of devices, such as brakes, clutches, dampers, and valves [2, 3]. Such devices have crucial importance in the industry of various fields.

The continuously shrinking size of devices resulted in the development of nanotechnology. Understanding the molecular mechanisms behind the workings of nanodevices is especially important because better understanding of microscopic mechanisms can lead to novel designs.

ER devices are also based on microscopic mechanisms leading to an emergent macroscopic pattern. No wonder that many modeling studies [4–22] aimed at investigating the microscopic processes behind chain formation and corresponding changes in measurable physical properties.

The properties of the ER fluid in the absence of an applied electric field have been investigated by Heyes and Melrose [23]. This means the investigation of the core potential that is either the Lennard-Jones (LJ) fluid or its cut-and-shifted version that is a purely repulsive potential. It has been demonstrated that the repulsive version reproduces experimental behavior better [4].

Cluster formation has been investigated via cluster size distribution [4, 9, 11, 12, 20, 22], order parameters [12–15, 19], mean square displacement and diffusion constant [4, 6, 12], pair distribution functions [6, 12], and relaxation times [5, 11, 12, 21]. In particular, Cao et al. [21] identified relaxation times corresponding to various subprocesses such as initial aggregation, chain formation, and column formation. Identifying these subprocesses is

*Correspondence: fertig.david92@gmail.com

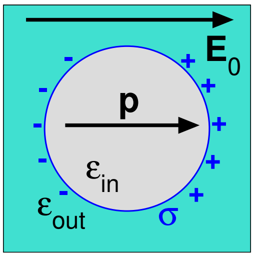


Figure 1: Sketch of an ER particle in an external electric field, $\mathbf{E}_0$. The dielectric constant inside the sphere is ϵ_{in}, while outside the sphere is ϵ_{out}. The surface charge distribution, $\sigma(\mathbf{r})$, induced on the dielectric boundary (Eq. 1) can be approximated by a point dipole, $\mathbf{p}$, in the center of the sphere (Eq. 2).

also our long-term goal. It is also our intention to simulate the ER system in the presence of shear as several authors did [5, 6, 10, 15, 21]. These authors investigated shear stress, various terms of viscosity, oscillatory strain, and dependence on strain rate.

In this paper, we do not apply stress, because our main interest is to study the dynamics of the formation of chains with a newly developed simulation package based on a novel Langevin integrator [24–26] as opposed to most studies from the 1990s that used the overdamped limit. We intend to test the program on the ER fluid in the absence and presence of an applied electric field and to follow the dynamics of chain formation when the field is switched on. We characterize this dynamics by plotting energy, mean square displacement, diffusion constant, average chain length, chain length distributions, and radial distribution functions as functions of time.

We use reduced units in this study (see Section 4) that are closely related to various parameters used in the literature. These parameters characterize the relations of various effects in the ER fluid. These effects are the polarization (dipole-dipole), thermal, and viscous forces.

The relation of the polarization and thermal forces is often denoted by λ and it practically corresponds to the square of the reduced dipole moment used in this study. It expresses the relation of the ordering effect of electrostatic forces and the disordering effect of thermal motion. The relation of the viscous force to the electrostatic force is called the Mason number (Ma). Many authors plot the characteristic physical quantities as functions of the Mason number [5,10,15]. The relation of the viscous and the electrostatic forces is called the Péclet number.

.

2. Model: the polarizable dielectric sphere

We model the ER fluid as dielectric spheres of dielectric constant ϵ_{in} inside the sphere immersed in a fluid of dielectric constant ϵ_{out} (Fig. 1). The radius of the spheres is R, while their diameter is $d = 2R$. When a constant electric field, $\mathbf{E}_0$ is applied to this system (in the z direction), the dielectric boundary on the sphere's surface becomes polarized. The polarization charge density is

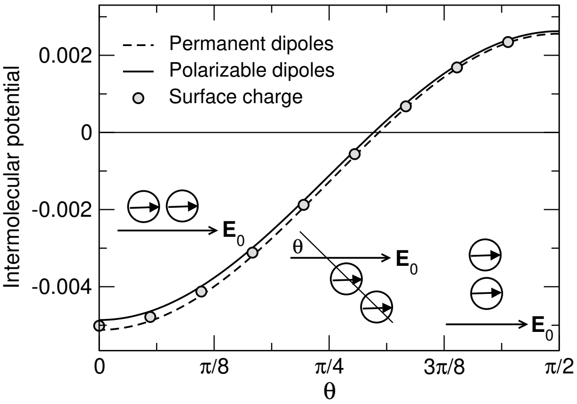


Figure 2: Interaction potential (arbitrary unit) between two dipoles at $r = 1.25d$ distance from each other at different mutual positions characterized by angle θ that is the angle between $\mathbf{E}_0$ and $\mathbf{r}_{ij}$. The potential is computed from the interaction of the charge distributions in Eq. 1 using the ICC method (symbols), from the interactions of the permanent point dipoles induced only by $\mathbf{E}_0$ (Eq. 2) (dashed line), and from the interaction of the polarizable dipoles when the sphere can be polarized by the electric field of other dipoles too (solid line).

$$\sigma(\theta) = 3\epsilon_0 \left(\frac{\epsilon_{\text{in}} - \epsilon_{\text{out}}}{\epsilon_{\text{in}} + 2\epsilon_{\text{out}}} \right) E_0 \cos\theta, \qquad (1)$$

where $E_0 = |\mathbf{E}_0|$, θ is the angle between the point of on the surface and the z-axis, and ϵ_0 is the permittivity of vacuum. As it was discussed in our previous publication [30], the effect of this surface charge distribution can be approximated with an ideal point dipole placed in the center of the sphere computed as [31]

$$\mathbf{p} = 4\pi\epsilon_0 \left(\frac{\epsilon_{\text{in}} - \epsilon_{\text{out}}}{\epsilon_{\text{in}} + 2\epsilon_{\text{out}}} \right) R^3 \mathbf{E}_0. \qquad (2)$$

In that paper, we showed that the point dipole model is a good approximation to the exact solution obtained from the polarization charge using the Induced Charge Computation method [32]. The agreement is better if the spheres are assumed to be polarizable by the electric fields of all the other particles, but even if it is assumed that an ER particle is polarized only by $\mathbf{E}_0$, the agreement is reasonable (Fig 2). The latter assumption means that the ER particles carry only the permanent dipoles of Eq. 2 that always point into the z-direction.

We further assume that the characteristic time of the rearrangement of the surface charge as the particles move is much smaller than the characteristic time of the rotation of the particles. This means that the $\mathbf{p}$ dipole always points into the z direction even if the sphere rotates, because the induced charges (that chiefly correspond to polarization of solvent molecules around the sphere) always have enough time to rearrange themselves according to the applied field, $\mathbf{E}_0$.

The potential produced by a dipole $\mathbf{p}_j$ (that is at $\mathbf{r}_j$)

at the position $\mathbf{r}_i$ of another dipole $\mathbf{p}_i$ is

$$\Phi_j(\mathbf{r}_i) = \frac{1}{4\pi\epsilon_0}\frac{\mathbf{p}_j \cdot \mathbf{r}_{ij}}{r_{ij}^3}, \tag{3}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $r_{ij} = |\mathbf{r}_{ij}|$. The electric field is

$$\mathbf{E}_j(\mathbf{r}_i) = \frac{1}{4\pi\epsilon_0}\frac{3\mathbf{n}_{ij}(\mathbf{n}_{ij}\cdot\mathbf{p}_j) - \mathbf{p}_j}{r_{ij}^3}, \tag{4}$$

where $\mathbf{n}_{ij} = \mathbf{r}_{ij}/r_{ij}$. The interaction potential between the two dipoles is

$$\begin{aligned} u_{ij}^{\mathrm{DD}}(\mathbf{r}_{ij}, \mathbf{p}_i, \mathbf{p}_j) = -\mathbf{p}_i \cdot \mathbf{E}_j(\mathbf{r}_i) = \\ = -\frac{1}{4\pi\epsilon_0}\frac{3(\mathbf{n}_{ij}\cdot\mathbf{p}_i)(\mathbf{n}_{ij}\cdot\mathbf{p}_j) - \mathbf{p}_i\cdot\mathbf{p}_j}{r_{ij}^3}, \end{aligned} \tag{5}$$

while the force exerted on dipole $\mathbf{p}_i$ by dipole $\mathbf{p}_j$ is

$$\begin{aligned} \mathbf{f}_{ij}^{\mathrm{DD}}(\mathbf{r}_{ij}, \mathbf{p}_i, \mathbf{p}_j) = -(\mathbf{p}_i \cdot \nabla_i)\mathbf{E}_j(\mathbf{r}_i) = \\ = \frac{1}{4\pi\epsilon_0}\frac{1}{r_{ij}^4}\{3\left[\mathbf{p}_i(\mathbf{n}_{ij}\cdot\mathbf{p}_j) + \mathbf{p}_j(\mathbf{n}_{ij}\cdot\mathbf{p}_i) +\right. \\ \left. +\ \mathbf{n}_{ij}(\mathbf{p}_i\cdot\mathbf{p}_j)\right] - 15\mathbf{n}_{ij}(\mathbf{n}_{ij}\cdot\mathbf{p}_i)(\mathbf{n}_{ij}\cdot\mathbf{p}_j)\}. \end{aligned} \tag{6}$$

Note that the forms of these equations are simplified when all the dipoles of magnitude p are aligned in the z direction:

$$u_{ij}^{\mathrm{DD}}(r_{ij}, \theta) = -\frac{p^2}{4\pi\epsilon_0}\frac{3\cos^2\theta - 1}{r_{ij}^3}, \tag{7}$$

and

$$\mathbf{f}^{\mathrm{DD}}(r_{ij}, \theta) = \frac{3p^2}{4\pi\epsilon_0}\frac{(2\cos\theta)\mathbf{k} + (1 - 5\cos^2\theta)\mathbf{n}_{ij}}{r_{ij}^4}, \tag{8}$$

where $\mathbf{k}$ is the unit vector in the direction of the z-axis and θ is the angle between $\mathbf{k}$ and $\mathbf{n}_{ij}$. There is also a torque acting on the dipole, but because the characteristic time of polarization charge formation is much smaller than the characteristic time of the rotation of the sphere, the rearrangement of surface charges is considered instantaneous without inertia. The torque, therefore, has been neglected.

The full interaction potential between two ER particles consists of this dipole-dipole (DD) term and a short-range core potential that defines the finite size of the particles:

$$u_{ij} = u_{ij}^{\mathrm{DD}} + u_{ij}^{\mathrm{WCA}}. \tag{9}$$

For the core potential, we use the cut & shifted LJ potential also known as the Weeks-Chandler-Anderson (WCA) potential that is

$$u_{ij}^{\mathrm{WCA}}(r_{ij}) = \begin{cases} u_{ij}^{\mathrm{LJ}}(r_{ij}) + u_{ij}^{\mathrm{LJ}}(r_{\mathrm{c}}) & \text{if} \quad r_{ij} < r_{\mathrm{c}} \\ 0 & \text{if} \quad r_{ij} > r_{\mathrm{c}} \end{cases}, \tag{10}$$

where

$$u_{ij}^{\mathrm{LJ}}(r_{ij}) = 4\varepsilon^{\mathrm{LJ}}\left[\left(\frac{d}{r_{ij}}\right)^{12} - \left(\frac{d}{r_{ij}}\right)^{6}\right] \tag{11}$$

is the LJ potential. The WCA force is

$$\mathbf{f}_{ij}^{\mathrm{WCA}}(\mathbf{r}_{ij}) = \begin{cases} \mathbf{f}_{ij}^{\mathrm{LJ}}(\mathbf{r}_{ij}) & \text{if} \quad r_{ij} < r_{\mathrm{c}} \\ 0 & \text{if} \quad r_{ij} > r_{\mathrm{c}} \end{cases}, \tag{12}$$

where

$$\mathbf{f}_{ij}^{\mathrm{LJ}}(\mathbf{r}_{ij}) = 24\varepsilon^{\mathrm{LJ}}\left[2\left(\frac{d}{r_{ij}}\right)^{12} - \left(\frac{d}{r_{ij}}\right)^{6}\right]\frac{\mathbf{r}_{ij}}{r_{ij}^2} \tag{13}$$

is the LJ force. In these equations the cutoff distance is $r_{\mathrm{c}} = 2^{1/6}d$ that is at the minimum of the LJ potential, so this potential is a smooth repulsive core potential used widely in dynamical simulations of large spherical particles.

3. Method: Brownian Dynamics simulation

When it comes to simulating the trajectories of particles in the phase space interacting with each other via a systematic force, $\mathbf{f}_{ij}$ (like those given in Eqs. 6 and 12), we use Newton's equation of motion in an MD simulation. In this case, the particles move in vacuum and the only forces that we take into account are those exerted by the particles themselves (plus, possibly, external forces).

When it comes to simulating the trajectories of particles immersed in a solvent, we use Langevin's equations of motion [33]

$$m\frac{d\mathbf{v}_i(t)}{dt} = \mathbf{F}_i\left(\mathbf{r}_i(t)\right) - m\gamma\mathbf{v}_i(t) + \mathbf{R}_i(t), \tag{14}$$

where $\mathbf{r}_i$, $\mathbf{v}_i$, m, and γ are the position, the velocity, the mass, and the friction coefficient of particle i, respectively. The mass and the friction coefficient are assumed to be the same for every particle, but, in general, they can depend on i.

The force has three components. In addition to the systematic force, $\mathbf{F}_i\left(\mathbf{r}_i(t)\right) = \sum_{j\neq i}\mathbf{f}_{ij}$, there are the frictional force, $-m\gamma\mathbf{v}_i(t)$, and the random force, $\mathbf{R}_i(t)$. The former describes friction, while the latter describes random collisions with surrounding solvent molecules.

The two additional forces represent the interactions with the heat bath and are coupled through the friction coefficient:

$$\langle\mathbf{R}(t)\rangle = 0 \tag{15}$$

$$\langle\mathbf{R}(t)\cdot\mathbf{R}(t')\rangle = 2kTm\gamma\delta(t - t') \tag{16}$$

This is also known as the fluctuation–dissipation theorem.

The Langevin equation is a stochastic differential equation that is solved numerically and, therefore, approximately. Several algorithms exist in the literature for its integration [34–37].

Here, we employ the simple and effective algorithm of Grønbech-Jensen and Farago (GJF). The original version [24] had a Verlet-type formalism. Recent modifications by Farago (GJF-F) [25] and Grønbech Jensen and

Table 1: Reduced quantities

Quantity	Symbol	Unit quantity	Reduced quantity
Time	t	$t_0 = d\sqrt{\frac{m}{kT}}$	$t^* = \frac{t}{d}\sqrt{\frac{kT}{m}}$
Distance	r	$r_0 = d$	$r^* = \frac{r}{d}$
Density	ρ	$\rho_0 = \frac{1}{d^3}$	$\rho^* = \rho d^3$
Velocity	v	$v_0 = \frac{d}{t_0} = \sqrt{\frac{kT}{m}}$	$v^* = v\sqrt{\frac{m}{kT}}$
Energy	u	$u_0 = kT$	$u^* = \frac{u}{kT}$
Force	F	$F_0 = \frac{kT}{d}$	$F^* = \frac{Fd}{kT}$
Dipole moment	p	$p_0 = \sqrt{4\pi\epsilon_0 kTd^3}$	$p^* = \frac{p}{\sqrt{4\pi\epsilon_0 kTd^3}}$
Friction coefficient	γ	$\gamma_0 = \frac{1}{t_0} = \frac{1}{d}\sqrt{\frac{kT}{m}}$	$\gamma^* = \gamma t_0 = \gamma d\sqrt{\frac{m}{kT}}$.

Grønbech-Jensen (GJF-2GJ) [26] have a leap-frog formalism using velocities in the half time steps. These modifications have the advantage that they accurately sample both kinetic and configurational properties even for large time steps within the stability limit. The authors demonstrated the efficiency of their algorithms for systems under linear and harmonic potentials. We use the GJF-2GJ version in this work that reads as

$$v^{n+\frac{1}{2}} = av^{n-\frac{1}{2}} + \frac{\sqrt{b}\Delta t}{m}f^n + \frac{\sqrt{b}}{2m}\left(R^n - R^{n+1}\right) \quad (17)$$

$$r^{n+1} = r^n + \sqrt{b}v^{n+\frac{1}{2}}\Delta t, \quad (18)$$

where $r^n = r(t^n)$ is any position coordinate of any particle, $v^n = v(t_n)$ is any velocity coordinate of any particle,

$$a = \frac{1 - \gamma\Delta t/2}{1 + \gamma\Delta t/2}, \quad (19)$$

$$b = \frac{1}{1 + \gamma\Delta t/2}, \quad (20)$$

Δt is the time step, $t_{n+\frac{1}{2}} = t_n + \frac{\Delta t}{2}$, and $t_{n-\frac{1}{2}} = t_n - \frac{\Delta t}{2}$. The discrete time noise

$$R^{n+1} = \int_{t_n}^{t_{n+1}} R(t')dt' \quad (21)$$

is a random Gaussian number with properties

$$\langle R^n \rangle = 0 \quad (22)$$

and

$$\langle R^m R^n \rangle = 2kT\gamma m\Delta t\delta_{mn} \quad (23)$$

with δ_{mn} being the Kronecker-delta.

4. Scaling and reduced units

Competing effects exist in an ER system. The DD interactions have an ordering effect. The head-to-tail position, in which the dipoles are aligned along $\mathbf{n}_{ij}$ ($\theta = 0$) at contact ($r_{ij} = d$), has a minimum energy with the value

$$u_0 = -\frac{1}{4\pi\epsilon_0}\frac{2p^2}{d^3}. \quad (24)$$

The magnitude of the force in this position is

$$f_0 = \frac{3p^2}{4\pi\epsilon_0 d^4}. \quad (25)$$

The Brownian motion has a disordering effect that expresses the coupling to a thermostat of temperature T and friction with the surrounding solvent with viscosity η. It is usual to characterize the disordering effect of the thermal motion energetically by kT. It is also usual to use reduced units in calculations. In reduced units our quantities are expressed as dimensionless numbers obtained by dividing a quantity in a physical unit by a unit quantity in the same unit, $t^* = t/t_0$, for example. Reduced quantities are useful not only because their values are close to 1, so it is easier to work with them, but also because they express relations between quantities in the numerator and the denominator, a kind of scaling [5].

There are different ways of defining reduced units. We use the convention of building the unit quantities from the mass, m, the particle diameter, d, and kT. Thus, the reduced units collected in Table 1 can be defined.

When we perform simulations in reduced units, these quantities can be chosen freely to see how the system behaves at the different combinations of the reduced parameters. How the reduced parameters are related to real-life physical parameters can be computed independently (see Section 5).

Table 2: Experimental parameters [38, 39].

ϵ_{in}	4
ϵ_{out}	2.7
η (Pa s)	0.5
E_0 (V/m)	10^6
$T(K)$	300
ρ_{out} (kg/m^3)	2650

The reduced quantities collected in Table 1 are determined by the real physical parameters of the system: the temperature, T, the mass density of the material of the ER particle, ρ_{in}, the diameter of the ER particle, d, the dielectric constant of the ER particle, ϵ_{in}, the dielectric constant of the solvent, ϵ_{out}, the viscosity of the solvent, η, and the strength of the applied electric field, E_0. For a specific ER fluid, these variables are tabulated in Table 2. This specific example is used because one of the coauthors (I.SZ.) published experimental results for this system [38, 39]. A wide variety of ER fluids exists, however.

The mass of a particle is computed as $m = \rho_{\text{in}}\pi d^3/6$, so it scales with d^3. The dipole moment of a particle is given by Eq. 2 that shows that p scales with d^3.

An important parameter is the ratio of the dipolar energy and the thermal energy that is expressed by the square of the reduced dipole moment:

$$(p^*)^2 = \frac{\pi\epsilon_0 E_0^2}{4kT}\left(\frac{\epsilon_{\text{in}} - \epsilon_{\text{out}}}{\epsilon_{\text{in}} + 2\epsilon_{\text{out}}}\right)^2 d^3 = Kd^3 \qquad (26)$$

that scales with d^3. If p^* is large, the dipolar interactions are strong enough to induce chain formation. If p^* is too large, the chains freeze, and the ER particles solidify (note that the fluid itself does not solidify). If p^* is small, thermal motion prevents chain formation and/or breaks the chains.

The friction coefficient can be computed from Stokes' law as

$$\gamma = \frac{3\pi\eta d}{m} = \frac{18\eta}{\rho_{\text{in}}} d^{-2}. \qquad (27)$$

so it scales with d^{-2}. The value of γ^* describes the strength of the coupling with the solvent and it scales with $d^{1/2}$. If γ^* is large, friction and the disordering effect of the random force are strong. The diffusivity of the particles in the fluid, therefore, will be smaller. The diffusion constant in the high coupling limit can be expressed by Einstein's relation:

$$D = \frac{kT}{m\gamma}, \qquad (28)$$

or, in reduced units, $D^* = 1/\gamma^*$. If $\gamma^* \to 0$, the frictional and the random forces vanish, and the Langevin equation goes into the Newton equation. The particles move in vacuum without a thermostat; this practically corresponds to an MD simulation in the microcanonical ensemble. If γ^* is small, we talk about an MD simulation with a Langevin thermostat.

In the case of the ER fluids, we are in the regime of large γ^*. As we will see, γ^* is in the order of $10^4 - 10^6$. In this case, our concern is how to make the simulation efficient in order to collect enough information about the dynamics of the system in a reasonable amount of computer time.

The parameter with which we can tune the speed of sampling is the time step, Δt^*. This parameter is also subject of optimization. If Δt^* is too small, the simulation will evolve slowly at the price of expensive computation time. If Δt^* is too large, the spheres might overlap and the repulsive core force (Eq. 12) becomes so large that the particles shoot apart resulting in unphysical movements. This leads to instabilities in solving the Langevin equation.

Various solutions have been proposed in the literature to cope with this problem. If the Langevin integration algorithm allows changing the time step during the simulation, it is a reasonable suggestion to reduce the time step if we observe problems (generally, big jumps) in the movements of particles [6, 13]. Displacements, velocities, or forces can be monitored for unusual events.

Berti et al. [40] used a uniform time step, while their solution for the jump-problem was that they went back the necessary number of time steps and started again with a different random number seed for the random force. If such a problem is rare, this can be a good solution, because the computational cost of going back a couple of times is balanced by the large time step used in the simulation. They used their simulations for ion channels whose selectivity filter is a high-density region, so overlaps can occur. Chain formation in the ER fluid also brings particles close to each other, so we need to be careful with large time steps.

We can estimate in advance the danger of overlap and judge the optimization between slow simulations (small Δt) and jumping particles (large Δt). We can introduce the average distance that a particle moves in a time step with the average thermal velocity, $\bar{v} = \sqrt{3kT/m}$. Let us introduce

$$\overline{\Delta s^*} = \frac{\bar{v}\Delta t}{d} = \sqrt{3}\Delta t^*, \qquad (29)$$

that characterizes the average distance with respect to the particle size. This is proportional to Δt^*. This reduced distance, and, consequently, the reduced time step should be smaller than 1. This imposes a strict limit to the time step.

The product $\gamma\Delta t = \gamma^*\Delta t^*$ characterizes how close we are to the overdamped limit. Basically, at a fixed γ^*, we can increase Δt^* up to the threshold limit to save computer time at the price of losing information about dynamics due to coarser time resolution.

The last parameter that we can choose relatively freely is the energy parameter of the LJ potential, ε^{LJ}, see Eqs. 10–12. Changing this parameter practically changes the effective diameter of the particles. Fig. 3 shows the curves of the core potential (Eq. 10) for varying values of ε^{LJ}. Smaller values of ε^{LJ} allows for the particles to

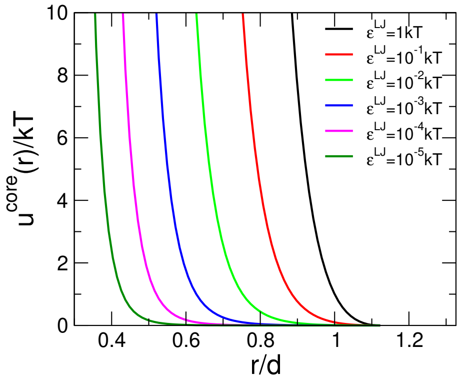


Figure 3: The core potential, $u^{\rm WCA}(r)$, for varying energy parameters, $\varepsilon^{\rm LJ}$.

approach each other closer: the r/d values at which the core potential reaches large values in kT are smaller for smaller $\varepsilon^{\rm LJ}$ values. The effective diameter, $d_{\rm eff}$, therefore decreases with decreasing $\varepsilon^{\rm LJ}$.

This results in larger dipole-dipole interactions at contact positions that, in turn, increases the weight of the dipolar interactions with respect to the thermal noise. Using smaller $\varepsilon^{\rm LJ}$, and, consequently, smaller $d_{\rm eff}$, however, makes our parameter d with which we reduced every variable meaningless. We would like the diameter used in the reduced quantities to be the real diameter of the spheres. For this reason, we do not change $\varepsilon^{\rm LJ}$ and fix it at the value of kT.

5. Relating reduced units to real ER fluids

To connect to a real system, we consider the ER fluid studied by Horváth and Szalai [38, 39] experimentally. The experimental parameters are collected in Table 2. Note that the diameters used in these studies were quite small in order to prevent sedimentation. Diameters used in other ER fluids are larger reaching 1 μm.

We change two parameters in this analysis, the particle diameter, d, and the reduced time step, Δt^*. According to Eq. 2, the dipole moment can be written as $p = Kd^3$, where $K = 1.922 \times 10^{-6}$ Cm^{-2} for the parameters in Table 2. Table 3 contains various quantities computed for different values of d.

It is seen that p^* falls into the regime simulated in this study around $d = 1$ μm. For diameters below 100 nm, at least, at the present value of K, the reduced dipole moment is too weak to counterbalance the thermal motion and to produce considerable chain formation.

The reduced friction coefficient also depends on d; it increases with $d^{1/2}$. It is in the regime of $\gamma^* \approx 10^5 - 10^6$. This looks simulatable, though it will require considerable computer time, because Δt^* is limited. The parameter $\overline{\Delta s^*}$ is the same for every diameter; it practically equivalent to Δt^*. To look at the effect of Δt^*, we show the same data for varying Δt^* at a fixed d (100 nm) in Table 4.

Table 3: Change of various variables as the diameter of spheres is changed from 10 to 10,000 nm for time step $\Delta t^* = 0.001$.

d (nm)	**10**	**100**	**1,000**	**10,000**
Δt^*	0.001			
m (kg)	1.387E-21	1.387E-18	1.387E-15	1.387E-12
t_0 (s)	5.788E-09	1.830E-06	5.788E-04	1.830E-01
$\bar{v}$ (m/s)	2.993E+00	9.463E-02	2.993E-03	9.463E-05
p (Cm)	1.922E-30	1.922E-27	1.922E-24	1.922E-21
p^*	**0.00283**	**0.0896**	**2.833**	**89.60**
Δt (s)	5.788E-12	1.830E-09	5.788E-07	1.830E-04
$\overline{\Delta s^*}$	0.00173	0.00173	0.00173	0.00173
γ (1/s)	3.396E+13	3.396E+11	3.396E+09	3.396E+07
γ^*	1.966E+05	6.216E+05	1.966E+06	6.216E+06
$\gamma\Delta t$	1.966E+02	6.216E+02	1.966E+03	6.216E+04

Table 4: Change of various variables as the reduced time step Δt^* is changed from 0.0001 to 0.1 for diameter d = 100 nm.

d (nm)	100			
Δt^*	**0.0001**	**0.001**	**0.01**	**0.1**
m (kg)	1.387E-18	1.387E-18	1.387E-18	1.387E-18
t_0 (s)	1.830E-06	1.830E-06	1.830E-06	1.830E-06
$\bar{v}$ (m/s)	9.463E-02	9.463E-02	9.463E-02	9.463E-02
p (Cm)	1.922E-27	1.922E-27	1.922E-27	1.922E-27
p^*	0.0896	0.0896	0.0896	0.0896
Δt	1.830E-10	1.830E-09	1.830E-08	1.830E-07
$\overline{\Delta s^*}$	**0.000173**	**0.00173**	**0.0173**	**0.173**
γ (1/s)	3.396E+11	3.396E+11	3.396E+11	3.396E+11
γ^*	6.216E+05	6.216E+05	6.216E+05	6.216E+05
$\gamma\Delta t$	6.216E+01	6.216E+02	6.216E+03	6.216E+04

6. Results and Discussion

In this study, we use a relatively small number of particles ($N = 128$) in order to save on computer time and be able to explore a wide range of parameters in reduced units. We also fix the packing fraction expressed in term of the reduced density at $\rho^* = 0.05$. At these values the width of the simulation cell is $L = 13.68\,d$.

The computer code has been written (in Fortran) in a way that we perform M_0 time steps in the absence of applied electric field ($\mathbf{E}_0 = 0$), and $M_{\rm E}$ time steps in the presence of it. That way, we can study the dynamics of chain formation after the electric field is switched on. To improve statistics, we can perform several of this $M_{\rm c} = M_0 + M_{\rm E}$ cycles and average over the cycles.

When we start a cycle over, we can choose between two options. We can either continue the simulation from the previous phase state point (configurations and velocities) only without dipoles, or we can restart from a freshly generated initial configuration. In this work, we choose the second option. This choice ensures that we start the simulation with nonzero E_0 in a completely disordered state without chains. The first option makes it possible to study the dynamics of the deconstruction of the chains.

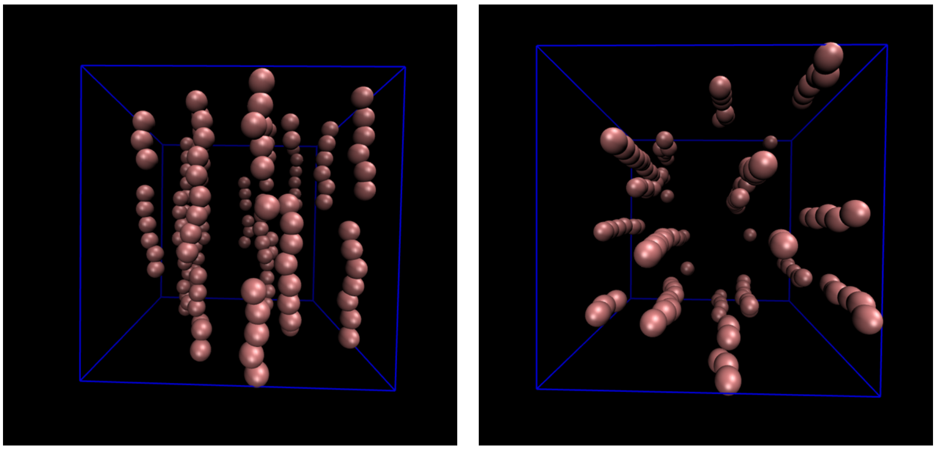


Figure 4: Typical snapshot of a simulation from the front (perpendicular to the z axis, left panel) and top (parallel to the z axis, right panel) for a state when chains are formed.

6.1 Quantities studied

As the chains are being formed, certain physical quantities change, so they directly or indirectly characterize chain formation quantitatively. In chains, particles are aligned into head-to-tail position along the z-axis as shown in Fig. 4. There are longer and shorter chains and the distribution of chains of various lengths changes continuously as the simulation evolves.

Since the head-to-tail position is the lowest energy configuration of the ER spheres (see Fig. 2 and Eqs. 24 and 25), the average one-particle dipole-dipole energy, $\langle u^{\mathrm{DD}}\rangle_{\mathrm{b}}/kT$, is a good indicator of chain formation. As it turns out, it is the best converging indicator.

By average, we mean average over a block in the simulation, denoted by $\langle \ldots \rangle_{\mathrm{b}}$. The length of a block (M_{b} is the number of time steps in a block), again, is a subject of optimization. If a block is too short, the physical quantities averaged over a block will have bad statistics. If a block is too long, we loose information about the dynamics of the system.

Diffusion constant When the particles are "frozen" into chains, their mobility decreases. Chains are frozen only at very large dipole moments, when even columnar structures are formed. In a moderate range of $(p^*)^2$, chains move around, break apart, and rejoin, see the video clip at https://youtu.be/OwXsuz6p0W4. A snapshot of this video clip is shown in Fig. 5.

The isotropic diffusion constant is computed as the slope of the mean square displacement (MSD) as a function of time:

$$D_{\mathrm{b}} = \frac{\langle \mathbf{r}^2(t)\rangle_{\mathrm{b}}}{2t_{\mathrm{b}}}, \tag{30}$$

where $\langle \ldots \rangle_{\mathrm{b}}$ denotes an average over time steps in a block and particles and t_{b} is the length of the block in time. The exact equilibrium diffusion constant is obtained in the limit of $t_{\mathrm{b}} \to \infty$.

Here, we must be satisfied with an approximate value of D_{b} obtained over a block of limited length. Fig. 6 shows the MSD as a function of t^* for six equidistantly chosen blocks. In this particular case, $\gamma^* = 5000$, so the slope is $D^* = \mathrm{MSD}/t^*_{\mathrm{b}} \approx 0.0002$ for the WCA fluid as

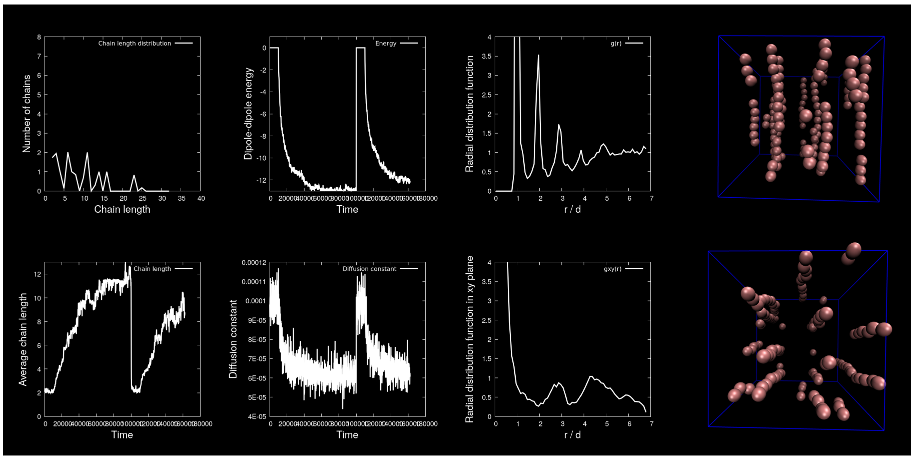


Figure 5: A snapshot of the video clip at https://youtu.be/OwXsuz6p0W4.

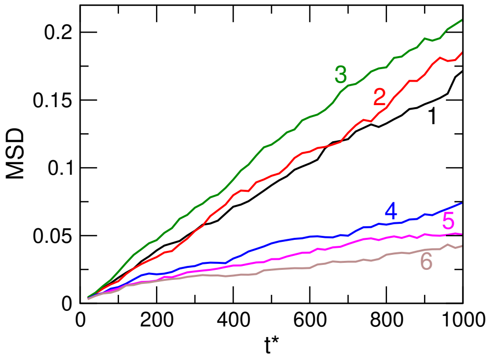


Figure 6: The mean square displacement for six selected blocks. The blocks are selected in equidistant time periods in way that the first three belong to the $E = 0$ phase, while the second three belongs to the ER phase. Parameters: $(p^*)^2 = 6$, $\gamma^* = 5{,}000$, $\Delta t^* = 0.02$, $M_\mathrm{b} = 50{,}000$.

also expressed by the Einstein relation ($D^* = 1/\gamma^*$, Eq. 28). Here, the time-length of the block is $t^*_\mathrm{b} = \Delta t^* M_\mathrm{b} = 1{,}000$, because $\Delta t^* = 0.02$ and $M_\mathrm{b} = 50{,}000$. The first three lines are in the $E = 0$ regime, while the second three lines are in the ER regime. The slope apparently is smaller in the ER case than in the WCA case, but the scattering is large.

The sampling can be improved by averaging over cycles, but this does not help on the problem of the diffusion constant being approximate obtained for a too short block.

Chain length distributions The chain formation can be directly followed by identifying chains in every configuration. If that is done, we can obtain the number of chains, n_s, having length s. The average chain length can be computed as

$$l = \frac{\sum_s s n_s}{\sum_s n_s}. \tag{31}$$

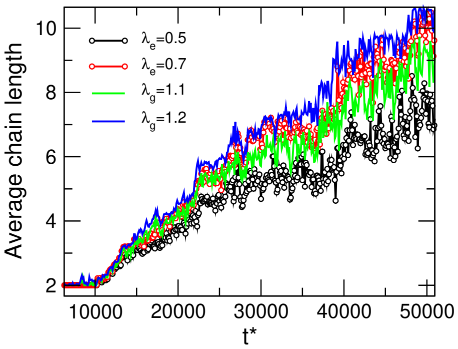


Figure 7: The trend of the change in the average chain length with various definitions of a chain: energetic with $\lambda_\mathrm{e} = 0.5$ and 0.7, geometrical with $\lambda_\mathrm{g} = 1.1$ and 1.2. Parameters: $(p^*)^2 = 6$, $\gamma^* = 10{,}000$, $\Delta t^* = 0.01$, $M_\mathrm{b} = 10{,}000$.

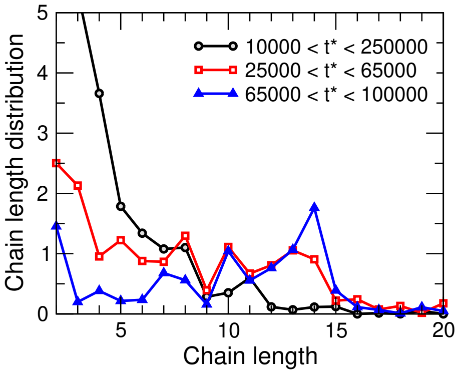


Figure 8: Chain length distribution averaged over three time intervals at the beginning ($10{,}000 < t^* < 25{,}000$), in the middle ($25{,}000 < t^* < 65{,}000$), and at the end ($65{,}000 < t^* < 100{,}000$) of chain formation. Parameters are the same as at Fig. 7.

This quantity than can be averaged over time steps in a block, so chain formation can be followed by plotting the average chain length, $\langle l \rangle_\mathrm{b}$, as a function of time in steps of t^*_b.

A chain, however, can be defined in various ways. One simple definition is geometrical. If two particles are closer to each other than a predefined distance:

$$r_{ij} < \lambda_\mathrm{g} d, \tag{32}$$

they are said to be part of the same chain. Another definition is energetic. If the dipole-dipole interaction energy is smaller than a predefined threshold:

$$u^\mathrm{DD}_{ij}(r_{ij}, \theta) < \lambda_\mathrm{e} u_0, \tag{33}$$

then they are said to be part of the same chain, where u_0 is the DD interaction energy in the head-to-tail position (Eq. 24).

Fig. 7 shows the increase of the average chain length as a function of time as obtained from different chain definitions and thresholds λ_e and λ_g. In general, the trends as shown by the various definitions are the same. The dynamic process of chains breaking up and reforming have the same effect in the cases of the various definitions. This process can be characterized by time constants obtained from fitting exponential functions. These time constants are insensitive to the choice of the chain definition. Here, we will use the geometrical definition with the parameter $\lambda_\mathrm{g} = 1.2$. The geometrical definition is advantageous, because it can also be used in the absence of an electric field.

The average chain length is an informative, but averaged quantity. From the simulations, we have the more detailed n_s vs. s chain length distributions that give the average number of chains of different lengths as a function of s. This function varies with time, see the video clip at https://youtu.be/OwXsuz6p0W4. To show the dynamics of this function, we average it for three distinct time intervals. The first one refers to the

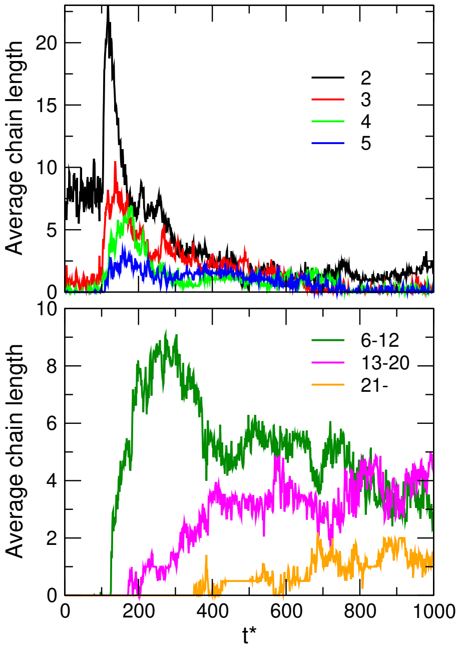


Figure 9: The variation of the number of chains of various lengths in time. Top: chains of lengths 2, 3, 4, and 5. Bottom: number of chains belonging to the ranges $6-12$, $13-20$, and above 21. Parameters are the same as at Fig. 7.

beginning of the time period in the presence of the field when the chains start forming. In the second, intermediate time interval (25,000 $< t^* <$ 65,000) longer chains are formed, while in the third time interval (65,000 $<$ $t^* <$ 100,000), full chains crossing the simulation box are formed.

Fig. 8 shows these three time-averaged functions. At the beginning, there are many pairs and short chains (black curve). In the intermediate time interval, the number of short chains decreases and longer chains are formed. In the third time interval, a well-defined peak at $s = 14$ appears that corresponds to the full chains crossing the simulation box of length $L = 13.68d$.

We can get a much better impression of the dynamics of chain formation, if we plot n_s as a function of time. Because there are too many possible n_s numbers to plot, again, we average over certain regions of chain lengths as seen in Fig. 9. In the top panel, the behavior of short chains from pairs to $s = 5$ is shown. The behavior of these chains is qualitatively similar. First, as the electric field is switched on, their numbers increase abruptly, then, as longer chains absorb them, or they fuse into longer chains, their numbers gradually decreases. Practically, they behave like reactive intermediates in chemical reactions: their formation is a first necessary step towards the formation of the end products.

The number of chains whose lengths are between 6 and 12 (bottom panel) behaves similarly. The curve for the chains whose lengths are between 13 and 20, however, saturates around $n_s = 4$. This means that there are generally about 4 full chains in the simulation box (this value, of course, depends on system size and packing fraction). They are often accompanied by shorter chains as seen in Fig. 4 and the video clip.

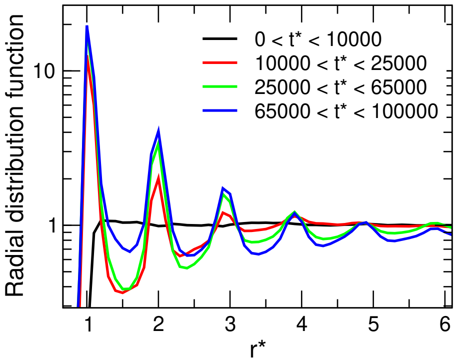


Figure 10: Radial distribution functions averaged over four time intervals in the absence of the electric field (0 $<$ $t^* <$ 10,000), at the beginning (10,000 $< t^* <$ 25,000), in the middle (25,000 $< t^* <$ 65,000), and at the end (65,000 $< t^* <$ 100,000) of chain formation. Parameters are the same as at Fig. 7.

Chains longer than 20 also exist. It can also occur that two chains are stuck together. Whether it is a stable, long time-span configuration, depends on the strength of the dipole moment (the electric field, in reality). A particle is attracted to another particle in a chain, if they are aligned in a way that $\theta = \pi/4$, see Fig. 2. This is a relatively weak attraction compared to the head-to-tail position. The chains displace due to thermal motion, so the chains move out of these mutual positions that favors aggregation of chains. If two chains move in a way that the particles get next to each other ($\theta = \pi/2$), a repulsive force replaces the weak attractive one. So, a strong dipole moment is needed to overcome the thermal motion if we want to see stable columnar aggregations of chains as seen many times in the literature.

Pair distribution functions As particles aggregate into chains, the structure of the fluid, generally expressed with pair distribution functions, changes. In an anisotropic dipolar fluid, we generally use the series expansion of the pair correlation function of axially symmetric molecules as

$$g(ij) = \sum_{nml} h^{mnl}(r_{ij})\, u^{mnl}(ij). \tag{34}$$

This expansion separates distance and angular dependence in such a way that the projections $h^{mnl}(r_{ij})$ depend only on the distance of particles and the projections $u^{mnl}(ij)$ are rotational invariants.

The projection $g(r_{ij}) = h^{000}(r_{ij})$ is the usual radial distribution function (RDF):

$$g(r_{ij}) = \int g(ij) d\Omega_i d\Omega_j\,, \quad \text{with} \quad u^{000} = 1, \tag{35}$$

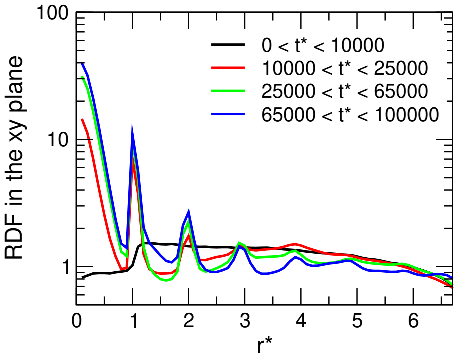


Figure 11: The xy-plane radial distribution functions averaged over four time intervals as in Fig. 10. Parameters are the same as at Fig. 7.

where Ω_i denotes molecular orientation. In a fluid phase, $h^{000}(r_{ij}) \to 1$ when $r_{ij} \to \infty$ both in isotropic and anisotropic phases. Other projections, called angular correlation functions, can also characterize chain formation, but we will discuss only the RDF in this study.

Similar to Fig. 8, we plot the RDF averaged over the time intervals discussed at the chain length distributions. In addition to those three time intervals, we also consider the time interval $0 < t^* < 10{,}000$ here, which is the time of the electric field being switched off. Fig. 10 shows that the $g(r)$ function behaves like a typical RDF for a dense real gas ($\rho^* = 0.05$) in the absence of E_0.

As the electric field is switched on, however, larger and larger peaks appear as time goes by and longer and longer chains are formed. The peaks appear at every integer multiples of d values that correspond to particles in the chain. A more detailed behavior of $g(r)$ can be followed in the video clip: `https://youtu.be/OwXsuz6p0W4`.

When the chains are formed, they are relatively stable, but they diffuse around in the xy plane. Therefore, we also define the RDF in the xy plane to follow how the chains are distributed over the xy plane. We will denote it with $g_{xy}(r)$ and is calculated the same way as the three-dimensional RDF.

Fig. 11 shows these functions averaged over the time periods as in Fig. 10. A similar conclusion can be drawn as from that figure except that the first peak now appears at $r^* = 0$, where now $r = \sqrt{\Delta x^2 + \Delta y^2}$. This peak represents particles belonging to the same chain. Peaks represent probable distances between chains. The shape of the curve indicates that this ER system ($(p^*)^2 = 6$) behaves like a two-dimensional fluid of chains.

At a given time (or, in a given block), these series of peaks are absent. Snapshots of $g_{xy}(r)$ show where the chains are in a given moment. This can be followed in the video clip: `https://youtu.be/OwXsuz6p0W4`. The $g_{xy}(r)$ function averaged over a longer time period characterizes the behavior of the chains as a two-dimensional fluid.

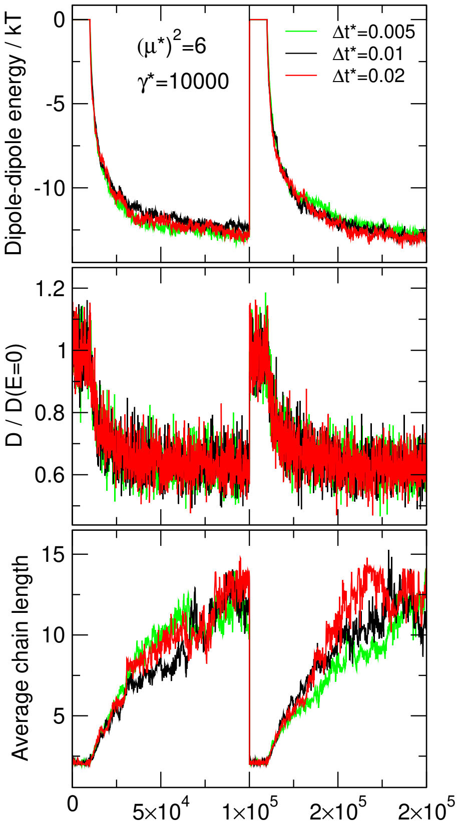


Figure 12: The one-particle dipole-dipole energy (top panel), the diffusion constant relative to its value in the absence of E_0 (middle panel), and the average chain length (geometrical definition with $\lambda_g = 1.2$, bottom panel) as functions of time using different time steps. Parameters: $(p^*)^2 = 6$ and $\gamma^* = 10{,}000$. The M_b is changed in a way that $\Delta t^* \times M_b$ is constant.

6.2 The effect of time step

First, let us consider the effect of the choice of the time step, Δt^*. Fig. 12 shows the variation of the one-particle dipole-dipole energy, the diffusion constant, and the average chain length (geometrical definition with $\lambda_g = 1.2$) for different values of Δt^*. The length of a block measured in t^* is kept fixed. It is seen that the measured quantities behave the same way as a function of time, which indicates that the BD simulation algorithm is robust and provides results that are independent of the time step.

Also, we monitored the temperature computed from the kinetic energy, $\langle T \rangle = m\langle v^2 \rangle / 3k$, and found that the algorithm reproduces the prescribed temperature very precisely even for this highly anisotropic fluid. This supports the claim of the developers that this algorithm provides a very good Langevin thermostat [24–26].

If we change Δt^*, but we keep M_b at the same value, meaning that we change the time length of the block, the dipole-dipole energy and the average chain length are still insensitive to the choice of Δt^* (data not shown). The diffusion coefficient, however, changes with the length of the blocks as already discussed above (at Fig. 6). This

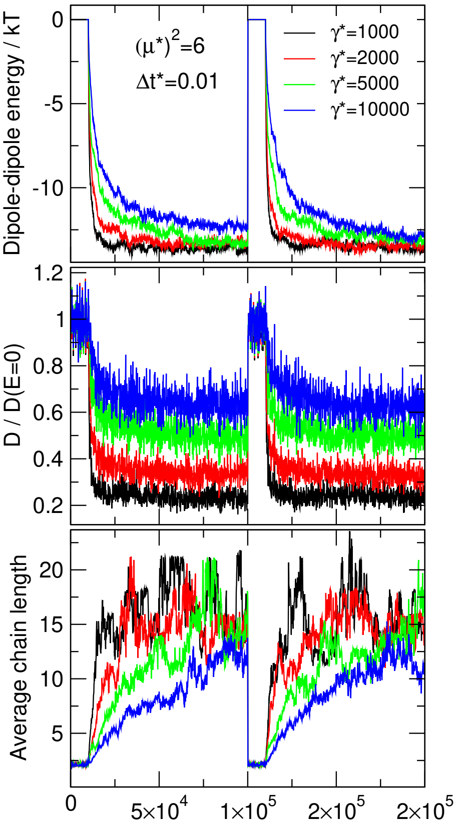


Figure 13: The one-particle dipole-dipole energy (top panel), the diffusion constant relative to its value in the absence of E_0 (middle panel), and the average chain length (geometrical definition with $\lambda_g = 1.2$, bottom panel) as functions of time using different friction coefficients. Parameters: $(p^*)^2 = 6$ and $\Delta t^* = 0.01$.

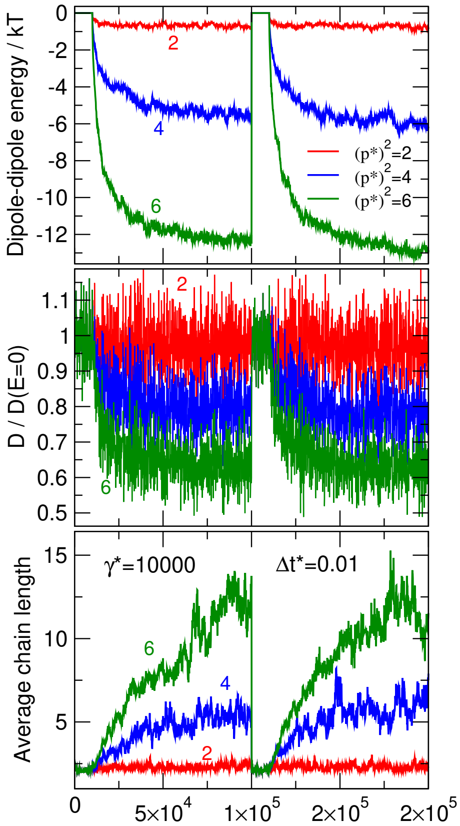


Figure 14: The one-particle dipole-dipole energy (top panel), the diffusion constant relative to its value in the absence of E_0 (middle panel), and the average chain length (geometrical definition with $\lambda_g = 1.2$, bottom panel) as functions of time using different dipole moments. Parameters: $\gamma^* = 10{,}000$ and $\Delta t^* = 0.01$. This figure shows two $M_0 + M_E$ cycles.

means that we have a trade-off between satisfactory sampling over a block and good resolution in time.

We do not consider viscosity in this paper; we refer it to future studies. This trade-off will be present in the case of the viscosity (and the stress tensor) as well. It will be even more serious, because the viscosity is even a more poorly converging quantity than the diffusion constant.

6.3 The effect of friction coefficient

We fix the dipole moment at $(p^*)^2 = 6$ and the time step at $\Delta t^* = 0.01$, and change the friction coefficient from $\gamma^* = 1{,}000$ to $10{,}000$. As discussed in the next section, realistic ER fluids have friction coefficients even larger than $10{,}000$, but we refer studying that regime to future publications.

As γ^* is increased, the curves tend to their equilibrium values as E_0 is switched on at a lower rate. Fitting exponential functions to these curves and identifying processes of different time lengths as parts of the complex process of chain formation will also be the subject of future studies.

The change in γ^* does not influence the value where the energy and the average chain length converge to. They converge to the same value but with a different rate. Changing friction, however, changes the diffusion constant. Fig. 13 shows the diffusion constant relative to its value in the absence of the field computed as $D^* = 1/\gamma^*$. The diffusion constant decreases to a smaller value relative to $D^*(E = 0)$ at smaller values of γ^*.

The average chain length shows that smaller γ^* results in a more wildly fluctuating system than a larger γ^*. The particles diffuse faster and produce larger variations in configurations during a given time period.

6.4 The effect of dipole moment

Our simulations show (Fig. 14) that the quantity that determines the structure of the ER fluid is the reduced dipole moment, namely, the relation of the dipole-dipole energy to the thermal energy unit, kT. Fig. 14 shows that these quantities converge to their equilibrium values exhibiting a similar trend.

The dipole moments studied in this work belong to the regime where the ER fluid considered as a collection of chains is still a fluid, namely, it does not solidify. Several papers in the literature study solidification of the ER chains [5, 6, 15, 16, 22].

7. Summary

In this work, we use a newly developed integrator algorithm to solve the Langevin equations and to perform BD simulations for ER fluids. Our focus was on the methodological development and identifying appropriate system parameters through which we can follow the dynamics of chain formation in the system.

The usefulness of computer simulations lies not only in the fact that we can follow the particles' trajectories, but also in the fact that we can gain a profound amount of information from these trajectories. In the BD simulations, for example, we can follow how the average number of chains of varying lengths changes in time. From that detailed information we can deduce time constants for characteristic processes during chain formation.

We intend to dig into those details in subsequent studies. Also, we want to examine the behavior of the chains under a stress.

Acknowledgement

This research was supported by the European Union, co-financed by the European Social Fund. EFOP-3.6.2-16-2017-00002. We also acknowledge the support of the National Research, Development and Innovation Office – NKFIH K124353.

REFERENCES

[1] Winslow, W.M.: Induced Fibration of Suspensions, *J. Appl. Phys.*, 1949, **20**(12), 1137–1140 DOI: 10.1063%2F1.1698285

[2] Duclos, T.G.; Carlson, J.D.; Chrzan, M.J.; Coulter, J.P.: Electrorheological Fluids — Materials and Applications, in Solid Mechanics and Its Applications (Springer Netherlands), 1992, **13**, 213–241 DOI: 10.1007%2F978-94-017-1903-2_5

[3] Havelka, K.O.; Filisko, F.E. (eds.): Progress in Electrorheology (Springer US), 1995. DOI: 10.1007%2F978-1-4899-1036-3

[4] Klingenberg, D.J.; van Swol, F.; Zukoski, C.F.: Dynamic simulation of electrorheological suspensions, *J. Chem. Phys.*, 1989, **91**(12), 7888–7895 DOI: 10.1063%2F1.457256

[5] Heyes, D.M.; Melrose, J.R.: Brownian Dynamics Simulations of Electro-Rheological Fluids, II: Scaling Laws, *Mol. Sim.*, 1990, **5**(5), 293–306 DOI: 10.1080%2F08927029008022415

[6] Whittle, M.: Computer simulation of an electrorheological fluid, *J. Non-Newtonian Fluid Mechanics*, 1990, **37**(2-3), 233–263 DOI: 10.1016%2F0377-0257%2890%2990007-x

[7] Klingenberg, D.J.; Zukoski, C.F.: Studies on the steady-shear behavior of electrorheological suspensions, *Langmuir*, 1990, **6**(1), 15–24 DOI: 10.1021%2Fla00091a003

[8] Jaggi, N.K.: Structure and dynamics of a dense dipolar system in an electric field and their relevance to electrorheological fluids, *J. Stat. Phys.*, 1991, **64**(5-6), 1093–1102 DOI: 10.1007%2Fbf01048816

[9] See, H.; Doi, M.: Aggregation Kinetics in Electro-Rheological Fluids, *J. Phys. Soc. Japan*, 1991, **60**(8), 2778–2782 DOI: 10.1143%2Fjpsj.60.2778

[10] Bonnecaze, R.T.; Brady, J.F.: Dynamic simulation of an electrorheological fluid, *J. Chem. Phys.*, 1992, **96**(3), 2183–2202 DOI: 10.1063%2F1.462070

[11] Toor, W.R.: Structure Formation in Electrorheological Fluids, *J. Colloid Interf. Sci.*, 1993, **156**(2), 335–349 DOI: 10.1006%2Fjcis.1993.1121

[12] Hass, K.C.: Computer simulations of nonequilibrium structure formation in electrorheological fluids, *Phy. Rev. E*, 1993, **47**(5), 3362–3373 DOI: 10.1103%2Fphysreve.47.3362

[13] Tao, R.; Jiang, Q.: Simulation of structure formation in an electrorheological fluid, *Phys. Rev. Lett.*, 1994, **73**(1), 205–208 DOI: 10.1103%2Fphysrevlett.73.205

[14] Tao, R.; Jiang, Q.: Simulation of Solid Structure Formation in an Electrorheological Fluid, *Int. J. Modern Phys. B*, 1994, **08**(20n21), 2721–2730 DOI: 10.1142%2Fs0217979294001081

[15] Baxter-Drayton, Y.; Brady, J.F.: Brownian electrorheological fluids as a model for flocculated dispersions, *J. Rheology*, 1996, **40**(6), 1027–1056 DOI: 10.1122%2F1.550772

[16] Gulley, G.L.; Tao, R.: Structures of an electrorheological fluid, *Phys. Rev. E*, 1997, **56**(4), 4328–4336 DOI: 10.1103%2Fphysreve.56.4328

[17] Jian, L.; Jiapeng, S.: Simulation of a three-dimensional electrorheological suspension, *J. Appl. Phys.*, 1996, **79**(9), 7312–7317 DOI: 10.1063%2F1.361447

[18] Wang, B.; Liu, Y.; Xiao, Z.: Dynamical modelling of the chain structure formation in electrorheological fluids, *Int. J. Eng. Sci.*, 2001, **39**(4), 453–475 DOI: 10.1016%2Fs0020-7225%2800%2900054-9

[19] Enomoto, Y.; Oba, K.: Simulation of structures and their rheological properties in electrorheological fluids, *Physica A*, 2002, **309**(1–2), 15–25 DOI: 10.1016%2Fs0378-4371%2802%2900599-x

[20] Climent, E.; Maxey, M.R.; Karniadakis, G.E.: Dynamics of Self-Assembled Chaining in Magnetorheological Fluids, *Langmuir*, 2004, **20**(2), 507–513 DOI: 10.1021%2Fla035540z

[21] Cao, J.G.; Huang, J.P.; Zhou, L.W.: Structure of Electrorheological Fluids under an Electric Field and a Shear Flow: Experiment and Computer Simulation, *J. Phys. Chem. B*, 2006, **110**(24), 11635–11639 DOI: 10.1021%2Fjp0611774

[22] Domínguez-García, P.; Melle, S.; Pastor, J.M.; Rubio, M.A.: Scaling in the aggregation dynamics of a magnetorheological fluid, *Phys. Rev. E*, 2007, **76**(5), 051403 DOI: 10.1103%2Fphysreve.76.051403

[23] Heyes, D.M.: Rheology of molecular liquids and concentrated suspensions by microscopic dynami-

cal simulations, *J. Non-Newton. Fluid*, 1988, **27**(1), 47–85 DOI: 10.1016%2F0377-0257%2888%2980004-1

[24] Grønbech-Jensen, N.; Farago, O.: A simple and effective Verlet-type algorithm for simulating Langevin dynamics, *Mol. Phys.*, 2013, **111**(8), 983–991 DOI: 10.1080%2F00268976.2012.760055

[25] Farago, O.: Langevin thermostat for robust configurational and kinetic sampling, *Physica A*, 2019, **534**, 122210 DOI: 10.1016%2Fj.physa.2019.122210

[26] Jensen, L.F.G.; Grønbech-Jensen, N.: Accurate configurational and kinetic statistics in discrete-time Langevin systems, *Mol. Phys.*, 2019, **117**(18), 2511–2526 DOI: 10.1080%2F00268976.2019.1570369

[27] Sherman, S.G.; Paley, D.A.; Wereley, N.M.: Massively Parallel Simulations of Chain Formation and Restructuring Dynamics in a Magnetorheological Fluid, in ASME 2011 Conference on Smart Materials, Adaptive Structures and Intelligent Systems, Volume 1 (ASMEDC), DOI: 10.1115%2Fsmasis2011-5188

[28] Sherman, S.G.; Paley, D.A.; Wereley, N.M.: Parallel Simulation of Transient Magnetorheological Direct Shear Flows Using Millions of Particles, *IEEE Transactions on Magnetics*, 2012, **48**(11), 3517–3520 DOI: 10.1109%2Ftmag.2012.2201214

[29] Fernández-Toledano, J.C.; Ruiz-López, J.A.; Hidalgo-Álvarez, R.; de Vicente, J.: Simulations of polydisperse magnetorheological fluids: A structural and kinetic investigation, *J. Rheology*, 2015, **59**(2), 475–498 DOI: 10.1122%2F1.4906544

[30] Boda, D.; Valiskó, M.; Szalai, I.: The origin of the interparticle potential of electrorheological fluids, *Cond. Matt. Phys.*, 2013, **16**(4), 43002 DOI: 10.5488/cmp.16.43002

[31] Jackson, J.D.: Classical Electrodynamics (Wiley, New York), 3rd edn., 1999. ISBN: 978-0471309321

[32] Boda, D.; Gillespie, D.; Nonner, W.; Henderson, D.; Eisenberg, B.: Computing induced charges in inhomogeneous dielectric media: Application in a Monte Carlo simulation of complex ionic systems, *Phys. Rev. E*, 2004, **69**(4), 046702 DOI: 10.1103/physreve.69.046702

[33] Lemons, D.S.; Gythiel, A.: Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)], *Am. J. Phys.*, 1997, **65**(11), 1079–1081 DOI: 10.1119%2F1.18725

[34] Schneider, T.; Stoll, E.: Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions, *Phys. Rev. B*, 1978, **17**(3), 1302–1322 DOI: 10.1103%2Fphysrevb.17.1302

[35] van Gunsteren, W.; Berendsen, H.: Algorithms for brownian dynamics, *Mol. Phys.*, 1982, **45**(3), 637–647 DOI: 10.1080%2F00268978200100491

[36] Brünger, A.; Brooks, C.L.; Karplus, M.: Stochastic boundary conditions for molecular dynamics simulations of ST2 water, *Chem. Phys. Lett.*, 1984, **105**(5), 495–500 DOI: 10.1016%2F0009-2614%2884%2980098-6

[37] Leimkuhler, B.; Matthews, C.: Rational Construction of Stochastic Numerical Methods for Molecular Sampling, *Appl. Math. Res. eXpress*, 2013, **2013**(1), 34–56 DOI: 10.1093%2Famrx%2Fabs010

[38] Horváth, B.; Szalai, I.: Structure of electrorheological fluids: A dielectric study of chain formation, *Phys. Rev. E*, 2012, **86**(6),061403 DOI: 10.1103%2Fphysreve.86.061403

[39] Horváth, B.; Szalai, I.: Dynamic dielectric response of electrorheological fluids in drag flow, *Phys. Rev. E*, 2015, **92**(4), 042308 DOI: 10.1103%2Fphysreve.92.042308

[40] Berti, C.; Furini, S.; Gillespie, D.; Boda, D.; Eisenberg, R.S.; Sangiorgi, E.; Fiegna, C.: A 3-D Brownian Dynamics simulator for the study of ion permeation through membrane pores, *J. Chem. Theor. Comput.*, 2014, **10**(8), 2911–2926 DOI: 10.1021/ct4011008

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 109–115 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-16

LOCALIZATION ACCURACY IMPROVEMENT OF AUTONOMOUS VEHICLES USING SENSOR FUSION AND EXTENDED KALMAN FILTER

ISTVÁN SZALAY*[1], KRISZTIÁN ENISZ[1], HUNOR MEDVE[1], AND DÉNES FODOR[1]

[1]Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

Advanced driver assistance systems and autonomous vehicles rely heavily on position information, therefore, enhancing localization algorithms is an actively researched field. Novel algorithms fuse the signals of common vehicle sensors, the inertial measurement unit and global positioning system. This paper presents a localization algorithm for vehicle position estimation that integrates vehicle sensors (steering angle encoder, wheel speed sensors and a yaw-rate sensor) and GPS signals. The estimation algorithm uses an extended Kalman filter designed for a simplified version of the single track model. The vehicle dynamics-based model only includes calculation of the lateral force and planar motion of the vehicle resulting in the minimal state-space model and filter algorithm. A TESIS veDYNA vehicle dynamics and MathWorks Simulink-based simulation environment was used in the development and validation process. The presented results include different low- and high-speed maneuvers as well as filter estimates of the position and heading of the vehicle.

Keywords: vehicle localization, extended Kalman filter, sensor fusion, dead reckoning

1. Introduction

Vehicle navigation systems are important components of autonomous driving solutions. These systems acquire the position, velocity and heading of the vehicle by using onboard or externally installed sensors such as wheel speed sensors, gyroscopes, accelerometers, inertial navigation systems (INS), compasses, radio frequency receivers, etc. [1].

The two most common vehicle localization techniques are dead reckoning and the use of a Global Navigation Satellite System (GNSS), like the Global Positioning System (GPS). In dead reckoning, distance and heading sensors are used to measure the vehicle displacement vector which is then integrated recursively to determine the current position of the vehicle. Measurement errors are accumulated by this integration, therefore, the accuracy of the position estimation is constantly decreasing. On the other hand, GPS provides absolute vehicle positions without the accumulation of errors associated with dead reckoning through the use of satellites as the reference points.

Localization methods typically integrate GPS with other sensors since GPS suffers from outages and errors. Many papers have integrated GPS with INS [2–5]. Others have integrated GPS with the inertial measurement unit (IMU) [6, 7]. This paper integrates GPS with vehicle sensors, similarly to Refs. [8–10]. The continuous availability of signals from vehicle sensors required by dead reckoning as well as the absolute positioning accuracy of GPS render them combinable to achieve better performance [8, 9].

In this paper, a vehicle localization algorithm is presented that uses an extended Kalman filter (EKF) to integrate dead reckoning with GPS. Dead reckoning is based on a simplified version of the Single Track Model (STM) and uses a steering angle encoder, wheel speed sensors and yaw rate measurements.

2. Modeling Strategy

The aim of the localization algorithm is to estimate the current position (x and y coordinates) and heading (yaw angle ψ) of the vehicle by fusing GPS measurements with dead reckoning based on vehicle sensor signals. GPS measurements provide the noisy x_{gps} and y_{gps} coordinates. Vehicle sensors provide the steering angle δ, the four wheel speed signals ω_i and the yaw rate $\dot{\psi}$. These measurements are available in most commercial vehicles. The lateral acceleration sensors were not used because the acceleration signal is usually less reliable and noisier than the other vehicle signals.

The localization algorithm fuses these signals using an extended Kalman filter which requires an appropriate system model. The system model connects the available and estimated signals with inputs, states and outputs of the system. The simplest and most practical solution is to

*Correspondence: szalay.istvan@mk.uni-pannon.hu

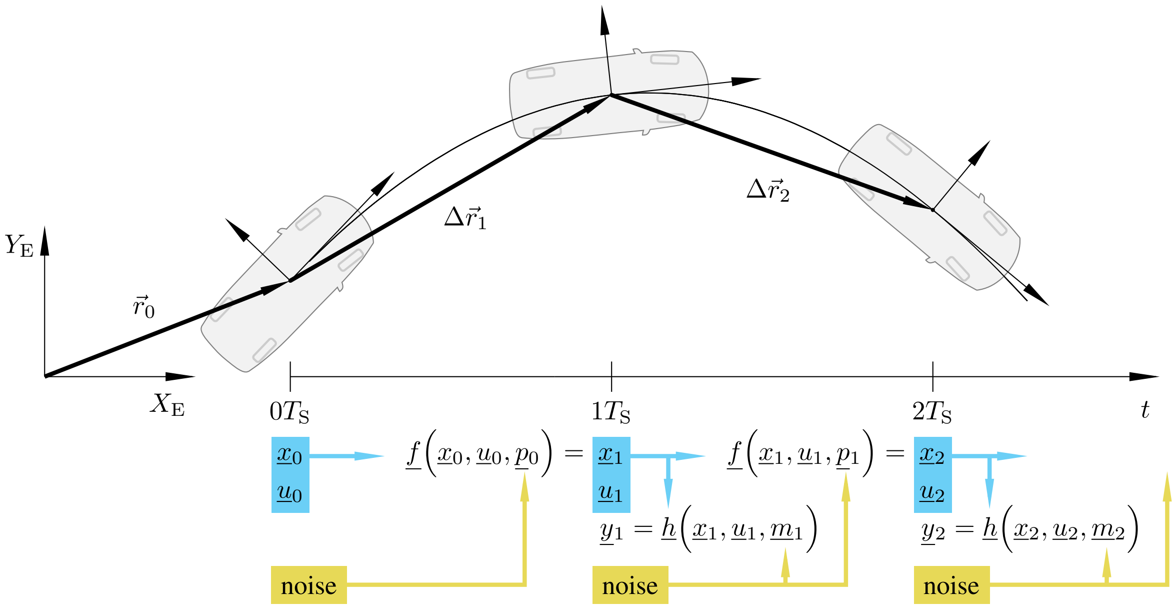

Y_{E}
X_{E}
$\vec{r}_0$
$\Delta\vec{r}_1$
$\Delta\vec{r}_2$
$0T_{\mathrm{S}}$
$1T_{\mathrm{S}}$
$2T_{\mathrm{S}}$
t
$\underline{x}_0$
$\underline{u}_0$
$\underline{f}\left(\underline{x}_0, \underline{u}_0, \underline{p}_0\right) = \underline{x}_1$
$\underline{u}_1$
$\underline{f}\left(\underline{x}_1, \underline{u}_1, \underline{p}_1\right) = \underline{x}_2$
$\underline{u}_2$
$\underline{y}_1 = \underline{h}\left(\underline{x}_1, \underline{u}_1, \underline{m}_1\right)$
$\underline{y}_2 = \underline{h}\left(\underline{x}_2, \underline{u}_2, \underline{m}_2\right)$
noise
noise
noise

Figure 1: Dead reckoning steps and propagation of the system state

formulate the system model in a way that produces the true-position and heading state variables, the input variables of vehicle sensor signals and the outputs of GPS signals.

3. Dead Reckoning Algorithm

Dead reckoning or path integration is the process of estimating the current position of a vehicle by using a previously determined position, and projecting that position based upon known or estimated speeds over a time period and course that has elapsed. Dead reckoning is subject to cumulative errors. In continuous time, dead reckoning results in the integration of the velocity vector or double integration of the acceleration vector with respect to time. In discrete time of sample time T_{S}, the equivalent of integration can be written as the following in vector-sum form:

$$\vec{r}_k = \vec{r}_{k-1} + \Delta\vec{r}_k = \vec{r}_0 + \sum_{i=1}^{k} \Delta\vec{r}_i \qquad (1)$$

The equivalent coordinate form of Eq. 1 for a vehicle in planar motion is

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \begin{bmatrix} \Delta x_k \\ \Delta y_k \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \sum_{i=1}^{k} \begin{bmatrix} \Delta x_i \\ \Delta y_i \end{bmatrix} \qquad (2)$$

In the vector form, $\vec{r}_k$ denotes the position of the vehicle's center of gravity (COG) at $t_k = kT_{\mathrm{S}}$ and $\Delta\vec{r}_k$ stands for the displacement between $t_{k-1} = (k-1)\,T_{\mathrm{S}}$ and $t_k = kT_{\mathrm{S}}$. The initial position in vector form is $\vec{r}_0$ and in coordinate form is $[x_0, y_0]^T$ (see Fig. 1).

Dead reckoning is performed in the Earth-Fixed coordinate system $X_{\mathrm{E}} - Y_{\mathrm{E}}$. Using the vehicle sensors, displacements in the vehicle axis system $X_{\mathrm{V}} - Y_{\mathrm{V}}$ can be calculated. To rotate the displacement into the Earth-Fixed coordinate system, changes in the heading or yaw angle ψ of the vehicle have to be tracked. The yaw angle can be estimated by integrating the yaw-rate signal $\dot{\psi}$ as

$$\psi_k = \psi_{k-1} + \dot{\psi}_k T_{\mathrm{S}} = \psi_0 + \sum_{i=1}^{k-1} \dot{\psi}_i T_{\mathrm{S}} \qquad (3)$$

The displacements $\Delta x_{V,k}$ and $\Delta y_{V,k}$ have to be calculated in the vehicle axis system $X_{\mathrm{V}} - Y_{\mathrm{V}}$. The displacements from the vehicle axis system to the Earth-Fixed coordinate system are transformed using a two-dimensional rotation matrix $\underline{\underline{R}}(\psi_{k-1})$ that rotates around the vertical axis Z_{E} by ψ_{k-1}.

$$\begin{bmatrix} \Delta x_k \\ \Delta y_k \end{bmatrix} = \underbrace{\begin{bmatrix} \cos\psi_{k-1} & -\sin\psi_{k-1} \\ \sin\psi_{k-1} & \cos\psi_{k-1} \end{bmatrix}}_{\underline{\underline{R}}(\psi_{k-1})} \begin{bmatrix} \Delta x_{V,k} \\ \Delta y_{V,k} \end{bmatrix} \qquad (4)$$

The simplest estimation of the longitudinal displacement $\Delta x_{V,k}$ involves the estimation of the longitudinal speed v_k as the average of the wheel speed signals divided by the effective radius of the tires R:

$$\Delta x_{V,k} = v_{k-1} T_{\mathrm{S}}, \quad \text{where} \quad v_{k-1} = \frac{R}{4} \sum_{i=1}^{4} \omega_{k-1,i} \qquad (5)$$

The lateral displacement Δy_{V} is estimated by the second integral of the lateral acceleration, assuming an initial lateral velocity of zero:

$$\Delta y_{V,k} = \frac{1}{2} a_{y,k-1} T_{\mathrm{S}}^2. \qquad (6)$$

The lateral acceleration a_y can be in the form of a sensor signal from an accelerometer or calculated based

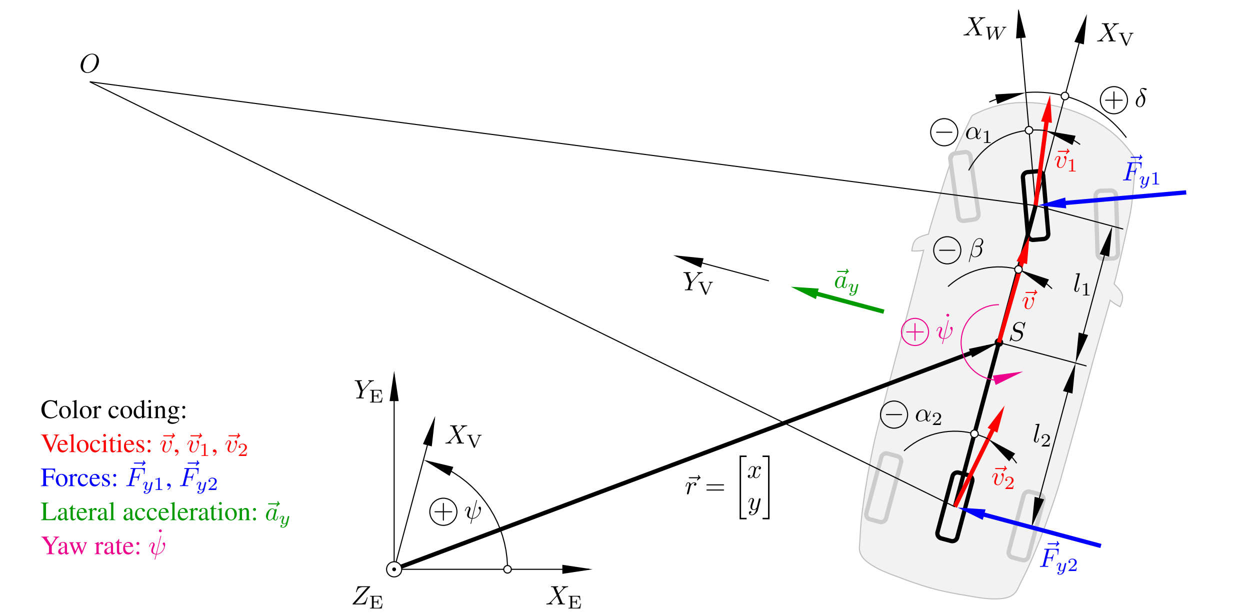


Figure 2: The physical quantities of the single track vehicle model following [11]

on the vehicle dynamics model and other vehicle sensor signals. Usually, the accelerometer signals are less reliable and noisier than those of the steering angle, velocity and yaw-rate sensors, therefore, a_y was calculated based on a vehicle model and signals from vehicle sensors, in a similar way to Ref. [9].

4. Lateral Vehicle Dynamics

As a basis for modeling the lateral vehicle dynamics, the well-known single track vehicle model defined by Riekert and Schunck [12] was used. Fig. 2 shows the physical quantities related to the single track vehicle model. This paper uses the sign conventions defined in Ref. [11].

The classical single track vehicle model includes several important simplifications. It only describes lateral motion and rotation around the vertical axis, while neglects vertical dynamics as well as rolling and pitching. Furthermore, the equations of motion of the single track vehicle model are linearized.

To define the lateral acceleration as a function of the steering angle δ, longitudinal velocity v and yaw rate $\dot{\psi}$, the lateral force equation of the model can be used:

$$ma_y = F_{y1}(\alpha_1) + F_{y2}(\alpha_1) \tag{7}$$

The linearized tire forces as the products of the cornering stiffnesses (c_1 and c_2) and tire slip angles (α_1 and α_2) are calculated as follows:

$$F_{y1}(\alpha_1) = -c_1\alpha_1 \quad \text{and} \quad F_{y2}(\alpha_2) = -c_2\alpha_2. \tag{8}$$

Tire slip angles (α_1 and α_2) are defined by the velocity triangles of the two axles in the following nonlinear forms (see Fig. 3):

$$\tan(\alpha_1 + \delta) = \frac{v\sin\beta + \dot{\psi}l_1}{v\cos\beta} \tag{9}$$

$$\tan\alpha_2 = \frac{v\sin\beta - \dot{\psi}l_2}{v\cos\beta} \tag{10}$$

Eq. 9 corresponds to the yellow part of the velocity triangle corresponding to the front axle and Eq. 10 to the rear axle.

The tire slip angle equations are linearized and the sideslip angle β omitted:

$$\alpha_1 \approx \beta + \frac{\dot{\psi}l_1}{v} - \delta \approx \frac{\dot{\psi}l_1}{v} - \delta \tag{11}$$

$$\alpha_2 \approx \beta - \frac{\dot{\psi}l_2}{v} \approx -\frac{\dot{\psi}l_2}{v} \tag{12}$$

The lateral acceleration a_y depends on the values of the vehicle sensors according to

$$a_y \approx -\frac{c_1}{m}\left(\frac{\dot{\psi}l_1}{v} - \delta\right) + \frac{c_2}{m}\frac{\dot{\psi}l_2}{v} \tag{13}$$

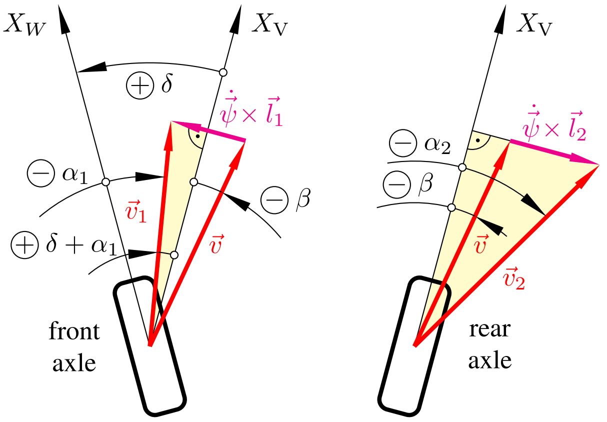


Figure 3: Velocity triangles in the single track vehicle model

In this way, the lateral acceleration depends on the usually available cornering stiffnesses (c_1 and c_2), the mass and length of the vehicle (m, l_1 and l_2), and the signals of the vehicle sensors. If available, the lateral acceleration signal from an accelerometer can be incorporated into the sensor fusion algorithm.

5. Extended Kalman Filter Design

The application of an extended Kalman filter requires a stochastic mathematical model of the system in a state-space representation including the statistical properties of the process and measurement noises [13, 14].

5.1 State-Space Model

The state-space representation includes both the state and measurement equations in vector form by introducing the functions $\underline{f}$ and $\underline{h}$, the known input vector $\underline{u}$, process-noise vector $\underline{p}$ and measurement-noise vector $\underline{m}$:

$$\underline{x}_k = \underline{f}\Big(\underline{x}_{k-1}, \underline{u}_{k-1}, \underline{p}_{k-1}\Big) \tag{14}$$

$$\underline{y}_k = \underline{h}\Big(\underline{x}_k, \underline{m}_k\Big) \tag{15}$$

The system is time invariant, therefore, $\underline{f}$ and $\underline{h}$ are not indexed. The discrete-time indexing of the difference equations is illustrated in Fig. 1.

To estimate the vehicle's position, the dead reckoning Eqs. 2–6 are used augmented to include the lateral acceleration (Eq. 13) as a computed value. The resulting state equation of the discrete-time state-space model is Eq. 18.

By examining Eq. 18, it can be concluded that the system is nonlinear and, therefore, a linear Kalman filter is unsuitable. In this paper, an extended Kalman filter is used.

The output vector $\underline{y}$ of the system includes the GPS measurements that consist of the true position and measurement noise $\underline{m}$ defined by the output function $\underline{h}$:

$$\underline{y}_k = \begin{bmatrix} x_{\mathrm{gps},k} \\ y_{\mathrm{gps},k} \end{bmatrix} = \begin{bmatrix} x_k + m_{x,k} \\ y_k + m_{y,k} \end{bmatrix} = \underline{h}\Big(\underline{x}_k, \underline{m}_k\Big) \tag{16}$$

5.2 Jacobians

The extended Kalman filter requires the Jacobians of both the state and measurement functions with respect to the state vector:

$$\underline{\underline{F}}_k = \left.\frac{\partial \underline{f}}{\partial \underline{x}}\right|_{\underline{\hat{x}}^+_{k-1}} \text{ and } \underline{\underline{H}} = \left.\frac{\partial \underline{h}}{\partial \underline{x}}\right|_{\underline{\hat{x}}^+_{k-1}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{17}$$

The elements of $\underline{\underline{H}}$ are constant, while the elements of $\underline{\underline{F}}_k$ depend on the discrete time k, as defined by Eq. 19.

The sample time T_S was $100\,\mathrm{ms}$, which was sufficiently small to capture the vehicle's movement but not too small for GPS sampling.

$$\underline{x}_k = \begin{bmatrix} x_k \\ y_k \\ \psi_k \end{bmatrix} = \begin{bmatrix} x_{k-1} + v_{k-1} T_\mathrm{S} \cos\left(\psi_{k-1}\right) - \dfrac{1}{2} a_{y,k-1} T_\mathrm{S}^2 \sin\left(\psi_{k-1}\right) + p_{x,k-1} \\ y_{k-1} + v_{k-1} T_\mathrm{S} \sin\left(\psi_{k-1}\right) + \dfrac{1}{2} a_{y,k-1} T_\mathrm{S}^2 \cos\left(\psi_{k-1}\right) + p_{y,k-1} \\ \psi_{k-1} + \dot{\psi}_{k-1} T_\mathrm{S} + p_{\psi,k-1} \end{bmatrix} = \underline{f}\Big(\underline{x}_{k-1}, \underline{u}_{k-1}, \underline{p}_{k-1}\Big) \tag{18}$$

$$\underline{\underline{F}}_k = \begin{bmatrix} 1 & 0 & -\sin\left(\psi_{k-1}\right) v_{k-1} T_\mathrm{S} - \dfrac{1}{2}\cos\left(\psi_{k-1}\right) a_{y,k-1} T_\mathrm{S}^2 \\ 0 & 1 & \cos\left(\psi_{k-1}\right) v_{k-1} T_\mathrm{S} - \dfrac{1}{2}\sin\left(\psi_{k-1}\right) a_{y,k-1} T_\mathrm{S}^2 \\ 0 & 0 & 1 \end{bmatrix}, \quad \underline{u}_k = \begin{bmatrix} a_{y,k} \\ v_k \\ \dot{\psi}_k \end{bmatrix} \tag{19}$$

5.3 Noise model

During the design of the extended Kalman filter, a time-invariant and normally distributed process as well as measurement noise were assumed:

$$\underline{p} \sim \mathcal{N}\Big(\underline{0}, \underline{\underline{Q}}\Big), \quad \underline{m} \sim \mathcal{N}\Big(\underline{0}, \underline{\underline{R}}\Big) \tag{20}$$

For the development and testing of our extended Kalman filter algorithm, a TESIS veDYNA vehicle dynamics-based simulation environment was used with configurable noise models (see Section 6). Although it was possible to match the noise and covariances of the filter perfectly, the filter and noise models were tuned in a slightly different way. The measurement- and process-noise covariances for both the filter and noise models have been estimated based on the characteristics of the sensor and GPS range error statistics [15]:

$$\underline{\underline{Q}} = \mathrm{diag}\Big(0.2\,\mathrm{m}^2, 0.2\,\mathrm{m}^2, 0.01\,\mathrm{rad}^2\Big) \tag{21}$$

$$\underline{\underline{R}} = \mathrm{diag}\Big(1.5\,\mathrm{m}^2, 1.5\,\mathrm{m}^2\Big) \tag{22}$$

Our algorithm performs the dead reckoning in the Earth-Fixed coordinate system and, therefore, no predetermined difference between the direction covariances X_E and Y_E is present, so $Q_{11} = Q_{22}$ and $R_{11} = R_{22}$. The displacement variance of $0.2\,\mathrm{m}^2$ corresponds to an error of $4.5\,\mathrm{m\,s}^{-1}$ in the wheel speed-based velocity calculation. In the presence of non-normally distributed noise, the filter is not optimal.

5.4 The extended Kalman Filter Algorithm

The extended Kalman filter algorithm follows well-known steps [14]. The prediction equations are

$$\hat{\underline{x}}_k^- = \underline{f}\left(\hat{\underline{x}}_{k-1}^+, \underline{u}_{k-1}, \underline{0}\right) \tag{23}$$

$$\underline{\underline{P}}_k^- = \underline{\underline{F}}_k \underline{\underline{P}}_{k-1}^+ \underline{\underline{F}}_k^T + \underline{\underline{Q}} \tag{24}$$

Calculation of the Kalman gain is

$$\underline{\underline{K}}_k = \underline{\underline{P}}_k^- \underline{\underline{H}}_k^T \left(\underline{\underline{H}}_k \underline{\underline{P}}_k^- \underline{\underline{H}}_k^T + \underline{\underline{R}}\right)^{-1} \tag{25}$$

The correction equations are

$$\hat{\underline{x}}_k^+ = \hat{\underline{x}}_k^- + \underline{\underline{K}}_k \left(\underline{y}_k - \underline{h}_k\left(\hat{\underline{x}}_k^-, \underline{u}_k, \underline{0}\right)\right) \tag{26}$$

$$\underline{\underline{P}}_k^+ = \left(\underline{\underline{I}} - \underline{\underline{K}}_k \underline{\underline{H}}_k\right) \underline{\underline{P}}_k^- \tag{27}$$

The initial state is determined based on the first several GPS observations whilst in motion.

6. Simulation Environment

The development of the localization algorithm was conducted in a MathWorks MATLAB/Simulink-based vehicle dynamics simulation environment. The TESIS veDYNA model library was used to model the vehicle, road system and maneuvers (Fig. 4).

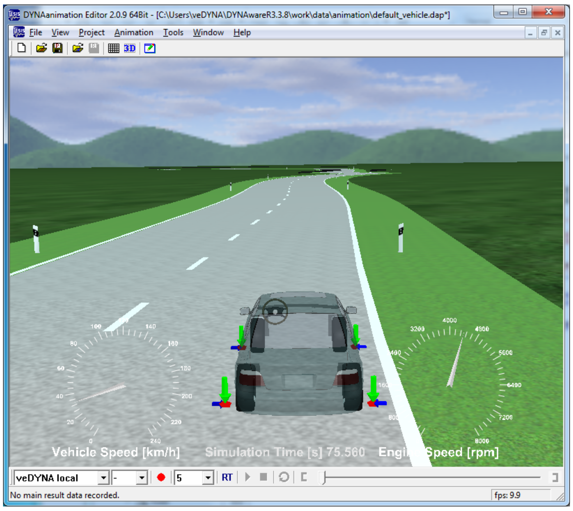
The advantage of our simulation platform lies in its integrated structure. Algorithms developed in MATLAB and Simulink can be integrated into the TESIS veDYNA vehicle dynamics simulation which enables Software-in-the-Loop testing and real-time Hardware-in-the-Loop analysis to be implemented by the National Instruments PXI hardware system.

In our platform, real streets using GPS coordinates or map databases can be modeled. This enables the direct comparison of simulations and real-world measurements.

7. Simulation Results

The extended Kalman filter was tuned and studied in simulated test maneuvers during which data were collected. The test maneuvers were performed on a virtual test track that included low- and high-speed driving, accelerating, braking and cornering. The size of the test track was around 2 by 2 km and the simulated test maneuvers lasted for 528 s.

Our extended Kalman filter was compared to simple dead reckoning and GPS observations. The true path, the result of dead reckoning and the position, estimated by the implementation of our extended Kalman filter, are shown in Fig. 5.



DYNAanimation Editor 2.0.9 64Bit - [C:\Users\veDYNA\DYNAwareR3.3.8\work\data\animation\default_vehicle.dap*]

Simulation Time [s] 75.560

No main result data recorded.

fps: 9.9

Figure 4: TESIS DYNAanimation user interface

7.1 Comparison with Dead Reckoning

Dead reckoning without the aid of the extended Kalman filter accumulated a large position error during the test maneuvers. The error of the dead reckoning reached 100 m, as can be seen in Fig. 5. The position estimated by our extended Kalman filter did not include error accumulation and performed better than dead reckoning alone.

7.2 Comparison with GPS

Due to the large difference between the size of the path and the position errors, the difference between GPS observations and the extended Kalman filter estimates were not as clearly visible as the error of dead reckoning (Fig. 5). Zooming in on certain parts of the path enables visual evaluation of a particular part of the path

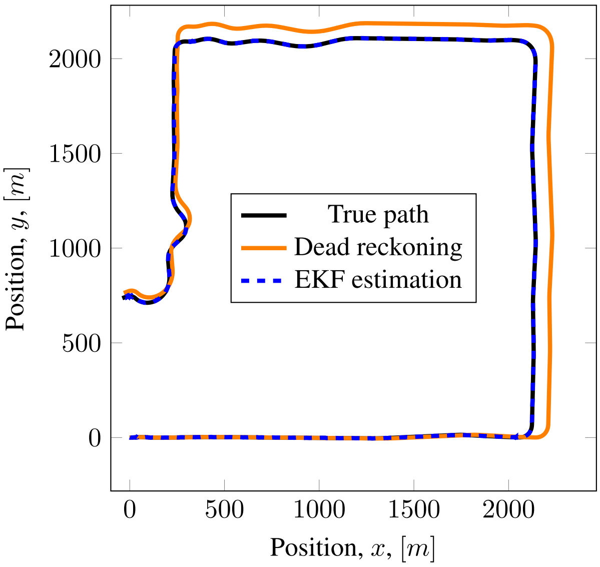


Figure 5: The real path, dead reckoning only and EKF estimation

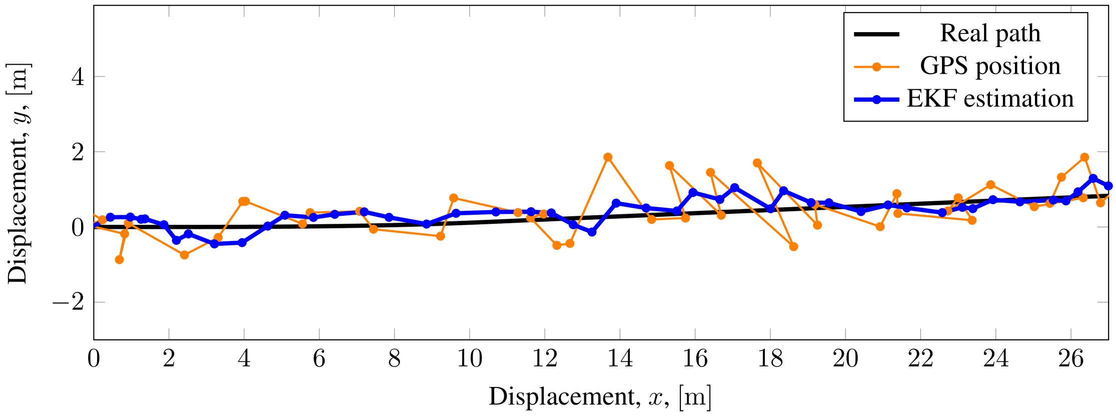


Figure 6: The initial section of the true, estimated and GPS-observed paths

(Fig. 6). To evaluate the extended Kalman filter estimate and make it numerically comparable to the GPS observations, the instantaneous and average distance errors were calculated.

The distance error d_k of the estimation is calculated as the distance between the estimated and true positions:

$$d_k = \sqrt{(\hat{x}_k - x_k)^2 + (\hat{y}_k - y_k)^2} \tag{28}$$

The distance error $d_{\text{gps},k}$ of GPS observations is

$$d_{\text{gps},k} = \sqrt{\left(x_{\text{gps}} - x\right)^2 + \left(y_{\text{gps}} - y\right)^2} \tag{29}$$

The average distance errors are

$$\bar{d} = \frac{1}{N}\sum_{k=1}^{N} d_k \quad \text{and} \quad \bar{d}_{\text{gps}} = \frac{1}{N}\sum_{k=1}^{N} d_{\text{gps},k} \tag{30}$$

The distance errors of the GPS observations and the extended Kalman filter position estimates are shown in Fig. 7. The timespan of the test maneuvers was 528 s, moreover, at a sample time of 100 ms, it produced 5280 points. The average distance error of the GPS observations was $\bar{d}_{\text{gps}} = 751\,\text{mm}$. Our extended Kalman filter reduced the average distance error significantly, $\bar{d} = 457\,\text{mm}$.

Besides the average values, the distributions of the distance errors are also of interest. The distance errors are discrete values whose distributions can be approximated by their histograms. Such histograms are shown in (Fig. 8) with a bin width of 10 cm.

8. Conclusions

In this paper, a real-time vehicle localization algorithm developed and implemented in a TESIS veDYNA vehicle dynamics simulation environment was presented. The localization algorithm used an extended Kalman filter to fuse GPS observations with vehicle sensor measurements of only the steering angle, yaw rate and wheel speeds. The underlying state-space model is based on planar dead reckoning that calculates the longitudinal displacement using wheel speed and lateral displacement in a simplified version of the single track model. The state-space model only includes the x and y coordinates and the yaw angle ψ to minimize the model and filter algorithm.

The performance of the position estimator was analyzed during different high- and low-speed maneuvers. Compared to the dead reckoning and GPS observations that were not integrated, the integrated system performed significantly better. The average distance error was reduced by 39 %.

Further improvement of localization accuracy would be possible by using a more sophisticated vehicle model resulting in a more complex implementation of an extended Kalman filter with a much higher computational burden.

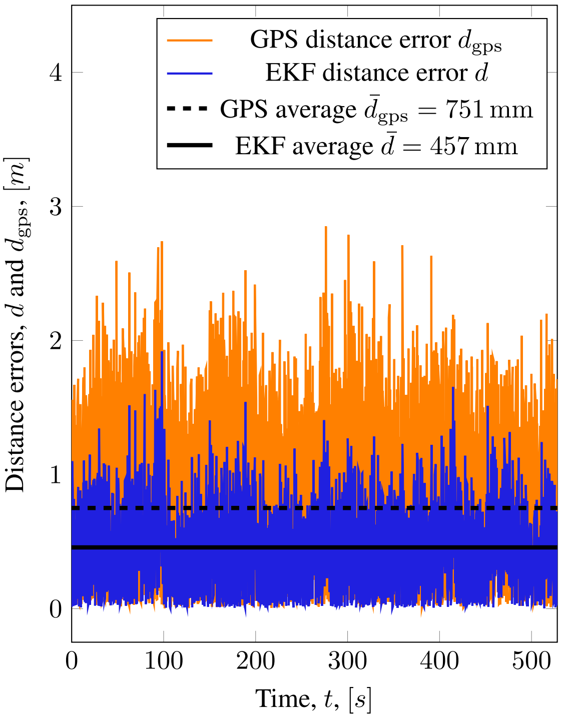


Figure 7: Distance error comparison

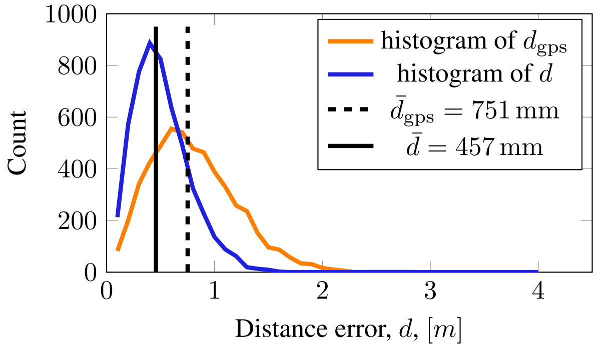


Figure 8: Distance error histograms

Acknowledgement

This research was supported by the project EFOP-3.6.2-16-2017-00002 entitled "Research of Autonomous Vehicle Systems related to the Autonomous Vehicle Proving Ground of Zalaegerszeg".

REFERENCES

[1] Karlsson, R.; Gustafsson, F.: The Future of Automotive Localization Algorithms: Available, reliable, and scalable localization: Anywhere and anytime, *IEEE Signal Processing Magazine*, 2017, **34**(2), 60–69 DOI: 10.1109/MSP.2016.2637418

[2] Farrell, J.A.; Givargis, T.D.; Barth, M.J.: Real-time differential carrier phase GPS-aided INS, *IEEE Transactions on Control Systems Technology*, 2000, **8**(4), 709–721 DOI: 10.1109/87.852915

[3] Qi, H.; Moore, J.B.: Direct Kalman filtering approach for GPS/INS integration, *IEEE Transactions on Aerospace and Electronic Systems*, 2002, **38**(2), 687–693 DOI: 10.1109/TAES.2002.1008998

[4] Yu, M.: INS/GPS Integration System using Adaptive Filter for Estimating Measurement Noise Variance, *IEEE Transactions on Aerospace and Electronic Systems*, 2012, **48**(2), 1786–1792 DOI: 10.1109/TAES.2012.6178100

[5] Liu, H.; Nassar, S.; El-Sheimy, N.: Two-Filter Smoothing for Accurate INS/GPS Land-Vehicle Navigation in Urban Centers, *IEEE Transactions on Vehicular Technology*, 2010, **59**(9), 4256–4267 DOI: 10.1109/TVT.2010.2070850

[6] Almeida, H.P.; Júnior, C.L.N.; d. Santos, D.S.; Leles, M.C.R.: Autonomous Navigation of a Small-Scale Ground Vehicle Using Low-Cost IMU/GPS Integration for Outdoor Applications, in 2019 IEEE International Systems Conference (SysCon), 1–8 DOI: 10.1109/SYSCON.2019.8836794

[7] Yu, M.; Guo, H.; Gao, W.: Realization of Low-Cost IMU/GPS Integrated Navigation System, in 2006 Japan-China Joint Workshop on Frontier of Computer Science and Technology, 189–195 DOI: 10.1109/FCST.2006.27

[8] Kao, W.W.: Integration of GPS and dead-reckoning navigation systems, in Vehicle Navigation and Information Systems Conference, 1991, vol. 2, 635–643 DOI: 10.1109/VNIS.1991.205808

[9] Rezaei, S.; Sengupta, R.: Kalman Filter-Based Integration of DGPS and Vehicle Sensors for Localization, *IEEE Transactions on Control Systems Technology*, 2007, **15**(6), 1080–1088 DOI: 10.1109/TCST.2006.886439

[10] Hohman, D.; Murdock, T.; Westerfield, E.; Hattox, T.; Kusterer, T.: GPS roadside integrated precision positioning system, in IEEE 2000. Position Location and Navigation Symposium (Cat. No.00CH37062), 221–230 DOI: 10.1109/PLANS.2000.838306

[11] ISO 8855:2011 Road vehicles – Vehicle dynamics and road-holding ability – Vocabulary, Standard, International Organization for Standardization, Geneva, CH, 2011

[12] Riekert, P.; Schunck, T.E.: Zur Fahrmechanik des gummibereiften Kraftfahrzeugs, *Ingenieur-Archiv*, 1940, **11**(3), 210–224 DOI: 10.1007/BF02086921

[13] Bucy, R.S.: Linear and nonlinear filtering, *Proceedings of the IEEE*, 1970, **58**(6), 854–864 DOI: 10.1109/PROC.1970.7792

[14] Simon, D.: Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches (Wiley-Interscience, New York, NY, USA), 2006 ISBN: 0471708585

[15] FAA William J. Hughes Technical Center: Global positioning system (GPS) standard positioning service (SPS) performance analysis report, 2019 https://www.nstb.tc.faa.gov/reports/

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 117–121 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-17

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
HJIC

GEOSPATIAL DATA MANIPULATION SUPPORTING THE IMPLEMENTATION OF DRIVING SIMULATION ENVIRONMENTS

FERENC SPEISER*[1], KRISZTIÁN ENISZ[1], AND DÉNES FODOR[1]

[1] Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

A lot of highly detailed geospatial information (obtained by mobile mapping and spatial data processing) is available that can be used to describe the exact road parameters for simulation and modeling. A gap between the freely available geospatial information and the descriptive standards of the road is present that is used in driving/traffic simulation as well as the systems of test vehicles. The toolset of GIS (Geographic Information Systems) provides wide-ranging functionality for spatial data processing, but is yet to offer support for standard formats of road description (OpenDRIVE data). This paper describes a method for gathering and converting road-network information from OpenStreetMap to OpenDRIVE data format. Using a similar conversion tool, the scenario generation and synthesis of realistic road networks for driving simulator applications can be more convenient and faster.

Keywords: OpenDRIVE, OpenStreetMap, road-network conversion

1. Introduction

Location information plays an important role in our everyday lives. To go somewhere, e.g. shopping, our exact position and destination is required. This location information can be obtained by GPS or a cellular network even with a high degree of accuracy, depending on the application area. However, other data (map, road network, terrain model) may be required to reach the destination. Detailed route planning can be achieved by these collected data (spatial databases) and has a great degree of significance in driver assistance.

The reliability of positioning and location awareness plays a crucial role in the field of autonomous vehicles and communication between them. It is important to validate the entire functionality of vehicles in all possible circumstances that could occur in real life during the development of all systems, subsystems and sensors of the Advanced Driver-Assistance Systems (ADAS) [1]. This kind of testing is expensive and requires a lot of resources to provide real-life conditions, e.g. the establisment of a proving ground like ZalaZone in Hungary.

During the first development stages of a function, it is appropriate to test basic functionality in a simulation environment. This environment can simulate all the information that is needed by the sensors of the control unit that runs the test build of the development. Of course, the more detailed the simulation environment, the more accurate the validated test result can be. This means that if a simulation environment were to be established, which can provide data in exactly the same mode as would occur in real life, the number of real-life tests necessary during the development of a functionality could decrease. Therefore, the rate of product development would accelerate and development costs decrease [2].

How can a simulation environment emulate real-life signals? Software-in-the-Loop (SIL), Hardware-in-the-Loop (HIL) and Model-in-the-Loop (MIL) solutions are able to emulate the signal level of real environments. All the necessary signals are obtained by the software, hardware or model components in the simulation as would be obtained in real life. The location information is a parameter that can also be loaded for the software, hardware or model component that is being tested.

What kind of location data can be used for testing, and how can the route, road as well as terrain on which the vehicle drives be described? A huge amount of location data is available, but the accuracy of data is the key question with regard to usability. Nowadays, many leading data-mapping and location-data service companies have access to a huge amount of data and services in the traffic-navigation industry partnered with large automotive manufacturers facilitating the development of ADAS, e.g. HERE HD Live Map [3]. The goal of these companies is to provide a high-quality, self-healing map that is a clear representation of the physical world for the navigation services. However, in terms of testing, not only navigation matters. It is important to have the tools necessary to create the data-mapping environment for the

*Correspondence: kohlrusz.gabor@mk.uni-pannon.hu

simulations. This paper elaborates on the options available to provide the proper format of location data for the TESIS DYNA4 HIL/SIL environment toolset of GIS (Geographic Information System). An overview will be given of the available road data formats as well as the available crowdsourced OpenStreetMap data. A lot of highly detailed geospatial information (obtained from mobile mapping and spatial data processing) is available that can be used to describe the exact road parameters for simulation and modeling. A gap exists between the freely available geospatial information and descriptive standards of roads that is used in driving/traffic simulation and systems of test vehicles. The goal is to provide an accessible, open-source and easy-to-use solution that is able to generate logical data concerning the description and visualization of roads for driving simulation environments based on available map data.

1.1 NDS Building Blocks

The Navigation Data Standard (NDS) is a standardized format for automotive-grade navigation databases established by automotive Original Equipment Manufacturers (OEMs), map-data providers and navigation-device/application providers. This is a standardized binary database format that allows the exchange of navigation data between different systems. The goal is to separate layers of data from the layers of software, mainly for navigation purposes. The following layers of data could be used from the NDS in a simulation environment:

- Digital terrain model
- Traffic information
- Basic map display
- Routing
- Lane information

1.2 OpenDRIVE road format

OpenDRIVE is an open format specification to describe the road networks developed by VIRES. It is a standard road description format that should help data exchange between different driving simulation environments. It has direct support without the necessity to convert into DYNA4, IPG's CarMaker or CarSim formats. OpenDRIVE has special features like complex road networks with junctions, crossfall, superelevation, road markings and barriers, traffic signs and lights, gantries, as well as the ability to integrate high-resolution road surface profiles in OpenCRG format [4]. The toolset of GIS (Geographic Information System) provides a wide range of functions for spatial data processing, but is yet to offer support for the OpenDRIVE standard road-description format [5]. This paper describes a way of gathering and converting road network information from OpenStreetMap into OpenDRIVE format. Using a conversion tool like this, the generation and synthesis of a scenario with regard to real-life road networks for driving simulator applications can be more convenient and faster.

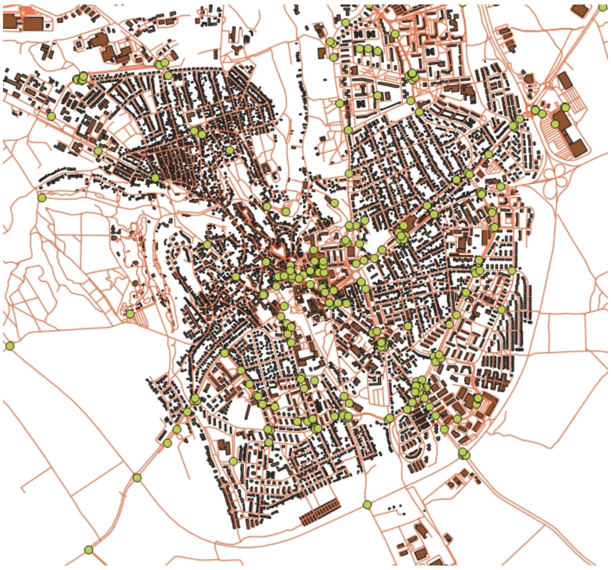


Figure 1: The filtered OSM dataset of Veszprém

1.3 RoadXML

Like OpenDRIVE, this is also an open file format for the logical description of road networks. The goal is to simplify the production of road databases as well as improve the consistency and ensure the interoperability of simulation models. This is an XML-based file format that can be easily edited to enhance road descriptions with custom data using a simple text editor. RoadXML is used by many driving simulator software programs as an alternative file format for describing road networks [3].

1.4 Freely available map data

Other commonly used formats for road networks, e.g. OpenStreetMap and LandXML, are available for geographical needs, but are not suitable for ADAS simulation purposes, however, they can be used as an input source since these have worldwide coverage.

OpenStreetMap is a free, editable map of the whole world that is being built, by and large, from scratch using crowdsourcing methods and has been released under a Creative Commons Share-Alike license. The data records are mostly based on GPS navigation data that can be used to create configurations of road networks based on the line segments of the database [6]. Further descriptive data is necessary to fulfill the needs of simulation environments. A question that might be raised is whether and how vector data can be used in geographic information systems?

Geofabrik is a free, community-maintained data extraction service from OpenStreetMap [7]. This can provide the necessary road-segments data in addition to other information to describe the segments, e.g. the number of lanes, traffic signs, etc. (Fig. 1).

Another option to obtain road-vector data is overpass-turbo.eu. Using this service, interesting map data can be

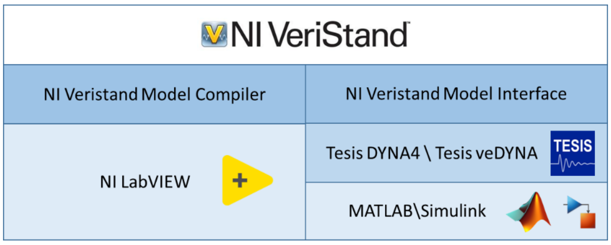

NI VeriStand
NI Veristand Model Compiler
NI Veristand Model Interface
NI LabVIEW
Tesis DYNA4 \ Tesis veDYNA
TESIS
MATLAB\Simulink

Figure 2: Main components of the NI-based system

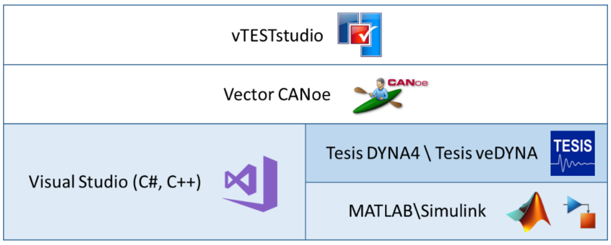


Figure 3: Main components of the Vector-based system

selected by a special query language, downloaded, displayed and used in GIS applications.

Other data sources can be used to obtain other descriptive information. Information concerning elevation can be derived from the Digital Terrain Model produced with the aid of satellites (ESA's Copernicus Sentinel-1 and Sentinel-2) and/or aerial images.

1.5 Open Source GIS software

If the necessary map data is downloaded, a data management and process system is needed to make further data preparations for this GIS software to be used. GIS software (Geographical Information System) is an application that enables geographical data to be managed on different platforms. Quantum GIS (QGIS) is one of the most popular open-source software solutions. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo) that runs on all operating systems and supports numerous vectors, rasters and database formats as well as functionalities, moreover, allows conversion between different reference systems. This has been selected to implement the data management tasks on the downloaded road data.

1.6 Simulation software environment

The road models are mainly necessary for testing and validating position estimation algorithms as well as advanced driving support systems for autonomous vehicles.

A hard real-time environment based on National Instruments devices has already been built for testing, and a soft real-time simulation system based on Vector's VT System is currently under construction. NI's LabVIEW as well as VeriStand are the base software programs, and MATLAB/Simulink-based TESIS DYNA4 software is used for automotive simulation purposes.

VeriStand provides the connection between the models and hardware components, while LabVIEW and DYNA4 provide opportunities to develop new models as well as other software components. In addition, NI's TestStand can be used to create tests and automated test sequences as needed (Fig. 2).

TESIS DYNA4 is also the development environment for vehicle simulations in the Vector-based simulation environment. Other software components and add-on models can be developed primarily in Visual Studio and CANoe in this environment, moreover, vTESTstudio provides a framework for creating tests and test sequences (Fig. 3).

TESIS DYNA4 serves as the vehicle simulation development environment in both cases, which is why the initial goal was to be able to use the data exported from the map databases in this environment.

2. Results and Analysis

An important aspect of simulation tests is how to model the different sections of road. This location information can be obtained from real measurements or map databases. The selected sampling area is a short track in Veszprém. Using the service of geofabrik.de, it is possible to download the latest dataset of the city from the OSM database (Fig. 4). This dataset provides the network level which consists of more data layers, roads and traffic signs, moreover, the buildings can probably be used as an information base for the description of roads. The necessary information was selected and processed by QGIS Desktop GIS software, moreover, the coordinate information of the points from the selected segment of road was exported in a suitable input format for data conversion.

After the data collection task, the next step is to convert the data into the appropriate road-description format

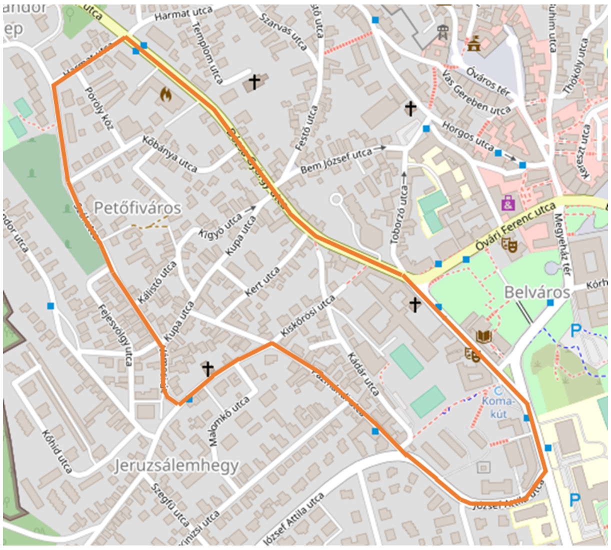


Figure 4: Selected track designated for testing in Veszprém (OSM)

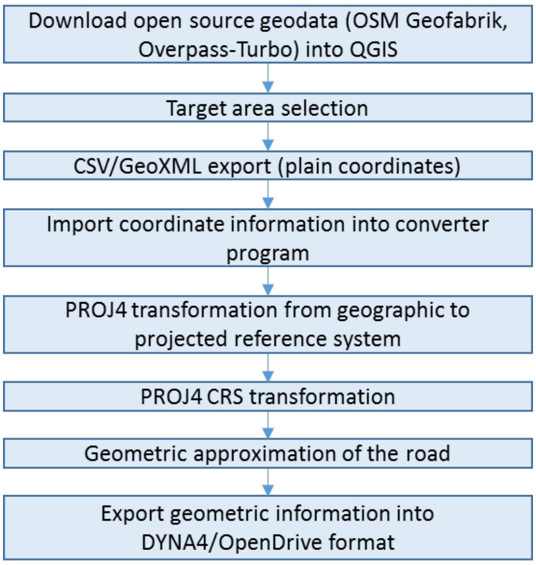

Download open source geodata (OSM Geofabrik, Overpass-Turbo) into QGIS
Target area selection
CSV/GeoXML export (plain coordinates)
Import coordinate information into converter program
PROJ4 transformation from geographic to projected reference system
PROJ4 CRS transformation
Geometric approximation of the road
Export geometric information into DYNA4/OpenDrive format

Figure 5: Transformation process

that can be imported by the selected simulation environment. DYNA4 has a conversion option that can create a track model in its own format from plain GPS coordinates, but has limitations. It does not allow the width of the road segments to be changed significantly, as well as data concerning the width of roads, number of lanes to specific coordinates or data concerning imported traffic signs and lights to be assigned, therefore, no descriptive information with regard to traffic levels can be found. Furthermore, the DYNA4-converted road description cannot be exported into other formats of road data exchange, e.g. OpenDRIVE or RoadXML.

2.1 Data conversion

To solve these problems, a self-developed conversion program was created that can create DYNA4-compliant path models from CSV and XLS file formats.

The necessary conversion steps are the following:

Import coordinates (OSM); convert input data and normalize (PROJ); define track segments and calculate; generate descriptive text; and define descriptive parameters, e.g. width, length and number of lanes as well as heights (Fig. 5).

The program reads the descriptive CSV/XLS file formats and then converts the coordinate data from the WGS84 geographic coordinate system to the Cartesian coordinate system to simplify the calculations. The HD-72 – Hungarian National Datum (EPSG:23700 SRID) was selected, which is a meter-based coordinate system. A proper coordinate transformation is necessary for transformation between the two coordinate reference systems that is provided by the PROJ library A special projection string was applied during the transformation, so the standard transformation error remained below 1m.

The program creates an approximate path according to the input points based on spatial geometry calculations

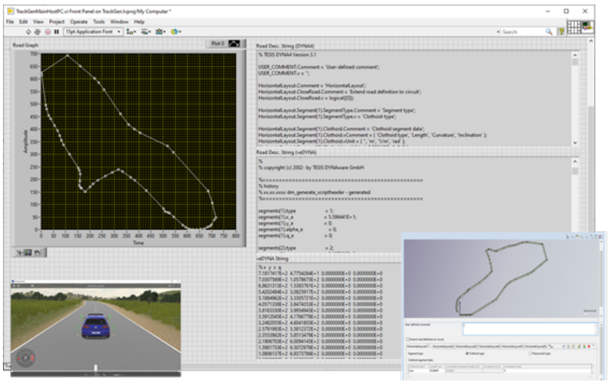


Figure 6: Testing track in Veszprém

compiled from lines, circles and splines corresponding to the DYNA4 format. Based on the path created in this way, DYNA4 files are created that can already be used directly in TESIS projects (Fig. 6).

3. Discussion

Even though the conversion program worked properly, it has one disadvantage. The DYNA4 format does not support the creation of complex intersections and multi-lane roads.

Further development is needed to solve this problem. The conversion tool has to be able to save data in Open-DRIVE format so much more complex track models can be created as well as in other simulation software formats such as IPG's CarMaker.

The coordinate transformation also causes a bottleneck because a meter-based coordinate system is needed for the geometric calculations. Not all projected and geographic coordinate systems are appropriate for simple meter-based calculations. It is important to apply an accurate transformation method that places the coordinate points accurately. The projection string that was used is only suitable in Hungary.

4. Conclusion

This paper introduced a map-tool that can be applied for the conversion of open-source map data to open road data formats. The converted road-description file can be used in different vehicle test systems as a basis for the functional testing of vehicle dynamics and driver assistance systems.

The main advantage of the proposed framework is that if some kind of road network data is available from an area, it is possible to convert it into a road-description format. The created description file can be imported into most vehicle test systems. However, some limitations should be taken into consideration. Although a lot of road data can be found from open-source map databases, currently not all of it is available, e.g. traffic signs and lights, and the quality of the data content needs to be enhanced. Our goal is to perform data collection tasks on the tracks that have been used in this case study.

Acknowledgements

The research was supported by EFOP-3.6.2-16-2017-00002 “Research of Autonomous Vehicle Systems related to ZalaZone autonome proving ground”.

REFERENCES

[1] Dupius, M.; Strobl, M.; Grezlikowski, H.: OpenDRIVE 2010 and beyond – status and future of the de facto standard for the description of road networks, *Proceedings of the Driving Simulation Conference Europe 2010*, 2010, pp. 231-242 ISBN: 978-2-85782-685-9

[2] Richter, A.; Fischer, M.; Frankiewicz, T.; Schnieder, L.; Köster, F.: Reducing the gap between simulated and real life environments by introducing high-precision data, *Driving Simulator Conference 2015 Europe*, 16–18 Sep. 2015, Tübingen, Germany ISBN 978-3-9813099-3-5

[3] Chaplier, J.; Nguyen, T.; Hewatt, M.; Galee, G.: Toward a standard: RoadXML, the road network database format, *Proceedings of the Driving Simulation Conference Europe 2010*, 2010, pp. 211-221 ISBN: 978-2-85782-685-9

[4] Richter, A.; Scholz M.: Deploying guidelines and a simplified data model to provide real world geodata in driving simulators and driving automation, *Transp. Res. Part F Traffic Psychol. Behav.* 2019, **61**, 305–313 DOI: 10.1016/j.trf.2017.04.004

[5] Despine, G.; Baillard, C.: Realistic Road Modelling for Driving Simulators using GIS Data, *Advances in Cartography and GIScience*, 2011, **2**, 431–448 DOI: 10.1007/978-3-642-19214-2_29

[6] Over, M.; Schilling, A.; Neubauer, S.; Zipf, A.: Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Comput. Environ. Urban Syst.*, 2010, **34**(6), 496–507 DOI: 10.1016/j.compenvurbsys.2010.05.001

[7] Kulawiak, M.; Dawidowicz, A.; Pacholczyk, M. E.: Analysis of server-side and client-side Web-GIS data processing methods on the example of JTS and JSTS using open data from OSM and geoportal, *Comput. Geosci.*, 2019, **129**, 26–37 DOI: 10.1016/j.cageo.2019.04.011

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 123–130 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-18

NEW HYBRID WAVELET AND CNN-BASED INDIRECT TIRE-PRESSURE MONITORING SYSTEM FOR AUTONOMOUS VEHICLES

ZOLTÁN MÁRTON*[1] AND DÉNES FODOR[1]

[1]Research Institute of Automotive Mechatronics and Automation, University of Pannonia, Egyetem u. 10, Veszprém, 8200, HUNGARY

Since the tire pressure has a significant influence on driving safety, even self-driving vehicles need to be aware of their current tire pressures. Two major types of methods for estimating tire pressures exist: direct and indirect methods. In spite of recent advancements in direct Tire Pressure Monitoring Systems (TPMSs), indirect pressure monitoring systems still play a significant role due to their low costs. Indirect systems rely on the processing of signals from wheel speed sensors. In most cases, a transformation is applied to generate a frequency spectrum from which the tire pressure-dependent eigenfrequency can be extracted. The most accurate methods apply the Fourier transform, but these require the highest computational power. After the spectrum of signals from the wheel speed sensor is created, the eigenfrequency must be extracted. Several methods are available to extract significant frequency components. One of the easiest methods is peak searching, however, it is susceptible to noise. On the other hand, more accurate methods that are less sensitive to noise require more computational power. If a transform that consumes less computational power can be applied, then the freed resources can be used by a better eigenfrequency identification method. In this paper, a Hybrid Wavelet-Fourier Transform and Convolutional Neural Network-based method is presented, which exhibits a promising level of noise tolerance.

Keywords: TPMS, Eigenfrequency, Hybrid Wavelet-Fourier Transform, Convolutional Neural Networks, Autonomous Vehicles

1. Introduction

The development and spread of autonomous and electric vehicles is about to change everything with regard to the internal structure and operation of cars. However, vehicles will always make contact with the road through tires so tires will continue to play an important role in terms of vehicle safety. The condition of tires is not only dependent on abrasion but on current tire pressures as well. Furthermore, tire pressures influence abrasion in addition to fuel or power consumption [1]. As a result, both the EU and USA have legislated that all new vehicles are equipped with Tire Pressure Monitoring Systems (TPMS) as standard. To achieve safe autonomous driving, the performance parameters of the vehicles must be in accordance with the conditions of the tires, e.g. the driving logic needs to know the current pressure of each tire. Nowadays, research is being done with regard to active tires which are capable of adapting to weather and road conditions by changing their own tire pressures according to the requirements [2]. On the other hand, active tires are extortionate compared to their passive counterparts, even those equipped with TPMS. An active tire must also include a compressor and be able to transmit the compressed air into the tires, which increases not only their cost but also their energy requirements. Active tire systems will only be available for high-end vehicles or military applications. Mainstream and more cost-effective vehicles will continue to use regular tires and be equipped with TPMS.

Two major types of TPMS are available: direct TPMS, which includes a pressure sensor in the tires themselves, and indirect TPMS, which uses Wheel Speed Sensors (WSS) required by the Anti-lock Braking System (ABS), Electronic Stability Program (ESP) and other driving safety systems. The indirect TPMS (iTPMS) can be accomplished by following two techniques. The simplest way is to compare the wheel speed signals from three different pairs, the bigger the tire pressure the bigger the radius of the tire will be. These systems are capable of detecting relative pressure changes and require the results to be filtered in such a way that the driving conditions and maneuvers do not affect the system. Since tires have different radii, these systems require a learning process to be implemented should any of the tires be changed [3]. Nowadays, comparable systems were replaced by signal processing-based systems and research is being done on model-based systems. Both approaches are based on the fact that a tire is like a complex system made of springs and masses, where each spring-mass pair has its own eigenfrequencies. One of the springs corre-

*Correspondence: marton.zoltan@mk.uni-pannon.hu

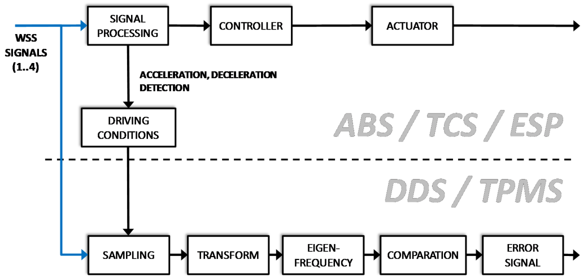


Figure 1: The system architecture of iTPMS

sponds to the air pressure inside the tire, therefore, one eigenfrequency is dependent on the tire pressure [3]. Signal processing-based iTPMSs consist of two major well-separated steps (Fig. 1).

The first step is to transform the signal from the time domain into the frequency or a frequency-related domain. These transforms can be Fourier [4], Cosine [5] or Hybrid Wavelet-Fourier (HWFT). The received frequency spectrum of the signal from the wheel speed sensor contains different eigenfrequencies as well as noises originating from the road surface, combustion engine and transmission. The second step is to detect and isolate the pressure-dependent eigenfrequency. Different algorithms are available to identify a specific eigenfrequency. Each algorithm is only performed on a given bandwidth. The simplest and fastest is the Peak Search (PS) algorithm. Another method is the so-called Center of Gravity (COG) algorithm which can virtually increase the frequency resolution. Although it is more resilient towards noise, noise still may cause the detected eigenfrequency to be shifted. This was subsequently observed in the results presented in Table 1. Since each eigenfrequency found in the spectrum of the WSS exhibits a unique pattern, a pattern-recognition method can also be applied to identify the eigenfrequency in a given bandwidth. One of the most popular and reliable pattern-recognition methods are the deep Convolutional Neural Networks (CNNs). Deep CNNs can learn to recognize and classify patterns through a process called Deep Learning.

In this paper, a novel iTPMS algorithm is proposed which consists of the HWFT and, as an alternative to the COG algorithm, a CNN is used to detect the eigenfrequency.

2. Transforms related to iTPMS

The first major step in signal processing-based iTPMS is always a transform of the WSS or other sensor signals from a time into a frequency or frequency-equivalent domain. Since one of the major aims of this paper is to compare different transforms and eigenfrequency-detecting methods, a short description will be given of the Fourier, Cosine and Hybrid Wavelet-Fourier transforms. For the iTPMS, a "new" HWFT will be proposed in this paper.

2.1 Fourier Transform

One of the most mainstream transforms in frequency analysis is the Fourier transform, which transforms the time-domain signals directly into the frequency domain [6]. The signals in the frequency domain contain information concerning frequencies, amplitudes and phases. Three major varieties of the Fourier transform exist depending on the nature of the signal and the number of samples. The Continuous Fourier Transform (CFT) is predominantly used in stability and control theory, signal processing theory, symbolic mathematics, electrical engineering, etc.

If a continuous signal is sampled, i.e. a discrete signal, by using the integral approximation to the sum, the Discrete Fourier Transform (DFT) can be obtained as

$$X\left[k\right]=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}x\left[n\right]e^{-i2\pi kn/N}, \qquad (1)$$

where k represents the discrete frequency, $x[n]$ denotes the sampled signal, N stands for the number of samples, $X[k]$ refers to the transformed discrete signal and i is the imaginary unit. Unlike the CFT, the DFT has many practical applications.

A special case of DFT is when the number of samples can be expressed as a power of two. In this case, DFT can be factorized using a divide and conquer approach. Hence the factorization of this transform requires less computational power and is referred to as the Fast Fourier Transform (FFT), it is the most widespread in signal processing and computer science.

2.2 Cosine Transform

The Cosine Transform (CT) originates from the Fourier transform by removing the imaginary components. Unlike the Fourier transform, where the phases are encoded in the complex amplitudes, the CT stores the phase information over the entire frequency spectrum if the signal cannot be synthesized entirely from a finite set of cosine functions [7]. CT also consists of three major variants: Continuous (CCT), Discrete (DCT) and Fast (FCT). In practice, FCT is predominantly used in the compression of lossy audio, images and motion pictures. DCT and FCT consist of four different variations, in our case, the so-called DCT or FCT II was implemented:

$$X\left[k\right]=\sum_{n=0}^{N-1}x\left[n\right]\cos\left[\pi\left(k+\frac{1}{2}\right)\frac{n}{N}\right]. \qquad (2)$$

In this paper, FCT was examined as an alternative to FFT.

2.3 Wavelet Transform

The Wavelet Transform (WT) can be seen as a transform which transforms a given signal from the time domain into the frequency-time domain. Like FT and CT,

Table 1: Test results of different combinations of transform- and eigenfrequency detection methods.

Test type	Ref. Freq. [Hz]	Transform	Eigenf. Det.	Average [Hz]	Deviation [Hz]	Correctness [%]
Sim.	47	FFT	Peak Search	46.181	3.808	75.475
		FCT		46.41	2.515	48.636
		HWFT-64		46.359	2.493	66.182
		FFT	CoG	46.682	0.828	79.986
		FCT		46.78	0.835	78.616
		HWFT-64		46.613	0.84	78.029
		FFT	CNN	46.995	0.132	99.853
		FCT		46.94	0.606	95.724
		HWFT-64		46.994	0.174	99.347
Sim.	43	FFT	Peak Search	42.997	4.717	99.539
		FCT		43.255	0.865	57.852
		HWFT-64		43.174	0.438	82.529
		FFT	CoG	43.73	0.144	99.214
		FCT		43.806	0.424	89.367
		HWFT-64		43.886	0.2	79.345
		FFT	CNN	43.011	0.18	99.414
		FCT		43.446	1.19	82.38
		HWFT-64		43.092	0.322	91.363
Real	45.58	FFT	Peak Search	45.21	7.224	The reference frequency was given in the metadata provided by the measurements and it was calculated by a closed source software in off-line processing. Hence, there is no guarantee that the reference frequency is always correct. The correctness values are omitted.
		FCT		46.013	5.172	
		HWFT-64		45.775	4.29	
		FFT	CoG	46.59	1.387	
		FCT		46.671	1.394	
		HWFT-64		47.225	1.407	
		FFT	CNN	45.202	3.186	
		FCT		45.254	3.249	
		HWFT-64		45.487	2.831	
Real	47.03	FFT	Peak Search	46.949	3.814	
		FCT		47.374	4.897	
		HWFT-64		46.901	4.53	
		FFT	CoG	47.339	1.683	
		FCT		47.259	1.584	
		HWFT-64		47.695	1.521	
		FFT	CNN	46.711	2.571	
		FCT		46.817	2.406	
		HWFT-64		47.103	2.194	

WT also consists of three variants, namely continuous (CWT):

$$X\left(t,s\right)=\frac{1}{\sqrt{s}}\int_{-\infty}^{\infty}x\left(t\right)\overline{\psi}\left(\frac{\tau-t}{s}\right)dt, \qquad (3)$$

discrete (DWT):

$$X\left[m,k\right]=\frac{1}{\sqrt{c_0^k}}\sum_{n=0}^{N-1}x\left[n\right]\overline{\psi}\left[\left(\frac{n}{c_0^k}-m\right)T\right], \qquad (4)$$

and fast (FWT), but unlike the aforementioned transforms, WT is not just a transformation, it resembles a whole family of transformations. In the above equations, where $\psi(t)$ is the so-called Mother Wavelet function, $x(t)$ denotes the continuous input signal or function, and $X(t,s)$ represents the WT.

CWT can be discretized. This process requires a scaling base c_0 and a time unit of the Mother Wavelet, T, to be defined:

$$X\left[m,k\right]=\frac{1}{\sqrt{c_0^k}}\sum_{n=0}^{N-1}x\left[n\right]\overline{\psi}\left[\left(\frac{n}{c_0^k}-m\right)T\right] \qquad (5)$$

If the scaling base is two and the total number of samples can be expressed as a power of two, the previously obtained DWT can be factorized. The obtained FWT can be implemented as a cascade of low pass filters (LPF), high pass filters (HPF) and downsampler banks [8].

Depending on the Mother Wavelet function, different WTs can be defined. These WTs share common properties, e.g. the frequency and time resolutions of the transform are dependent on each other, they can be viewed as bands where both resolutions are in logarithmic steps. Their other properties depend on which Mother Wavelet is selected. The most widespread WTs are the Mexican hat Wavelet, Haar (Wavelet) Transform and Cohen-

Daubechies-Feauveau (CDF) Wavelets. WTs are applied in data compression, e.g. JPEG2000, DjVu, CineForm, etc., and transient analysis.

2.4 Hybrid Wavelet-Fourier Transform

While WT was applied in signal processing for transient and Fourier transforms in frequency spectrum analysis, demand for both grew simultaneously. One of the solutions to meet this requirement was HWFT. The basic idea behind it was that the WT decomposes the signal of the time domain into bandwidths. Inside a bandwidth, characteristic information concerning the time domain in the given bandwidth can be found. Therefore, it is possible to perform FTs on each bandwidth [9].

Unlike the Continuous HWFT:

$$X\left(f,s\right)=\frac{1}{\sqrt{s}}\int_{-\infty}^{\infty}x\left(t\right)\overline{\pi}\left(\frac{\tau-t}{s}\right)dt, \tag{6}$$

which is scarcely applied, the Discrete HWFT:

$$X\left[f,k\right]=\sum_{m=0}^{N-1}\frac{e^{-i2\pi fm/N}}{\sqrt{Nc_0^k}}\sum_{n=0}^{N-1}x\left[n\right]\overline{\psi}\left[\left(\frac{n}{c_0^k}-m\right)T\right] \tag{7}$$

has some applications in biometry [10]. Fast HWFT (FHWFT) can be synthesised by applying FFT to the output of each bandwidth of the FWT.

From the perspective of our research, FHWFT is suitable to calculate a frequency spectrum of a WSS signal for iTPMS applications. Furthermore, FHWFT has a special property which can be exploited to reduce the computational complexity by only performing the FFT on those bandwidths that are significant with regard to the given application. This property allows it to be used as a lightweight alternative to the traditional FFT. However, since the filter banks of the FWT are imperfect, a phenomenon known as spectral leakage can be observed (Fig. 2). Spectral leakage essentially means that frequency components from one bandwidth also appear in the neighboring bandwidths [9]. This causes additional disturbances in the spectrum necessitating further investigation into how much the detection of eigenfrequencies in an iTPMS is affected. For this reason, FHWFT is compared with FFT and FCT using different eigenfrequency detection methods in this paper.

3. Eigenfrequency Detection Methods

The most important step in an iTPMS is eigenfrequency detection because this identifies the frequency component inner pressure dependent. Several methods are capable of identifying or detecting peaks in a given frequency spectrum. The sensitivity, ability to handle multiple peaks, noise susceptibility, etc. of these methods differ.

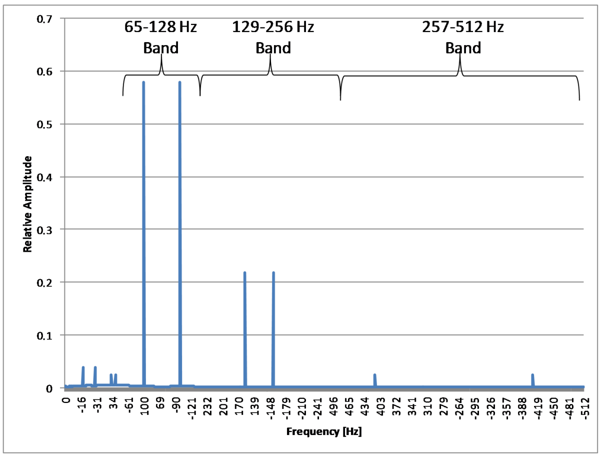


Figure 2: Spectral leakage of the HWFT demonstrated on a 100 Hz signal

3.1 Peak Search Algorithm

One of the simplest methods is the Peak Search (PS) algorithm. In the first step, a bandwidth must be determined in which the interesting and/or important eigenfrequencies can be present. In this bandwidth, a search is performed to identify the frequency which exhibits the maximum amplitude. This simple algorithm is presented in

$$F_p=\left\{f\in I\,\middle|\,\max_i|S(i)|=|S(f)|\right\}, \tag{8}$$

where F_p denotes the set of peaks, I represents the interval of interest, and $S(f)$ stands for the frequency spectrum of a signal. Under certain circumstances, multiple frequency components can be found. Depending on the application, mostly vibration analysis, error detection as well as searching for local maxima and multiple peaks is required. In such applications, usually a minimum value ε is also defined to restrict the number of possible peaks:

$$F_p=\{f\in I\,||S(f)|>\varepsilon\text{ and }||S(f)|\text{ is local max}\} \tag{9}$$

If noise or disturbances are present, the PS algorithm might identify the wrong peaks. To reduce the impact of the noise, a Sliding-Window Median Filter might be applied. If the PS algorithm in Eq. 9 is used, then it is more resilient to noise than the global maximum method (Eq. 8), but a priori information is required to identify the correct eigenfrequency.

Neither of the PS methods are as accurate in such systems as the iTPMS because pressures about 30 % lower than the optimum shift the pressure-dependent eigenfrequency by only approximately $3-4$ Hz. Due to disturbances originating from the road surface and transmission, multiple peaks might be present in the interval of interest.

3.2 Center of Gravity algorithm

A more noise-resistant method is the Center of Gravity (CoG) algorithm since it involves weighted averaging.

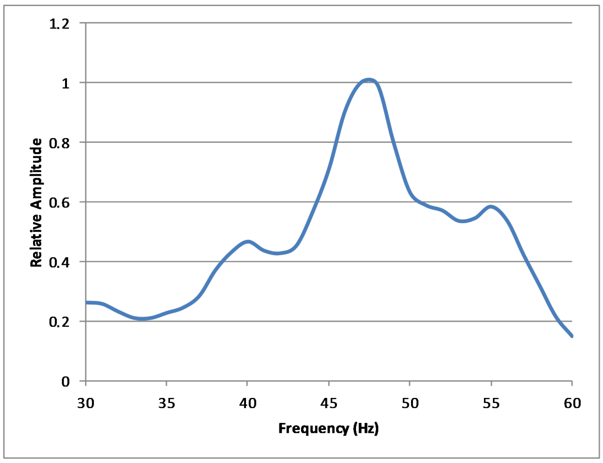


Figure 3: The pressure-dependent eigenfrequency of a tire in the spectrum of the WSS signal

Unlike the PS, COG always yields only one frequency, therefore, this eliminates the problems that originate from multiple peaks present in the Interval of Interest. Furthermore, the eigenfrequency theoretically has a higher resolution than would have resulted from the sampling frequency and sampling time. This theoretically increased frequency resolution can only be achieved in the presence of an error and only in that case when exactly one frequency component is present in the Interval of Interest. The COG can be calculated by

$$f_c = \frac{\int_I f\,|S(f)|^p\,df}{\int_I |S(f)|^p\,df}, \tag{10}$$

where f_c denotes the center frequency and p represents an exponent whose value, depending on the application, is usually between 2 and 3. Although this method is more robust than PS, it is still susceptible to noise.

3.3 Convolutional Neural Networks

The eigenfrequency of the tire exhibits a very distinct pattern in the frequency spectrum of the WSS signal (Fig. 3) which facilitates its detection using pattern-matching algorithms. One of the most popular pattern-matching algorithms involves deep (multilayered) CNNs. Our research focused on developing an algorithm for eigenfrequency detection using CNN. On the downside, CNNs demand much more computational power and memory than PS or COG. To compensate for the higher computational power, a FHWFT was strictly developed to carry out the transform in the bandwidth where the tire pressure-dependent eigenfrequency is located (HWFT-64).

The Artificial Neural Networks (ANNs), e.g., CNN, mimic the structure and supposed operation of neurons and neural systems found in living beings. The differences concern the actual operation of neurons and, unlike the central nervous system of a living being, ANNs often structure Artificial Neurons (ANs) into layers. Since in most cases one layer of neurons is insufficient for most applications, multilayer ANNs are often constructed. ANs often have multiple inputs, each of which has a unique weight assigned to it. Unlike living neurons, the inputs and outputs of ANs can be vector-valued. ANNs store the recognizable patterns in the weight values of their ANs. The analytical determination of the correct weight values is taxing but like the central nervous system of a living being, ANNs can learn the desired patterns. The layers of an ANN can have different functions. In the case of CNNs, at least one of those layers implement spatial discrete convolution, hence the name Convolutional Neural Networks. The structure of a layer depends on its function as well as the connection between the given layers and the previous layer. Each AN in the layer has a layer-specific Activation Function (AF) which acts as the output function of the AN. Depending on the function of the layer, the following types of layers can be distinguished: input, convolutional, polling and fully connected layer. Since ANs can have vector output values, each layer can be regarded as if each component of the output vector has its own parallel layer with its own weights, which are grouped together as matrices. The only thing in common would be the AF associated with these virtual parallel layers [11].

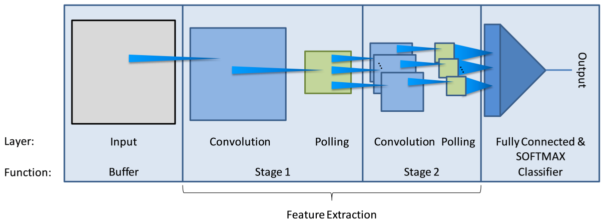


Figure 4: The structure of a simple image classification CNN [12].

Pattern recognition that applies CNNs usually consists of four different types of AN layers: input, convolutional, polling and fully connected (Fig. 4) [12]. The simplest layer is the Input Layer. It has no input weights, moreover, its AF is the identity function and serves as a buffer layer. In the case of Convolutional Layers, each AN shares the same input weights matrices which are referred to as convolution kernels. These layers are applied for noise reduction, filtered downsampling, upsampling and feature extraction. The Polling Layer is used for downsampling or dimension reduction of data. Like the Input Layer, it has no input weights but a kernel radius and step distance. Its specific AFs are average, minimum and maximum functions. The Fully Connected (FC) Layer is the most important part of an ANN-based classifier. The number of ANs must be identical to the number of classes the ANN has to distinguish. Its name originates from the fact that each of its neurons are connected to each of the neurons in the previous layer and each connection has a unique weight assigned to it. The ANs in this layer also have a so-called class index attached to them. During the classification, the neurons contain the possibility of their attached classes. This decision can be

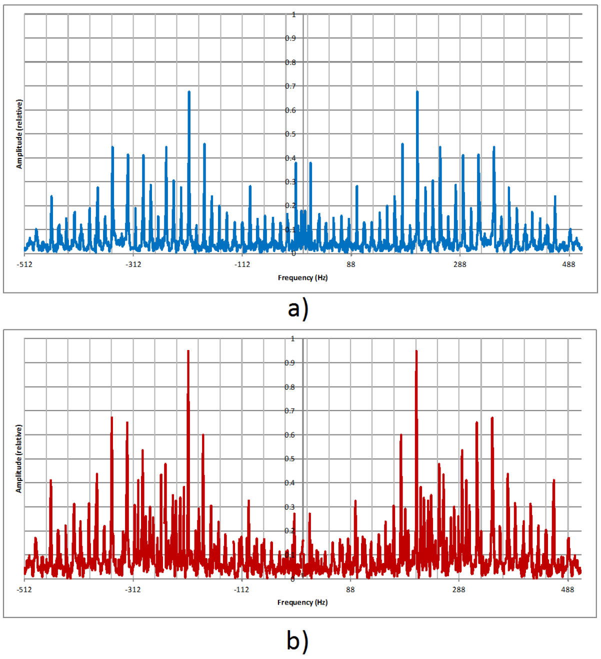


Figure 5: The frequency spectrum of a WSS signal produced by FFT (a) and FHWFT (b) after reordering

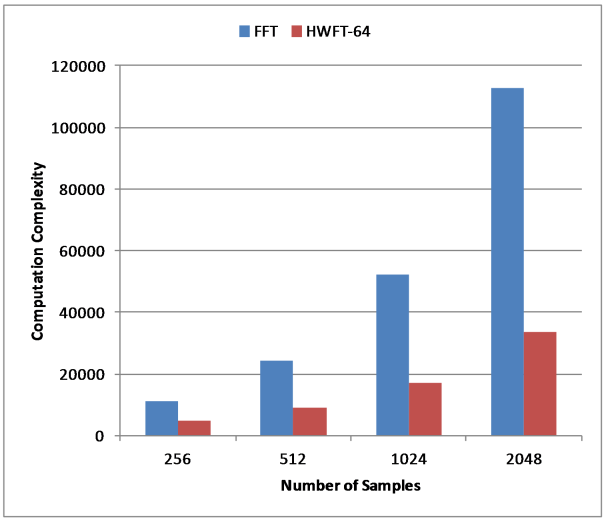


Figure 6: Computational complexity of FFT and HWFT-64

influenced by the chosen AF, from which different functions are available. The most commonly used is the so-called softmax function [11].

4. Our New iTPMS Algorithm

4.1 CNN-based Determination of Eigenfrequencies

Since similar results can be achieved to FFT using FHWFT, it can be concluded that FHWFT can be safely used as an alternative to FFT in iTPMSs (Fig. 5). The eigenfrequency, depending on the tire pressure, is between 42 and 48 Hz. Using this information, a FHWFT can be optimized which only calculates the FFT within the bandwidth of 32–63 Hz. If the sampling rate is 1024 samples per second, the FFT has to be performed on only 64 samples. This FHWFT optimized for TPMS was labeled by us as HWFT-64. The computational power requirements of HWFT-64 are about four times less than in the case of a FFT with a sampling rate of 1024 (Fig. 6). The freed up resources make it possible to use more advanced eigenfrequency detection methods than PS or COG algorithms. Since the CNN can learn different patterns, it is possible to construct such learning patterns which include various frequency disturbances. Naturally, the CNN must also be constructed in such a way that such patterns could be learnt correctly. Furthermore, the CNN should be as simple as possible.

As the first step of the design using the a priori information and by taking the available free and open-source CNN software tool into consideration, an interval of 16 Hz was selected as the interval of interest with 47 Hz (the eigenfrequency of the non-deflated tire) at the center. One of the goals of this research was to create a better eigenfrequency detection method than PS. This meant that 16 classes had to be and were specified, thus the last layer of the CNN had to be a fully connected layer that consisted of 16 neurons with a softmax function. Since the input interval is comprised of 16 elements, the input layer also had to consist of 16 neurons. The structure of the pattern recognition-based eigenfrequency-detecting CNN can be even simpler than the simple image classifier shown in Fig. 4 since much less input data is used.

The first attempt just applied a two-layer approach, with an Input and an FC Layer. This design could not be validated during the learning process. To enhance the capabilities of the CNN, an additional layer had to be inserted. The new layer was a Convolutional Layer consisting of four neurons and a 3x3 convolution kernel. The outputs were 16D vectors (Fig. 7). The resulting CNN was capable of passing the validation tests and robust against the simulated disturbances. Implementation of the CNN had to be capable of running on an ECU which meant no OpenCL, Compute Unified Device Architecture (CUDA), Compute Shaders or multiprocessor-based implementation could be used. The MOJO-CNN was chosen, which is an open-source implementation with different built in solving algorithms such as Adam, SGD and AdaGrad [13]. Only the Adam solving algorithm was used.

The learning samples were created by frequency shifting of the spectrum of a non-deflated tire (Fig. 3). Additional learning samples were created by injecting a spike at 54 Hz because, in most cases, disturbances would occur at this frequency. To prevent overfitting, data augmentation was used. The validation samples were slightly altered versions of learning samples. The amplitude was altered to such a degree of different frequency components that the location of the peaks remained unchanged.

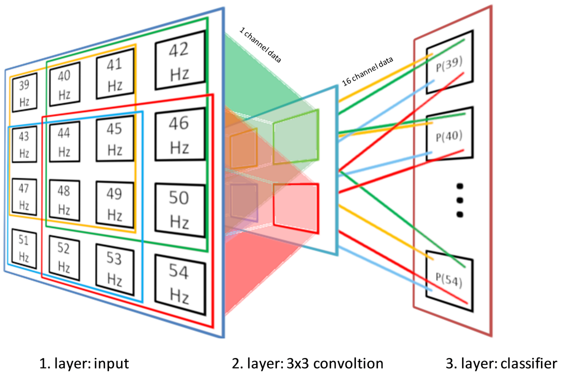


Figure 7: Structure of eigenfrequency-detecting CNN

After the design of the CNN, the new algorithm was tested. For the tests, FFT, FCT and HWFT-64 were selected as the transforms. The eigenfrequency identification algorithms were PS, COG and our new CNN-based pattern recognition. The test data used consisted of both artificial, referred to as “simulated”, and real measurement data.

4.2 Simulations and Measurements

The first sample tests were conducted in such a way to facilitate a known eigenfrequency. This was accomplished by taking real measurement data, which were filtered by a non-ideal Band-Stop Filter (BSF). The BSF being non-ideal makes it possible for the original frequency components to be still present but attenuated and, therefore, act as disturbances. Following filtration by the BSF, known frequency components were injected (47 Hz and 43 Hz for the non-deflated and deflated tires, respectively).

As previously stated, the other set of test data were unmodified measurement data. The data was provided by Continental AG and the expected eigenfrequency was used as a reference frequency. Two “simulated” and two unmodified measurement sets of test data were used for the tests. During the tests, the average frequency deviation and accuracy were evaluated. The accuracy was defined as how many times the algorithm would yield the reference frequency compared to the number of times the algorithm was executed. This could only be determined on the “simulated” data set, since on real measurements, the eigenfrequency cannot be guaranteed to be always the same as given in the metadata sheet. The noises and disturbances were uncontrolled and originated from the road surface, transmission and combustion engine.

5. Results and Discussion

The results can be seen in Table 1. As can be observed, the CNN yields better results with regard to approaching the reference frequency more appropriately and with less deviation than PS and comparable results to COG in the case of both “simulated” and real measurement data.

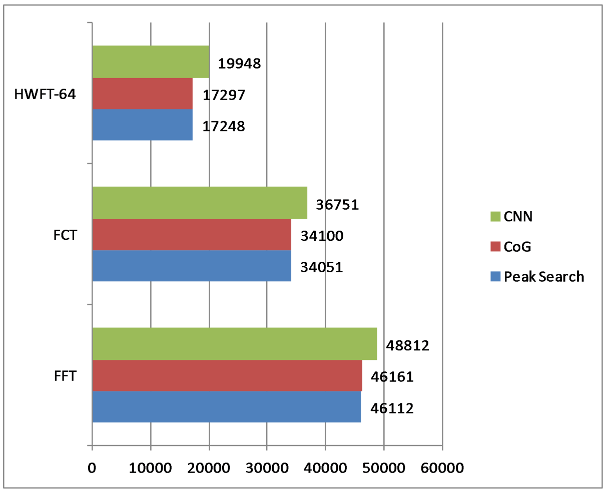


Figure 8: Computational complexity of different combinations of transform- and eigenfrequency-detecting methods, when the number of samples was 1024 and the interval of interest consisted of 16 frequency spectral amplitudes

The most significant improvement of the CNN-based pattern recognition method can be observed in the results from the FCT during the “simulated” data tests. However, except for the first test on the 47 Hz “simulated” data, the COG yielded the smallest deviations, the differences between the average frequencies of COG and reference frequencies are bigger than in the case of PS or CNN-based pattern recognition methods. In the case of the CNN-based method, the deviation was about 50 % greater than in the case of COG. This is most likely due to the fact that the output of the CNN-based method yielded a lower frequency resolution than COG. On the other hand, the HWFT-64 with CNN shows promising and comparable results using just less than half of the computational power required by FFT and COG.

6. Conclusion

In this paper, a new pattern recognition-based eigenfrequency detection algorithm for iTPMS was presented. By combining this new algorithm with an optimized FHWFT, the resulting system is of lower computational complexity, moreover, the reliability and accuracy is almost the same as that of the currently industrial mainstream FFT and COG combination (Fig. 8). The computational complexity was calculated by how often simple mathematical operations supported natively by the CPU/FPU were used in the implementation and how those embedded in iterations were affected by the input data size. In this case, the input data size was 1024 samples. The COG can also work with the HWFT-64 but is slightly less reliable (Table 1). Since more free resources are still available, when using this new transform optimized for iTPMS, further improvements or more complex eigenfrequency detection methods can be used without exceeding the computational complexity of the FFT.

Acknowledgements

The research was supported by EFOP-3.6.2-16-2017-00002 programme of the Hungarian National Government.

REFERENCES

[1] National Highway Traffic Safety Administration, Federal Motor Vehicle Safety Standards, Tire Pressure Monitoring Systems, Controls and Displays, March 2009, http://www.nhtsa.dot.gov/cars/rules/rulings/tirepresfinal/TPMSfinalrule.pdf

[2] Schoettle, B.; Sivak, M.: The Importance of Active and Intelligent Tires for Autonomous Vehicles, The University of Michigan Sustainable Worldwide Transportation Report, 2017, Report No. SWT-2017-2 http://umich.edu/~umtriswt/PDF/SWT-2017-2.pdf

[3] Silva, A.; Sánchez, J. R.; Granados, G. E.; Tudon-Martinez, J. C.; Lozoya-Santos, J. J.: Comparative Analysis in Indirect Tire Pressure Monitoring Systems in Vehicles, *IFAC-PapersOnLine*, 2019, **52**(5), 54–59 DOI: 10.1016/j.ifacol.2019.09.009

[4] Gustafsson, F.; Drevo, M.; Forssell, U.; Lofgrën, M.; Persson, N.; Quicklund, H.: Virtual Sensors of Tire Pressure and Road Friction, *SAE Technical Paper Series*, 2001. DOI: 10.4271/2001-01-0796

[5] Márton, Z.; Fodor, D.; Enisz, K.; Nagy, K.: Frequency Analysis Based Tire Pressure Monitoring, *2014 IEEE International Electric Vehicle Conference (IEVC)* DOI: 10.1109/IEVC.2014.7056187

[6] Oberst, U.: The Fast Fourier Transform, *SIAM Journal on Control and Optimization*, 2007, **46**(2), 496–540 DOI: 10.1137/060658242

[7] Strang, G.: The Discrete Cosine Transform, *SIAM Review*, 1999, **41**(1), 135–147 DOI: 10.1137/S0036144598336745

[8] Goswami, J. C.; Chan, A. K.: Fundamentals of Wavelets, John Wiley & Sons, Inc., 2011 DOI: 10.1002/9780470926994

[9] Tarasiuk, T.: Hybrid Wavelet-Fourier Spectrum Analysis, *IEEE Transactions on Power Delivery*, 2004, **19**(3), 957–964 DOI: 10.1109/TPWRD.2004.824398

[10] Ziółko, B.; Kozłowski, W.; Ziółko, M.; Samborski, R.; Sierra, D.; Gałka, J.: Hybrid Wavelet-Fourier-HMM Speaker Recognition, *International Journal of Hybrid Information Technology*, 2011, **4**(4), 25–41

[11] Albawi, S.; Mohammed, T. A.; Al-Zawi, S.: Understanding of a Convolutional Neural Network, *2017 International Conference on Engineering and Technology (ICET)*, 2017. DOI: 10.1109/ICEngTechnol.2017.8308186

[12] Hijazi, S.; Kumar, R.; Rowen, C.: Using Convolutional Neural Networks for Image Recognition, IP Group, Cadence, 2015. https://ip.cadence.com/uploads/901/cnn/

[13] https://github.com/gnawice/mojo-cnn/wiki

HUNGARIAN JOURNAL OF
INDUSTRY AND CHEMISTRY
Vol. 48(1) pp. 131–138 (2020)
hjic.mk.uni-pannon.hu
DOI: 10.33927/hjic-2020-19

XBRL UTILIZATION AS AN AUTOMATED INDUSTRY ANALYSIS

Alex Suta*[1] and Árpád Tóth[1]

[1]Research Center of Vehicle Industry, Széchenyi István University, Egyetem tér 1, Győr, 9026, HUNGARY

In the last two decades, electronic financial reporting went through a significant evolution, where to date, eXtensible Business Reporting Language (XBRL) has become the leading platform that is already obligatory for listed entities in the United States and was also legislated in the European Union from January 1, 2020. The primary objective of this research was to review the US-listed companies' 2018 quarterly reports. The study generated an automated industry analysis for the automotive industry from the aspect of four main financial item categories as an alternative to statistics-based, manually prepared industry analyses. Statistical tests were carried out between two industrial classification methodologies, the securities' industry identification marks and the reported Standard Industrial Classification (SIC) codes. The results showed a significant difference between the industry classification methodologies. Automated reporting was more precise with regard to the identification of the listed and reporting entities, however, the data fields of SIC codes within the XBRL data set provided an inaccurate classification, which is a potential area of improvement along with additional recommendations outlined in the Conclusion.

Keywords: XBRL, US-listed entities, ACL, Automated data analytics, Industry analysis, SIC codes

1. Introduction

The electronic reporting and automated fundamental reviews in the field of financial reporting is becoming increasingly important considering the difficulties and error-prone procedure of manual analysis from the available source of information. The eXtensible Business Reporting Language (XBRL) provides a standardized platform for this activity, which supports automated and digitalized reviews compared to the paper-based reports from the previous manual. This electronic reporting platform is already used as the official reporting form in the United States for listed entities, therefore, the application of a proper industry classification is essential. Even though XBRL reporting is required by the U.S. Securities and Exchange Commission (SEC), research institutions can choose from various generally accepted industry classifications. Despite the lack of regulation, it is a primary interest of research institutions to protect their reputations by adequately representing companies from the various industries. The two different approaches might provide different results, which can lead to inaccurate trend projections or unreliable industry comparisons. The validation of the XBRL classification and reports by marketing research firms can only be reconciled and validated to statistical industry reports which identify discrepancies. To date, Standard Industrial Classification (SIC) codes are used in the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system to define the type of business of companies. Based on its primary activity, each company assigns a four-digit code to itself when registering an Initial Public Offering (IPO) with the SEC [1]. The four digits indicate levels of description of the industry classification, e.g. the location hierarchy for car manufacturers is Division D - Manufacturing (codes 20-39), code 37: Transportation equipment, and code 3711: Motor vehicles and Passenger Car Bodies [1]. The objective of this research was to review the US-listed companies, where XBRL reports are already required and implemented. Subsequently, through an automated review of the automotive industry, to then identify how these reports can be compared to the European listed entities. This information is crucial for stakeholders and regional policymakers to gain a clear view of the conditions of the target industry. According to the European Securities and Markets Authority (ESMA), from January 1, 2020 onwards, new requirements on the stock exchange-listed companies in the European Union came into effect to provide respective financial statements in a new European Single Electronic Format (ESEF). This is a significant change to the application of XBRL as companies now have to provide reports in this specific reporting language. The data sets include structured information; for this reason, a new wave of research initiatives is expected in this academic area that could follow on from inconsistent industry classifications, further hindering comparability.

*Correspondence: suta.alex@ga.sze.hu

2. Literature review

2.1 XBRL utilization in industry-specific data analysis

Prior literature has documented uses of XBRL in a variety of data analysis environments, generally in the research areas of accounting and financial reporting. Systematic financial data provides data analysts and investors with the ability to measure performance and risks, as well as create comparisons, ratings and other value-added products [2]. Connected to comparability aspects, several sources have been reviewed that are related to the semantic issue of industrial classification. Being a driver of electronic data interchange, XBRL data sets are constructed from multiple identifying tags and numerical data that can be processed by computer software [3]. While the technical background on data-centric analysis is available [4] 2013 [5] 2014, [6], it is uncommon in the industry-specific research literature that XBRL databases are used as the primary source of data.

Chychyla-Leone-Meza [7] measured financial reporting complexity by comparing the quantity of text in US Generally Accepted Accounting Principles (GAAP) and SEC regulations of textual data from XBRL filings. In this study, the variation with regard to the data content of different taxonomy versions (denominative tags, labels, documentation) is emphasized. For this reason, the annual changes in published taxonomy updates have to be taken into consideration [8]. Despite the existence of the XBRL Industry Resource Group established by the FASB [8], Standard Industrial Classification (SIC) codes are not part of taxonomy updates and their current 2007 form seems to be generally accepted for statistical use. Felo-Kim-Lim [9] observed changes in the information environment of analysts by the overuse of customized tags, creating assumptions based on industrial classification as a factor. Zhang-Guan-Kim [10] proposed an expected investor crash risk model based on financial information gathered from XBRL-based SEC databases. In terms of the estimation of the impacts, the industry median of customized tags is generated by 2-digit SIC codes as an adjustment tool in the regression model. Similarly, industrial classification was taken into account as a dummy variable during the analysis with regard to the XBRL adoption of reductions in audit fees as per Shan-Troshani-Richardson [11].

In other XBRL-based studies, industry-specific assumptions required solutions other than basic SIC codes. Liu-Luo-Wang [12] reviewed the effect of XBRL adoption on information asymmetry, where SIC was reclassified to identify high-technology industries.

2.2 Discrepancies between industrial classification systems

Since the emergence of the North American Industry Classification System (NAICS) in 1997 as a sound replacement of Standard Industrial Classification (SIC) codes in U.S. industrial statistics, papers have reviewed the impacts of different frameworks in financial research. Effective comparative statistics require the use of a standardized classification system [13]. The U.S. Economic Census Bureau has made regulatory, business and academic purposes of performing economic research on historical data possible. In 1997, the existing framework, the SIC, was replaced by the NAICS [14]. Unlike the SIC's mixed production/market system, NAICS introduced a production-oriented economic concept that supports the examination of industry-specific indicators such as productivity, input-output relationships and capital intensity [15]. The specific rearrangements between industrial classes primarily affected manufacturing industries, where the SIC functions as a somewhat outdated alternative. U.S. government departments, namely the Bureau of Labor Statistics (BLS), Internal Revenue Service (IRS) and Social Security Administration (SSA), alongside the U.S. Securities and Exchange Commission (SEC), continue to use the most recent 2007 update of four-digit SIC codes. While maintaining a unified classification system is necessary for government departments, the lack of conceptual harmony between industrial classification systems creates a discrepancy with academic research [16]. Several papers have been collected that present empirical evidence of disharmonious schemes based on Financial Statement Data Sets. Kahle-Walking [17] observed differences in financial variables gathered from two statistical databases (CRSP and Compustat) using four-digit SIC codes to be substantial, moreover, showed that commonly used methods of industrial classification disagree due to frequent changes in the SIC codes of firms. Bhojraj-Lee-Oler reviewed the capital market applications of four broadly available industrial classification schemes and found that a significant degree of variance with regard to the number of companies represented in industry divisions exists. The study argues that the six-digit Global Industry Classification Standard (GICS), followed by NAICS, offers better comparability between firms concerning SIC in terms of the critical evaluation of financial ratios and that industrial classification is essential in instances of fundamental analysis. While GICS reflects the dynamic changes in industry sectors, being a privately available system mainly involved in investment processes, it is unlikely to be suitable in statistical research [18]. Kelton-Pasquale-Rebelein [19] referred to SIC codes as outdated in the field of industry cluster analysis and prepared an updated framework using NAICS. As opposed to classifying establishments according to similar products (SIC), the groups are formed from identical production processes (NAICS). Hrazdil-Zhang [20] and Hrazdil-Trottier-Zhang [21] published empirical results on the heterogeneity of industry concentration with the use of SIC and other classification schemes based on the market shares of sales and financial ratios of companies in the manufacturing sector (SIC 2000-3999). According to their findings, the SIC system remains inferior to GICS and NAICS in terms of industrial homogeneity.

Instead of ordinary company databases such as Compustat or S&P 1500, Papagiannidis et al. proposed an exploratory big data method to gather regional research of industry clusters based in the UK. In this study, keywords connected to business operations were collected from official websites to enhance the level of detail provided by single SIC codes, supporting the formation of regional clusters. It is a common conclusion in the reviewed literature that the sole use of SIC codes in industry analysis could lead to the loss of information and false estimation of market forces; in this context the potential of XBRL as a primary data source of financial statements has been reviewed.

2.3 The multi-tier supply chain approach

One possible outcome of the barriers of traditional statistical classification systems is the addition of extra information to existing schemes. In an industrial analysis, especially in the automotive industry, it is essential to differentiate between operational properties, e.g. their position in the automotive supply chain. The contemporary position of an industry must be judged by the different weights of its market players. Assumptions about financial information are heavily affected by the final product, whether it is a part of the interorganizational supply chain, or sold to dealerships or directly to consumers in the form of passenger cars. In terms of a supply chain, manufacturers and suppliers can be classified into multi-tiered groups based on their position in the production chain, as well as the state of raw materials (tier 3 and additional sub-tiers) in addition to finished or semi-finished components (suppliers from tiers 1 and 2) compared to fully finished products (Original Equipment Manufacturers (OEMs)). Concerning the automotive industry, sources from both academia, business and governments [22–24] agree that market players from multi-tier supply chain structures can be ranked as follows:

1. OEMs: a concentrated group of companies accountable for the main manufacturing, assembly and design processes that possess a large market share and well-known brand names;

2. Suppliers from Tiers 1 & 2: potentially several hundred large or small companies, accountable for the supply of automotive parts and systems to OEMs. The range of sold goods is diverse and includes engine components, interior, exterior, transmission as well as cooling and electronic systems. Although their role in the supply chain is consistent, suppliers from Tiers 1 and 2 vary in their direct/indirect (through other participants) nature of interaction with OEMs, therefore, from a statistical viewpoint, can be aggregated;

3. Tier 3 and sub-tier suppliers: several thousand smaller companies are accountable for the supply of raw materials to suppliers from Tiers 1 & 2.

In the scientific literature, several utilizations of the multi-tier supply chain approach exist. Mena-Humphries-Choi [25] reviewed the existing literature at the time on structural arrangements (buyer-supplier-customer) and prepared three cases of theoretical linkage. According to the study, the most typical structure of the automotive industry is the “closed triad”, where the buyer (OEM) can insist on certain requirements (either assurance or training function) not only from Tier 1 but sub-tier suppliers as well. Masoud-Mason [26] used the multi-tier system in the automotive industry to simulate cost optimization on a supply-chain level. Thomé et al. [27] adopted a similar approach of representing many tiers and their interactions that affect selected flexibility measures (product, responsiveness, sourcing, delivery and postponement). Other popular fields of use are sustainability-related questions and green supply chains [28–30].

The available literature clarifies the widespread applicability and general acceptance of tiered levels of suppliers, which supports the methodology examined in the current study. Despite its academic use, the application of the well-established OEM / tiered system of suppliers in automotive business reports published by major consulting firms [31–33] is common practice.

3. Data collection and methods used

The SEC has published XBRL data sets containing raw aggregate financial statement data quarterly since 2009. At the same time, as a premium service, the SEC offers a professional version of its search engine [34] designed specifically to fit the goals of professional financial analytics. However, in line with tendencies identified from the literature review, even a discrepancy on the same platform exists between the Standard Industrial Classification codes current in XBRL data sets and the EDGAR search tool. To perform an automated industry analysis, a suitable classification is required. In this study, a possible classification using the software program ACL (Audit Command Language) Robotics Professional version 14.1.0.1581 is evaluated. From the listed U.S. entities, those operating in the automotive industry were selected to measure deviance in terms of crucial financial indicators between the two data sources. The specific choice of the automotive industry lies in its accurate definability, while the goal of the study was to provide an industry-independent methodology of data analysis that can be applied to several other fields. The two main platforms of data collection were EDGAR Pro Online (2019) operated by the SEC, which is equivalent to the quasi-manual download process of financial statements, and the obligatory quarterly reports of aggregate data sets in the XBRL format available on the SEC website. To avoid existing industrial classification issues, a multi-tier supply chain approach was introduced by grouping companies as OEMs and suppliers from Tiers 1 & 2 (T1&2 S).

4. Results

4.1 Data categorization: number of companies and industries

By using the EDGAR Pro Online search tool, market segments can be filtered, of which three categories connected to the automotive industry are available. At the same time, in the XBRL data set, companies are provided with much general information, including SIC codes that can be used for categorization. According to the list of codes provided by the SEC, six four-digit codes cover the automotive industry (and related services with the exception of retail) that were reviewed in the quarterly reports of 2018. A summary of publicly listed entities is presented in Table 1.

All entities listed on the New York Stock Exchange (NYSE), National Association for Securities Dealers Automated Quotations (NASDAQ) and Better Alternative Trading System (BATS) from the entire population are supposedly consistent data sources and regulated by the SEC. In addition to the variance in the number of listed entities in the automotive industry, the size of the entire population between the two sources is inconsistent and differs by over 24%. In terms of industrial classification, the taxonomy behind SIC codes in XBRL data sets is valid but incomparable to the customary EDGAR approach in the case of the identification of specific activities. Therefore, two additional categories were created to fit the measurement process; OEMs and suppliers from Tiers 1 & 2 (other automotive suppliers).

4.2 Errors in terms of the consistency and availability of samples

Listed entities from both data sources that are unmatched as a result of their supposedly consistent counterparts were found. Out of the sample sizes of 103 and 74, 50 companies are common in both which raises concerns over reliability.

Furthermore, data availability raised concerns in terms of search results from the SEC EDGAR Pro Online system. Out of the strong sample size of 103, 13 annual reports concerning 2018 were unavailable in the electric filing system of the SEC, while an additional 10 required data collection from official websites. Four financial statement items concerning the wealth and profitability of companies were selected for analysis in order to evaluate the differences between the two industrial classification schemes. The values of Total assets, Total Equity, Net sales revenue and Profit after-tax are central financial factors of investor decision-making. When necessary, exchange rates of the Federal Reserve were used according to the ASC (Accounting Standards Codification) standards issued by the Financial Accounting Standards Board (FASB) [35].

4.3 Comparison of financial information on an industrial level

Based on the financial statement data, descriptive statistics were calculated on the selected reporting lines. Differences were summarized in terms of both absolute values between the two data sources and percent deviations as presented in the Tables 2 and 3. A general observation of the data source is that the intervals between the minimum and maximum values are substantial for all four financial statement items. It is likely that – when used as a statistical sample – a normal distribution cannot be assumed. The standard deviation exceeds the mean values in the case of Total assets, therefore, the set of values (especially for the financial data of suppliers from Tiers 1 & 2) is highly dispersed.

A pattern can be observed in the deviation between the two data sources. The total values of OEM financial statement items are higher in the XBRL data set, in contrast to data derived from the online SEC source, while the opposite is seen in the case of suppliers from Tiers 1 & 2, where the total values are dominated by online

Table 1: Industrial specification of data sources

SEC's EDGAR Online Pro			SEC's EDGAR XBRL data set		
	Industrial specification	Number of companies	Industry (SIC)	Number of companies	
OEMs	Auto & Truck Manufacturers	26	3711. Motor Vehicles & Passenger Car Bodies	20	OEMs
Suppliers from Tiers 1 & 2	Automobiles, Parts & Service Retailers	24	3713. Truck & Bus Bodies	2	
	Auto, Truck & Motorcycle Parts	53	3714. Motor Vehicle Parts & Accessories	41	Suppliers from Tiers 1 & 2
			3715. Truck Trailers	1	
			3716. Motor Homes	2	
			3751. Motorcycles, Bicycles, and Parts	8	
	Total	103	Total	74	
	Entire population size	5,736	Entire population size	7,133	

sources. As an attempt to generalize the automotive industry, mean values were calculated where XBRL represents higher values except for the net sales revenues of suppliers. These deviations are partly validated by the amount of incompletely matched samples, but the 103:74 sample-size ratio is not represented by the results. The Table 4 summarizes the difference between the results of descriptive statistics in the form of percentages.

Despite former expectations, OEMs do not represent the majority of the financial item totals (between 45.1 and 55.9%), total equity (between 31.8 and 48.6%), net sales revenue (44.7 and 57.6%) and profit after-tax (35.4-52.5%), the differences between data sources can be measured on a scale of 6.7% to 31.9% as seen in Table 4. Suppliers from Tiers 1 & 2 match to an even lesser extent, so percent deviations are typically higher, especially in the case of net sales revenue (56.7%). Based on the matrix, the individual averages of companies cannot be used for industry generalization, both in terms of absolute mean values and standard deviations. The deviation "hotspots" are clearly centered around the suppliers from Tiers 1 & 2.

4.4 *Chi-square statistical testing*

To support our assumptions of statistically significant deviation between data sources, Pearson's chi-squared test was implemented, a full description of the steps is available in Appendix A [36, 37]. Selected categories of OEMs and suppliers from Tiers 1 & 2 were differentiated along with expected (data derived from online SEC-based financial statements) vs. observed (data derived from XBRL data sets) values. Based on the performed Chi-square test, the results highlighted that the differences between the expected and observed values of financial statement items (Total assets, Total Equity, Net sales revenue and Profit after-tax) were significant. With a 95% confidence interval (α = 0.05), OEMs and suppliers from Tiers 1 & 2 both exceeded the critical value of 16.92 with 7 degrees of freedom (df = 7). It is important to note the very significant (almost 10 times higher) impact of suppliers from Tiers 1 & 2 in terms of the total level of deviance.

5. Conclusions

XBRL preparation is obligatory, however, the content can include differences from the reported and published financial statements. Conclusions can be summarized in the following points:

- Potential duplication of lines in XBRL sources (e.g. 8 lines of certain financial statement items from China Automotive Systems, Inc.);
- Lack of standardization in tags: The XBRL platform manages to integrate more financial reporting taxonomy (different annual versions of IFRS and US GAAP). Due to the different (and potentially customized) tags, the definitions of some financial statement items converge; the structure of financial statements has yet to be fully harmonized between annual reports and XBRL statements;
- Errors in the reporting period (temporal differences): in some cases, outdated (1 or 2 years prior to

Table 2: Automotive market share of the entire population (%, number)

Representation % (n)	**SEC's EDGAR Pro Online**	**SEC's EDGAR XBRL data set**	
Entire population	100% (5,736)	100% (7,113)	
OEMs	0.45% (26)	0.31% (22)	
Suppliers from Tiers 1 & 2	1.34% (77)	0.73% (52)	
Total	**1.80% (103)**	**1.04% (74)**	
OEMs (matched)	0.28% (16)	0.20% (14)	Total assets Total equity Net sales revenue Profit after-tax
Suppliers from Tiers 1 & 2 (matched)	0.59% (34)	0.50% (36)	
Total (matched)	**0.87% (50)**	**0.70% (50)**	

Table 3: Absolute financial data from data sources (left – SEC's EDGAR Pro Online / right – SEC's XBRL data set) (USD in millions)

Statistics	**Category**	**Total assets value**		**Total equity attributable to company owners**		**Net sales revenues**		**Profit after taxes**	
Totals	OEM	⇩ 783.346	⇧ 874.139	⇩ 153.668	⇧ 221.111	⇩ 606.225	⇧ 649.422	⇩ 23.941	⇧ 35.125
	Tier 1&2 S.	⇧ 953.149	⇩ 688.812	⇧ 329.224	⇩ 234.365	⇧ 749.694	⇩ 478.399	⇧ 43.717	⇩ 31.834
Mean	OEM	⇩ 27.012	⇧ 39.734	⇩ 5.299	⇧ 10.050	⇩ 20.904	⇧ 29.519	⇩ 826	⇧ 1.597
	Tier 1&2 S.	⇩ 12.880	⇧ 13.246	⇩ 4.449	⇧ 4.507	⇧ 10.131	⇩ 9.200	⇩ 591	⇧ 612
St. dev	OEM	⇩ 66.884	⇧ 77.251	⇩ 11.802	⇧ 18.956	⇩ 45.992	⇧ 52.970	⇩ 1.875	⇧ 3.007
	Tier 1&2 S.	⇩ 56.749	⇧ 63.640	⇩ 21.527	⇧ 23.684	⇩ 33.241	⇧ 34.997	⇩ 2.971	⇧ 3.262

Table 4: Deviation matrix between data sources (%)

Statistic	Category	Total assets value	Total equity attributable to company owners	Net sales revenues	Profit after taxes
Totals	OEM	10.39%	30.50%	6.65%	31.84%
	T1&2 S.	38.38%	40.48%	56.71%	37.33%
Mean	OEM	32.02%	47.28%	29.18%	48.29%
	T1&2 S.	2.76%	1.29%	10.12%	3.50%
St.dev	OEM	13.42%	37.74%	13.18%	37.65%
	T1&2 S.	10.83%	9.11%	5.02%	8.93%

the current fiscal year) financial information is presented in current filings (e.g. an entity presents information from the 2017 fiscal year in the Q4 2018 filing as the most current);

- The inability to fully and feasibly automate data analysis in the case of automotive suppliers. Mean values are inconsistent between data sources due to the varying sample size of automotive suppliers. To perform a comprehensive industry analysis, error terms need to be defined clearly. Otherwise such an analysis would be performed with many predefined assumptions, leading to a decrease in the overall explanatory power and raising concerns about reliability/reproducibility.

Financial analysts should use XBRL datasets with concern, these points kept in mind. As a currently available best practice, the methodology of the U.S. Securities and Exchange Commission is a precedent for the building of inline XBRL statements into integrated datasets. An emerging challenge of regulatory bodies such as the European Securities and Markets Authorities is the supervision of companies uploading their data to a central system of a similar nature to produce well-structured databases for automated financial analytics.

Acknowledgements

The research presented in this paper was financed by the “Research Area Excellence Program – 2019 (TUDFO/51757/2019-ITM)” and University of Győr.

REFERENCES

[1] NAICS Association (2019). Common SIC Questions, https://www.naics.com/frequently-asked-questions/#NAICSfaq

[2] XBRL International (2019). An Introduction to XBRL, https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/

[3] Yaghoobirafi, K.; Nazemi, E.: An Approach to XBRL Interoperability Based on Ant Colony Optimization Algorithm. *Knowledge-Based Systems*, 2019, **163**, 342–357 DOI: 10.1016/j.knosys.2018.08.038

[4] Wenger, M. R.; Elam, R.; Williams, K. L.: A Tour of Five XBRL Tools: Products That Help Make Tagged Data Work for You and Your Clients. *Journal of Accountancy*, 2013, **215**(4), 48–55

[5] Radzimski, M.; Sanchez-Cervantes, J. L.; Garcia-Crespo, A.; Temiño-Aguirre, I.: Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data. *International Journal of Software Engineering and Knowledge Engineering*, 2014, **24**(5), 801–823 DOI: 10.1142/S0218194014500314

[6] Joyner, D.: How to Transform XBRL Data into Useful Information. *Review of Business and Technology Research*, 2017, **14**(1), 34–41

[7] Chychyla, R.; Leone, A. J.; Minutti-Meza, M.: Complexity of Financial Reporting Standards and Accounting Expertise. *Journal of Accounting and Economics*, 2019, **67**(1), 226–253 DOI: 10.1016/j.jacceco.2018.09.005

[8] SEC Reporting Taxonomy Technical Guide (version 2019), Financial Accounting Standards Board, Financial Accounting Foundation, 2019 https://www.fasb.org/jsp/FASB/Document_C/DocumentPage&cid=1176171806866

[9] Felo, A. J.; Kim, J. W.; Lim, J. H.: Can XBRL Detailed Tagging of Footnotes Improve Financial Analysts’ Information Environment? *International Journal of Accounting Information Systems*, 2018, **28**, 45–58 DOI: 10.1016/S0740-624X(98)90003-X

[10] Zhang, Y.; Guan, Y.; Kim, J. B.: XBRL Adoption and Expected Crash Risk. *Journal of Accounting and Public Policy*, 2019, **38**(1), 31–52 DOI: 10.1016/j.jaccpubpol.2019.01.003

[11] Shan, Y. G.; Troshani, I.; Richardson, G.: An Empirical Comparison of the Effect of XBRL on Audit Fees in the US and Japan. *Journal of Contemporary Accounting & Economics*, 2015, **11**(2), 89–103 DOI: 10.1016/j.jcae.2015.01.001

[12] Liu, C.; Luo, X. R.; Wang, F. L.: An Empirical Investigation on the Impact of XBRL Adoption on Information Asymmetry: Evidence from Europe. *Decision Support Systems*, 2017, **93**, 42–50 DOI: 10.1016/j.dss.2016.09.004

[13] Kaya, D.; Pronobis, P.: The Benefits of Structured Data Across the Information Supply Chain: Initial Evidence on XBRL Adoption and Loan Contracting of Private Firms. *Journal of Accounting and Public Policy*, 2016, **35**(4), 417–436 DOI: 10.1016/j.jaccpubpol.2016.04.003

[14] Comparing Economic Data, U.S. Economic Census Bureau, 2016 https://www.census.gov/programs-surveys/economic-census/guidance/historical-data.html

[15] Chychyla, R.; Leone, A. J.; Minutti-Meza, M.: Complexity of Financial Reporting Standards and Accounting Expertise. *Journal of Accounting and Economics*, 2019, **67**(1), 226–253 DOI: 10.1016/j.jacceco.2018.09.005

[16] Krishnan, J.; Press, E.: The North American Industry Classification System and its implications for accounting research. *Contemporary Accounting Research*, 2003, **20**(4), 685–717 DOI: 10.1506/N57L-0462-856V-7144

[17] Kahle, K. M.; Walkling, R. A.: The Impact of Industry Classifications on Financial Research. *Journal of Financial and Quantitative Analysis*, 1996, **31**(3), 309–335

[18] MSCI (2018). Global Industry Classification Standard (GICS§) https://www.msci.com/gics

[19] Kelton, C. M.; Pasquale, M. K.; Rebelein, R. P.: Using the North American Industry Classification System (NAICS) to Identify National Industry Cluster Templates for Applied Regional Analysis. *Regional Studies*, 2008, **42**(3), 305–321 DOI: 10.1080/00343400701288316

[20] Hrazdil, K.; Zhang, R.: The Importance of Industry Classification in Estimating Concentration Ratios. *Economics Letters*, 2012, **114**(2), 224–227 DOI: 10.1016/j.econlet.2011.10.001

[21] Hrazdil, K.; Trottier, K.; Zhang, R.: A Comparison of Industry Classification Schemes: A Large Sample Study. *Economics Letters*, 2013, **118**(1), 77–80 DOI: 10.1016/j.econlet.2012.09.022

[22] Tassey, G.; Brunnermeier, S. B.; Martin, S. A.: Interoperability Cost Analysis of the US Automotive Supply Chain. *Research Triangle Institute*, 1999, Report 7007-03

[23] ILO (2005). Automotive Industry Trends Affecting Component Suppliers, Report for Discussion at the Tripartite Meeting on Employment, Social Dialogue, Rights at Work and Industrial Relations in Transport Equipment Manufacturing. International Labour Organization, pp. 26–42

[24] Silver, D.: The Automotive Supply Chain, Explained. 2016

[25] Mena, C.; Humphries, A.; Choi, T. Y.: Toward a Theory of Multi-Tier Supply Chain Management. *Journal of Supply Chain Management*, 2013, **49**(2), 58–77 DOI: 10.1111/jscm.12003

[26] Masoud, S. A.; Mason, S. J.: Integrated Cost Optimization in a Two-Stage, Automotive Supply Chain. *Computers & Operations Research*, 2016, **67**, 1–11 DOI: 10.1016/j.cor.2015.08.012

[27] Thomé, A. M. T.; Scavarda, L. F.; Pires, S. R.; Ceryno, P.; Klingebiel, K.: A Multi-Tier Study on Supply Chain Flexibility in the Automotive Industry. *International Journal of Production Economics*, 2014, **158**, 91–105 DOI: 10.1016/j.ijpe.2014.07.024

[28] Jia, F.; Gong, Y.; Brown, S.: Multi-Tier Sustainable Supply Chain Management: The Role of Supply Chain Leadership. *International Journal of Production Economics*, 2019, **217**, 44–63 DOI: 10.1016/j.ijpe.2018.07.022

[29] Sauer, P. C.; Seuring, S.: Extending the Reach of Multi-Tier Sustainable Supply Chain Management – Insights from Mineral Supply Chains. *International Journal of Production Economics*, 2019, **217**, 31–43 DOI: 10.1016/j.ijpe.2018.05.030

[30] Sarkis, J.; Gonzalez, E. D. S.; Koh, S. L.: Effective Multi-Tier Supply Chain Management for Sustainability. *International Journal of Production Economics*, 2019, **217**, 1–10 DOI: 10.1016/j.ijpe.2019.09.014

[31] The Future of the Automotive Value Chain - Supplier Industry Outlook 2025, Deloitte, 2017

[32] Automotive News - North America top suppliers, Tenneco, 2019 https://www.autonews.com/assets/PDF/CA116090622.PDF

[33] Global Automotive Supplier Study 2018, Roland Berger, 2017 https://www.rolandberger.com/it/Publications/Global-Automotive-Supplier-Study-2018.html

[34] EDGAR Pro Online (2019) operated by Donnelley Financial, LLC, https://pro.edgar-online.com/

[35] Foreign Exchange Rates - H.10, Federal Reserve, 2019. https://www.federalreserve.gov/releases/H10/hist/

[36] Field, A.; Miles, J.; Field, Z.: Discovering Statistics Using R. Sage Publications, 2012, pp. 814–816 ISBN: 978-1-446-20046-9

[37] Agresti, A.: An Introduction to Categorical Data Analysis. John Wiley & Sons. 2018, pp. 34–40 ISBN: 978-0-471-22618-5

Appendix A – Chi-square test steps

1) Contingency table

Financial statement items		OEMs	Suppliers from Tiers 1 & 2
Total assets	Expected	783,346	953,149
	Observed	874,139	688,812
Total equity attributable to the owners of companies	Expected	153,668	329,224
	Observed	221,111	234,365
Net sales revenue	Expected	606,225	749,694
	Observed	649,422	478,399
Profit after-tax	Expected	23,941	43,717
	Observed	35,125	31,834
Total		**3,346,977**	**3,509,195**

2) H0 The financial values of the XBRL data source (observed values) are not significantly different from the values of the online SEC EDGAR Pro Online data source (expected values). Therefore, industry totals from the two sources are consistent.

3) Calculated marginal totals for the observed table

4) Expected value calculation - based on the specific financial statement item's proportion in the whole population

5) Degree of freedom

$\text{df} = (r-1)(c-1)$ $\text{df} = (8-1)(2-1) = 7$

6) Calculation of chi-square values

Financial statement items	$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$	OEMs	Suppliers from Tiers 1 & 2	
Total assets	E-O	-90,792	264,337	
	$(\text{E-O})^2$	8,243,257,799	69,874,243,884	
Total equity attributable to the owners of companies	E-O	-67,442	94,860	
	$(\text{E-O})^2$	4,548,480,637	8,998,332,516	
Net sales revenue	E-O	-43,197	271,295	
	$(\text{E-O})^2$	1,865,960,356	73,600,784,685	
Profit after-tax	E-O	-11,184	11,883	
	$(\text{E-O})^2$	125,073,800	141,206,159	Chi-square total
Chi-square		**4,417**	**43,490**	**47,907**

7) Determination of Significance level and Critical value:

significance level (alpha)	0.05
critical value	**16.92**

8) The chi square test result showed as the H0 hypothesis should be rejected with a 95% confidence interval (degree of freedom = 7).

The Chi-squared test results showed that the hypothesis H0 should be rejected.