Regensburger Applied Research Conference 2020

RARC 2020

July 31st, 2020

Organizer:

Ostbayerische Technische Hochschule Regensburg

Proceedings

Jürgen Mottok, Ingo Ehrlich, Anton Haumer (Eds.)

Proceedings of the Regensburg Applied Research Conference 2020

Regensburg, July 31st, 2020

DOI: **10.35096/othr/pub-641**

Organized by:

Ostbayerische Technische Hochschule Regensburg

Editors:

Jürgen Mottok, Ingo Ehrlich, Anton Haumer

Conference Chair:

Prof. Dr. Jürgen Mottok

Program Committee:

Prof. Dr. Ingo Ehrlich

Prof. Anton Haumer

Prof. Dr. Anton Horn

Prof. Dr. Alfred Lechner

Prof. Dr. Wolfgang Mauerer

Prof. Dr. Jürgen Mottok

Prof. Dr. Armin Sehr

FOREWORD

The Applied Research Conference which is held every year at another University of Applied Sciences in Bavaria is the main event for all students in the Master of Applied Research program. They come together to present their work in oral presentations and full papers, which are published in the proceedings, as well as poster presentations of the students after their 1st semester. For sure it is interesting for Professors and interested people to see the results of upcoming scientists and researchers.

Due to the Corona pandemic, this year it is not possible to organize the conference as usual in presence. We have to refrain from face-to-face discussions, having together a cup of coffee.

We as Professors at the OTH Regensburg wanted to give our students the chance to finish this semester successfully – despite all limitations due to the pandemic situation. Therefore we decided to organize the RARC 2020 – Regensburg Applied Research Conference 2020 – for the Master of Applied Research students of OTH Regensburg as an online conference. However, for a Technical University the situation is rather a challenge than a problem. Using a variety of online tools for teaching during this semester, we have enough experience to find a setup for RARC2020.

We received 28 submissions for full papers, which were peer reviewed and 26 of them were accepted – you will find them in this proceedings, and they will be presented orally on July 31st, 2020. Additionally, there are 22 posters which will be presented on the same day.

During the Plenary Opening Session, after a welcome by our President Prof. Dr. Wolfgang Baier we will have 3 Keynote speakers:

- Prof. Dr. phil. habil. Karsten Weber: Erkenntnistheorie für Ingenieure
- Prof. em. Georg Scharfenberg: 11 Years Master of Applied Research Alumnis
- Veronika Fetzer: Entrepreneurship

Each session has its own virtual meeting room where the students give their presentations, followed by discussions as usual. There are two oral sessions in parallel to a poster session. In the poster sessions, a virtual meeting room is preserved for each poster. The authors present their posters, waiting for visitors to have discussions about their work.

Walk around virtually and grab the chance to have interesting discussions with our upcoming scientists in our meeting rooms.

I wish you a pleasant and interesting Applied Research Conference 2020 – stay healthy!

Prof. Dr. Jürgen Mottok
Conference Chair

CONTENTS

SESSION A1

Uncertainty of musculoskeletal model predictions considering variances of moment arms

Amelie Jungtäubl
Laboratory for Biomechanics
Ostbayerische Technische Hochschule
(OTH) Regensburg
Regensburg, Germany
ajungtaeubl@gmx.de

Maximilian Melzner
Laboratory for Biomechanics
Ostbayerische Technische Hochschule
(OTH) Regensburg
Regensburg, Germany

Sebastian Dendorfer
Laboratory for Biomechanics
Ostbayerische Technische Hochschule
(OTH) Regensburg
Regensburg, Germany

Abstract— For motion analysis and musculoskeletal simulation a realistic model with lifelike anatomical structure is important. For measuring muscle moment arms frequently the tendon excursion method is used. The resulting data often shows a high deviation. The aim of this work is to evaluate the sensitivity of the joint reaction forces and the muscle activity towards this variances of the moment arms. In a first step the experimental set up for measuring the moment arms in the elbow with the tendon excursion method is described. In a second step the moment arms in the AnyBody Modeling System™ standing model were adjusted to the recorded data. In a second model the moment arms are increased to observe the sensitivity of the joint reaction forces towards higher changes. The results of the adapted, increased and default moment arms are compared. The changes between the three models are with a variance of maximal 4 N very small, except for the proximal-distal forces. The results indicate that the variances which occur while measuring with the tendon excursion method are acceptable.

Keywords— *Elbow joint, Musculoskeletal modelling, Tendon excursion method, Moment arms*

I. INTRODUCTION

An essential part of the musculoskeletal system are tendons. They transmit the forces generated by the muscles to the skeletal system in order to perform a movement. Furthermore they determine the line of action. In the interest of a realistic calculation of the occurring forces, simulating a correct anatomical structure is important. Hence, measuring the exact moment arms of the tendons is fundamental. According to Spoor et al [1] it is preferable to measure the tendon displacement as a function of joint angle instead of directly measuring moment arms. Therefore, frequently the tendon excursion method is used. Here the moment arm r can be brought into connection with the tendon excursion E and the joint rotation φ. While the method is commonly used, the results often show a big deviation.

The aim of this work is to evaluate the influence of these variances of the moment arms and thus the maximal muscle activity as well as the joint reaction forces.

II. MATERIALS & METHODS

In a first step an experiment is performed to calculate muscle moment arms along the elbow joint during extension/flexion. Therefore a cadaveric left arm (age 65+, gender unknown) is used which is fixed with ethanol and stored at 4°C. Before the experiment it is stored at room temperature for one hour to assure full flexibility for the measurement. The hand was cut off and a 0.7mm aramid string was sewn on the distal end of each measured tendon. Then the arm was screwed to a board in order to ensure a stable position (see Fig. 1). The distal end of the aramid string was attached to a 5 N weight to ensure a pre-tensioning of the muscles. Following muscles are measured:

- extensor carpi radialis brevis (ECRB)
- extensor carpi ulnaris (ECU)
- extensor digitorum (ED)
- flexor carpi radialis (FCR)
- flexor carpi ulnaris (FCU)
- palmaris longus (PM)

For measuring the excursion of a muscle, its string was lead through a pulley which was secured in a vice. The vice was attached to the table in a way, so the tendons remain in a realistic range of motion. During the experiment, the arm was flexed three times in a pronating position for each tendon in a range between 0° and maximum 90°. In order to measure the tendon excursion and the joint angle, a 12 - camera motion capture system (Vicon Motion Systems Ltd., Oxford, England) and a set of reflecting markers are used (see Fig. 2 and 3). The reflecting markers are attached to the arm with 1.2x3mm countersunk screws to obtain the excursion of the tendon and flexion angle of the elbow during the experiment (see Fig. 3). They are also attached to the weight in order to calculate the tendon excursion. There are three markers each on the weight, the upper and lower arm in order to get a determined model (see Fig. 4).

For calculating the joint angle and determining the correct joint center location the Vicon Nexus "SCoRe" (Symmetrical Center of Rotation Estimation) and "SARA" (Symmetrical Axis of Rotation Analysis) operations are used. SCoRE is an optimization algorithm to estimate the center of point rotation [2].

Figure 2: Experimental set-up to obtain the tendon excursion of the muscles

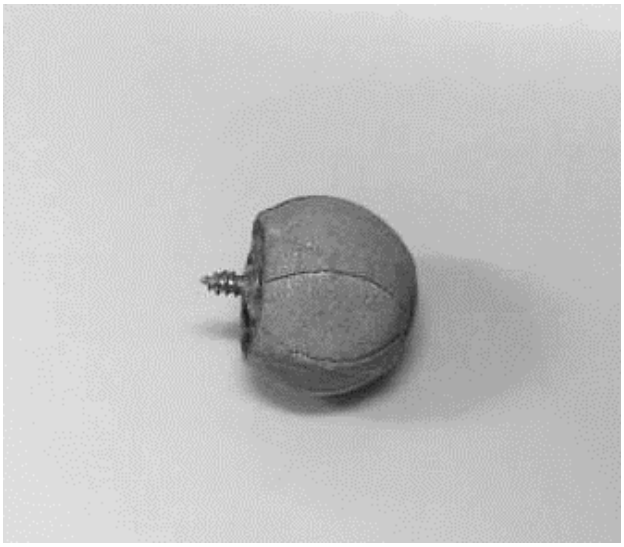


Figure 3: Marker with screw to attach to muscle

With the optimization algorithm SARA the axis of rotation can be estimated.

For the evaluation of the sensitivity the AnyBody Modeling System™ v.7.2 ((64-bit version) AnyBody Technology A/S, Aalborg, Denmark) AMMR (AnyBody Managed Model Repository) Standing model is used. The legs and the right arm are omitted in order to shorten the calculation time. The arm is adjusted to a pronating position in the same way it was in the experiment. The wrapping obstacles of the muscles are varied as well as the insertion points and the via points in order to see the influence on the moment arm. The change of the coordinates of the insertion and via points are kept small in a range between 1mm and 20mm.

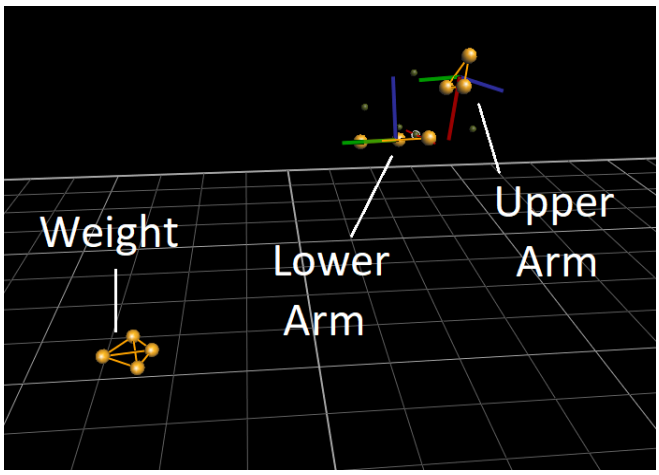


Figure 4: Vicon model of the arm and the weight

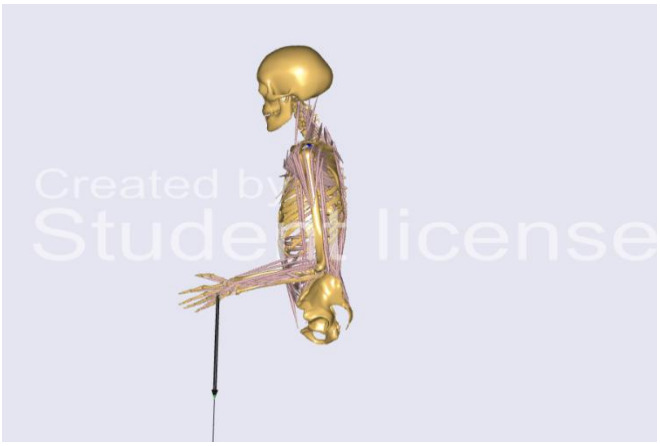


Figure 5: AnyBody™ Standing Model with force applied to the palm

The moment arms were adjusted in a way to match with the data from the experiment (see Fig. 6 – 11). In a second model the moment arms are increased to evaluate the sensitivity of the joint reaction forces towards bigger changes. The difference between the adapted and the default moment arms is up to 10mm, between the adapted and the increased moment arms there is a maximum change of 19mm. Afterwards a force of 10 N along the gravity is applied on the palm and a flexion from 0° to 90° is executed with default, increased and adapted moment arms (see. Fig. 5).

With the resulting differences of the joint reaction forces the sensitivity of the calculations towards the moment arms can be evaluated. Therewith the issue of the high deviation of the tendon excursion method can be assessed.

III. Results

In the appendix in Figure 6 to 11 there are the moment arms of the different adjustments for ECRB, ECU, ED, FCR, FCU and PM. The graphs show the measured data of the experiment with the tendon excursion method (‘Measure TEM’), the aligned moment arms (‘Adapted’), the increased moment arms (‘Increased’) and the preset moment arms of the AnyBody™ standing model (‘DefaultStandingModel’). The figures 12 to 17 show the

- humeroulnar anterior-posterior force (HUAPF)
- humeroulnar medio-lateral force (HUMLF)
- humeroulnar proximal-distal force (HUPDF)
- radiohumeral proximo-distal force (RHPDF)
- humeroulnar axial moment (HUAM)
- humeroulnar lateral moment (HULM)

	Default	Adapted	Increased
HUAPF	0.7 – 20.0	1.0 – 20.0	4.2 – 20.5
HUMLF	-4.0 – 3.3	-3.3 – 5.3	-6.5 – 3.0
HUPDF	-90.3 – -28.4	-91.6 – -28.1	-77.8 - -16.3
RHPDF	-150.0 – 8.1	-143.1 – 7.4	-185.9 – 7.3

Table 1: Values of forces in [N] for the three different models

	Default	Adapted	Increased
HUAM	-0.08 – 0.39	-0.07 – 0.39	-0.31 – 0.29
HULM	0.64 – 1.31	0.51 – 1.33	0.44 – 1.16

Table 2: Values of the axial and lateral moments in [Nm] for the three different models

In the HUAPF, HUMLF, HUAM and HULM the three values are all very similar and show no notable change. In the HUPDF the adapted moment arms and the default moment arms are almost identical. The increased moment arms show an increased force of about 10%. In the RHPDF also the adapted moment arms and the

default moment arms show an identical outcome of the force. The increased moment arms show a decrease of the force of about 20%.

Figure 18 shows the maximal muscle activity, which is for all three models between 0.22 and 0.24.

IV. Discussion and Limitations

A. Moment arms

With the experiment moment arms of muscles were measured, which were not regarded previously in the literature. All of the observed muscles are not main flexors of the elbow. Therefore, the maximal muscle activity is not affected by a change of the moment arms and the changes in the forces are rather small. The adapted muscle arms show a large correspondence with the default AnyBody™ Standing Model. The increased moment arms, which are showing the sensitivity of the forces towards higher changes of the moment arms, don't result in a change of the values except for the HUPDF and the RHPDF which indicates a small influence of these muscle moment arms on the elbow joint reaction forces. As the change of the wrapping obstacles and the via points did not show a large influence of the moment arms, the main adaption of them was carried out through the origin points.

B. Limitations of the tendon excursion method

The movement of the arm was executed manually, so the velocity is not steady. As the aramid string is not stiff, there might be a movement of the weight, which is not based on the flexion or excursion of the arm but for example on the rotation of the weight.

C. Fixation

According to Kim et al. [3] usually the alteration of cellular components also affects proteins, the membrane and other intracellular structures. So due to the fixation with ethanol the mechanical properties of the arm have changed. Therefore, the excursion of the tendons is different to the in vivo excursion. Furthermore, the flexibility of the joints is limited. The extensor carpi radialis longus can't be measured, because it ruptured. As there is no surrounding tissue the movement of the tendons changes. During the extension of the fingers, the tendon loses the contact to the bone. Therefore, the tendons are attached to a 5 N weight to keep them strained.

D. Measurement of the deep muscles

As the motion capture system is an optical measurement it is not possible to measure the deep muscles as the markers would be covered by the superficial muscles.

V. Conclusion

In a next step the observation of the main flexors of the arm is interesting as the variances of the joint reaction forces are not very notable in this work. The variances of the moment arms, which occur while measuring with the tendon excursion method, are acceptable regarding the presented outcome.

References

[1] C. W. Spoor, J. L. van Leeuwen, C.G.M. Meskers, A. F. Titulaer, and A. Huson. Estimation of instantaneous moment arms of lower-leg muscles. Journal of biomechanics, 23(12):1247–1259, 1990.

[2] Vicon Motion Systems, "About ScoRE and SARA in Vicon Nexus", Vicon Documentation, [Online] Available: https://docs.vicon.com/pages/viewpage.action?pageId=508888 67.

[3]S.-O. Kim, J. Kim, T. Okajima, and N.-J. Cho, "Mechanical properties of paraformaldehyde-treated individual cells investigated by atomic force microscopy and scanning ion conductance microscopy," Nano convergence, vol. 4, no. 1, p. 5, 2017.

APPENDIX

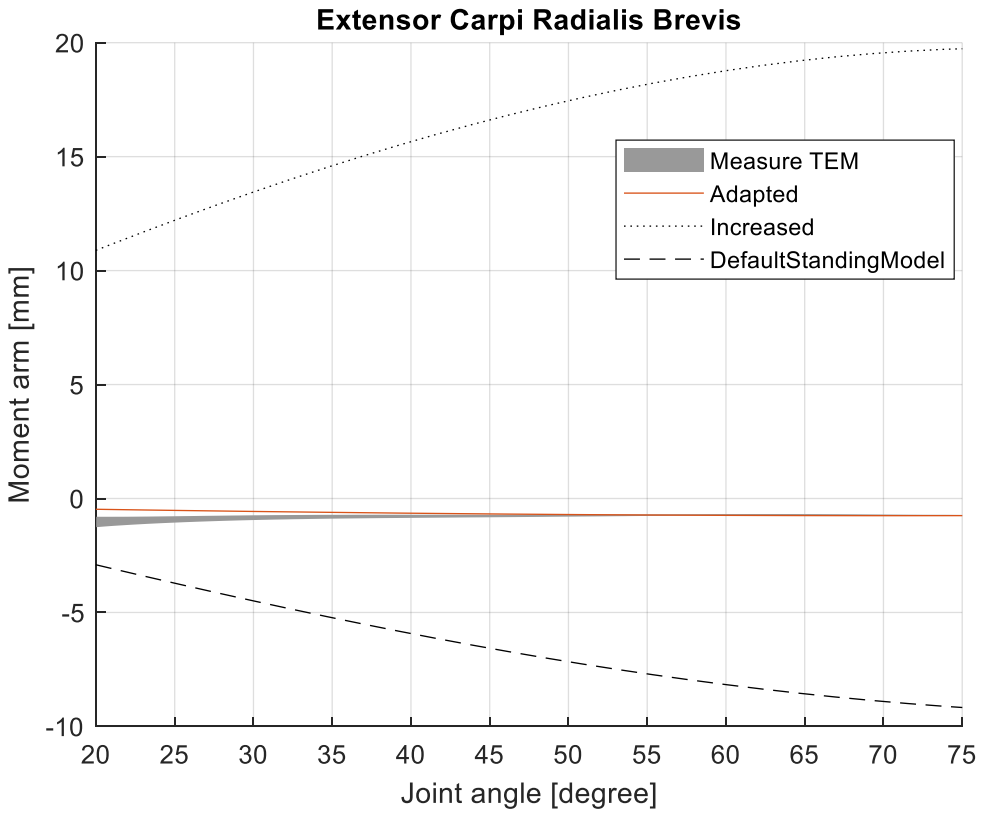


Figure 6: Moment arms of the ECRB

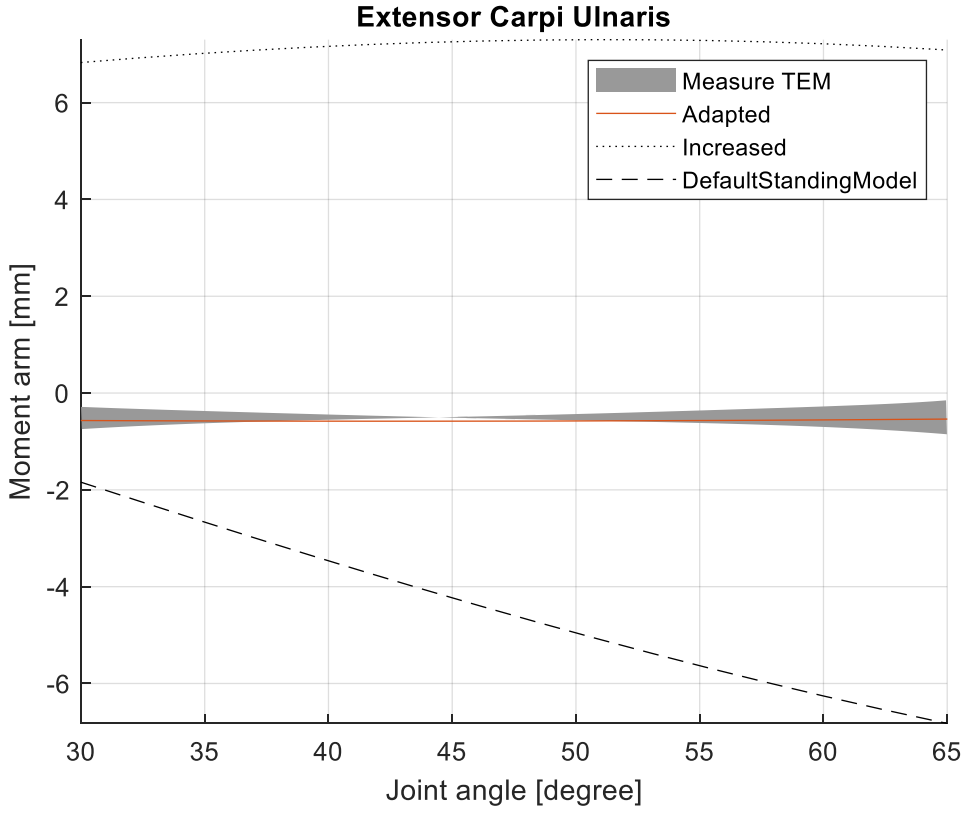


Figure 7: Moment arms of the ECU

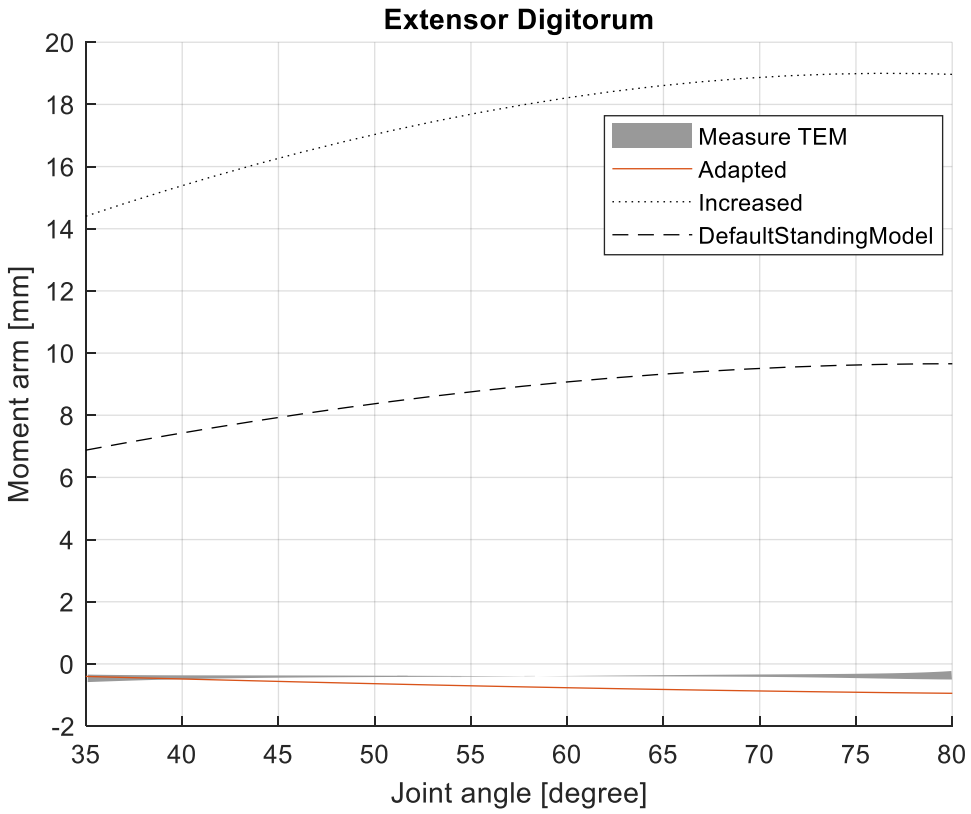


Figure 8: Moment arms of the ED

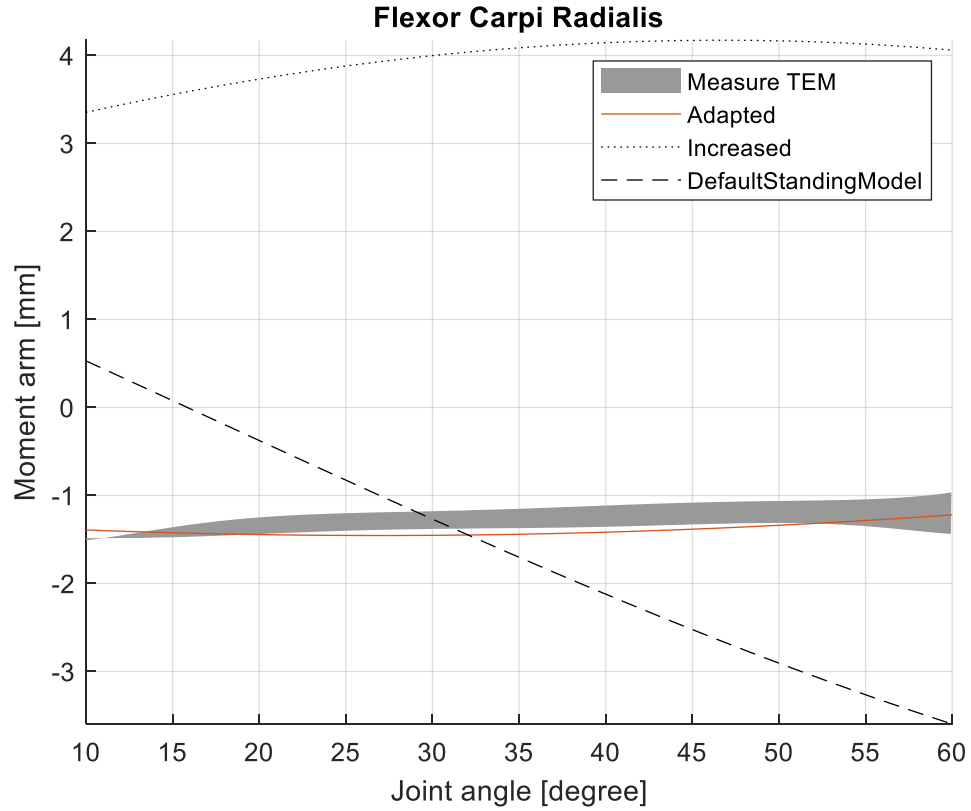


Figure 9: Moment arms of the FCR

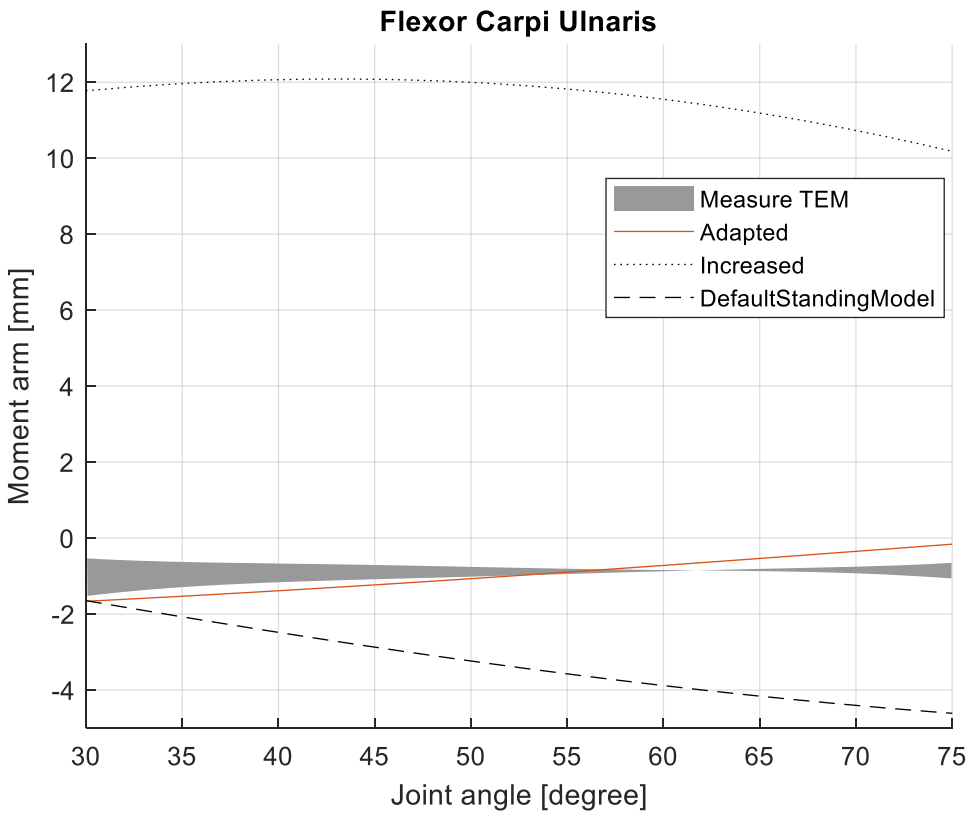


Figure 10: Moment arms of the FCU

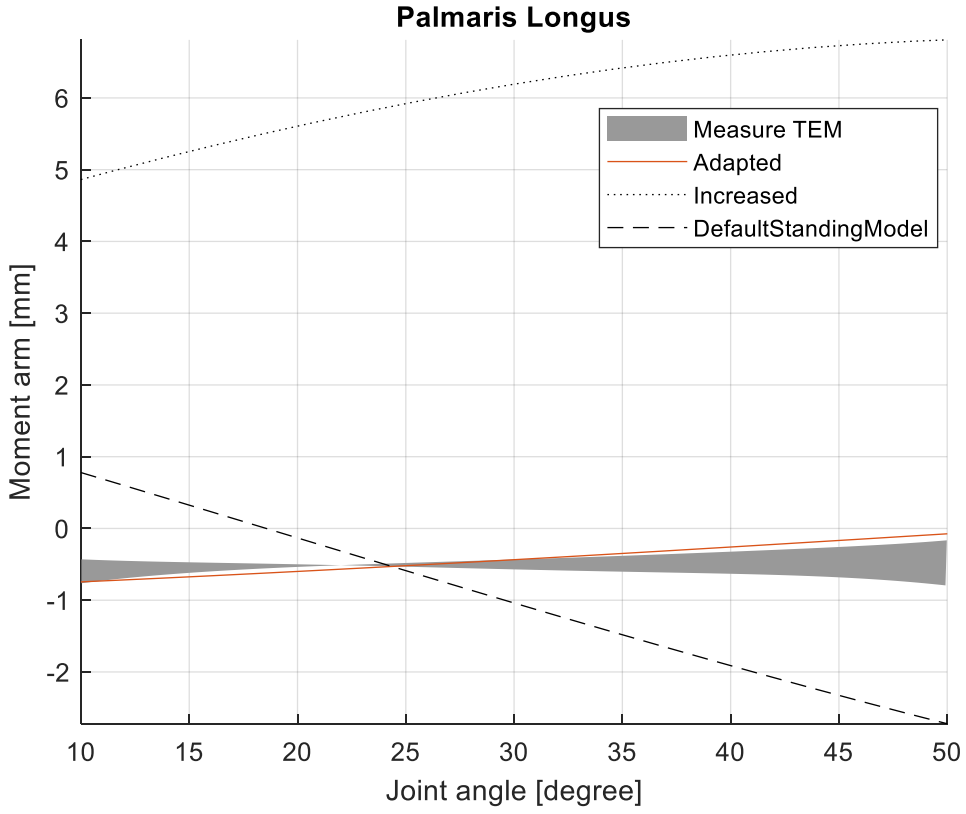


Figure 11: Moment arms of the PL

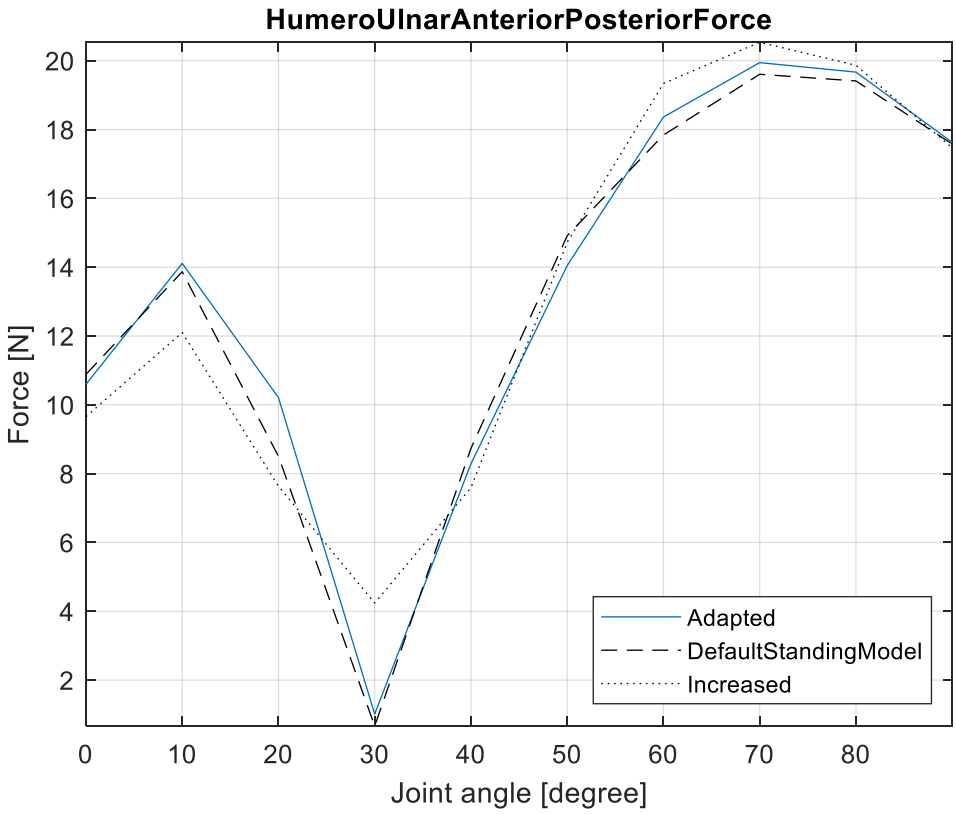


Figure 12: Humeroulnar anterior-posterior force of the three different models

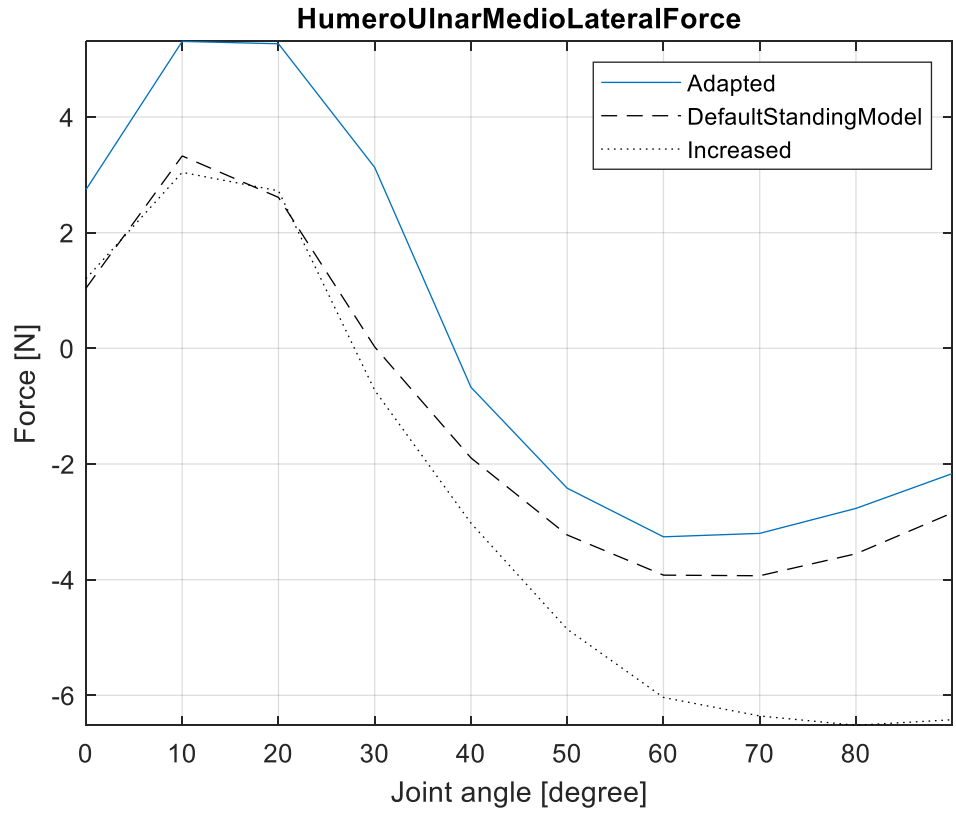


Figure 13: Humeroulnar medio-lateral force of the three different models

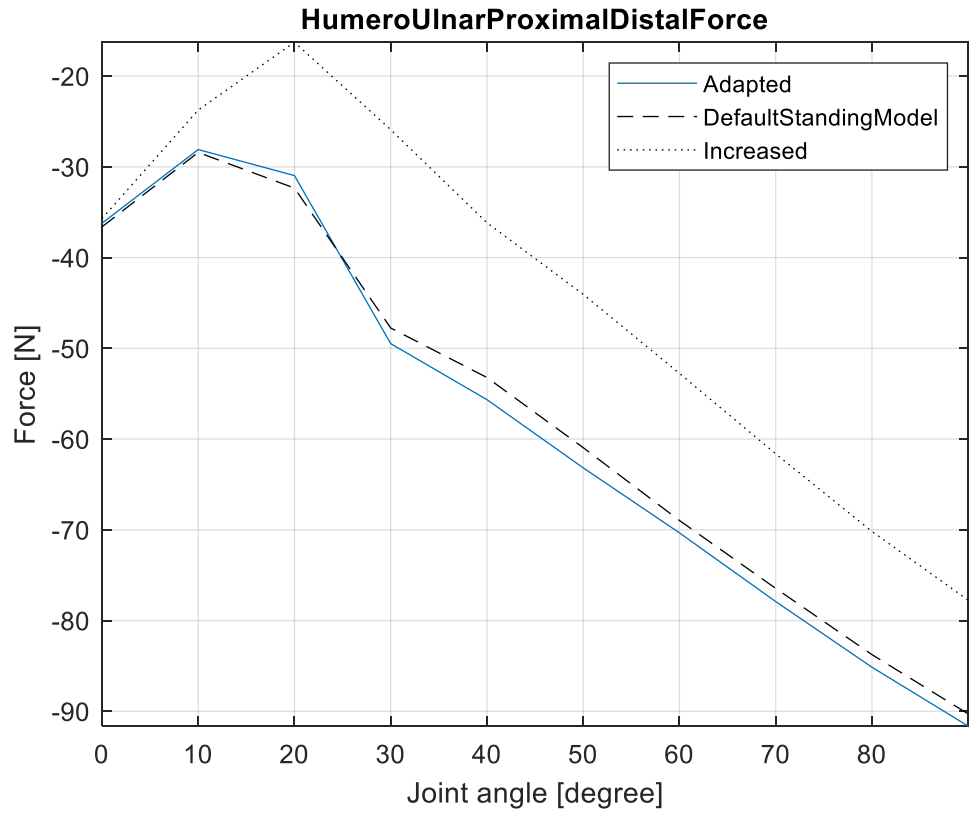


Figure 14: Humeroulnar proximal-distal force of the three different models

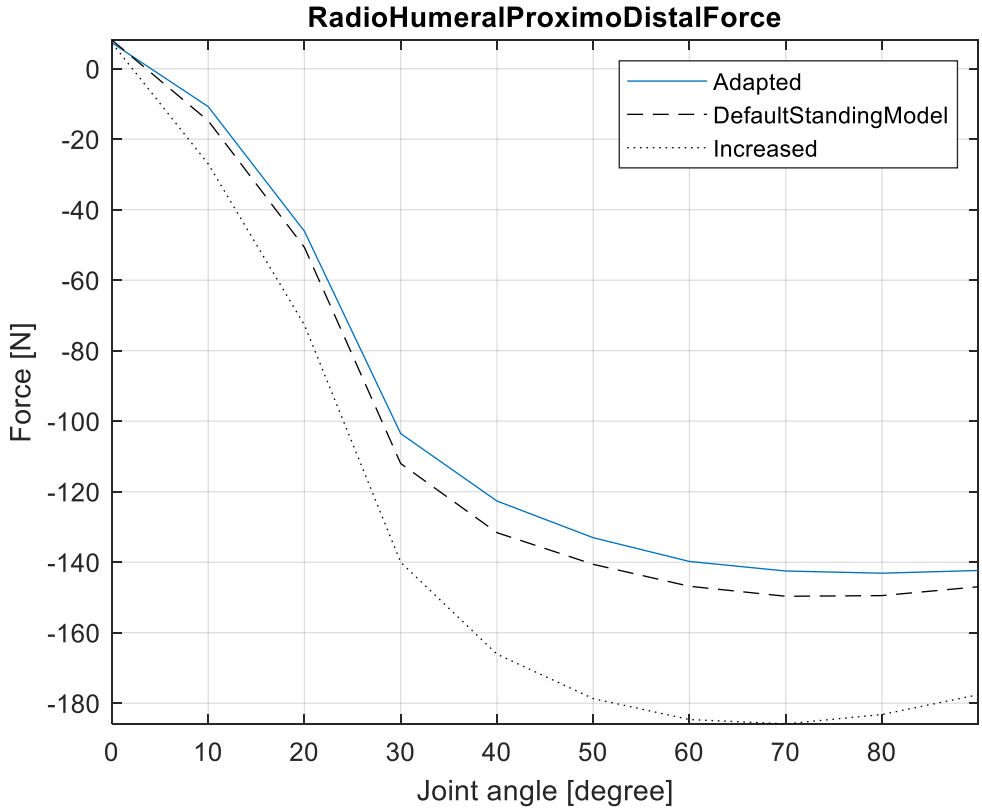


Figure 15: Radiohumeral proximo-distal force of the three different models

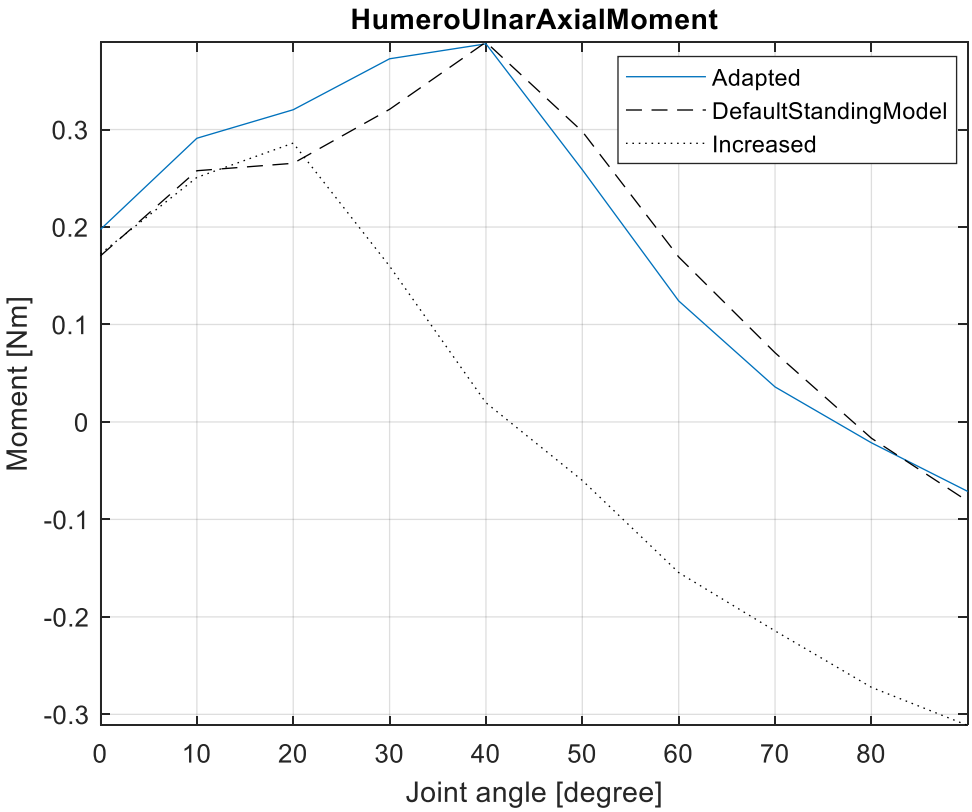


Figure 16: Humeroulnar axial moment of the three different models

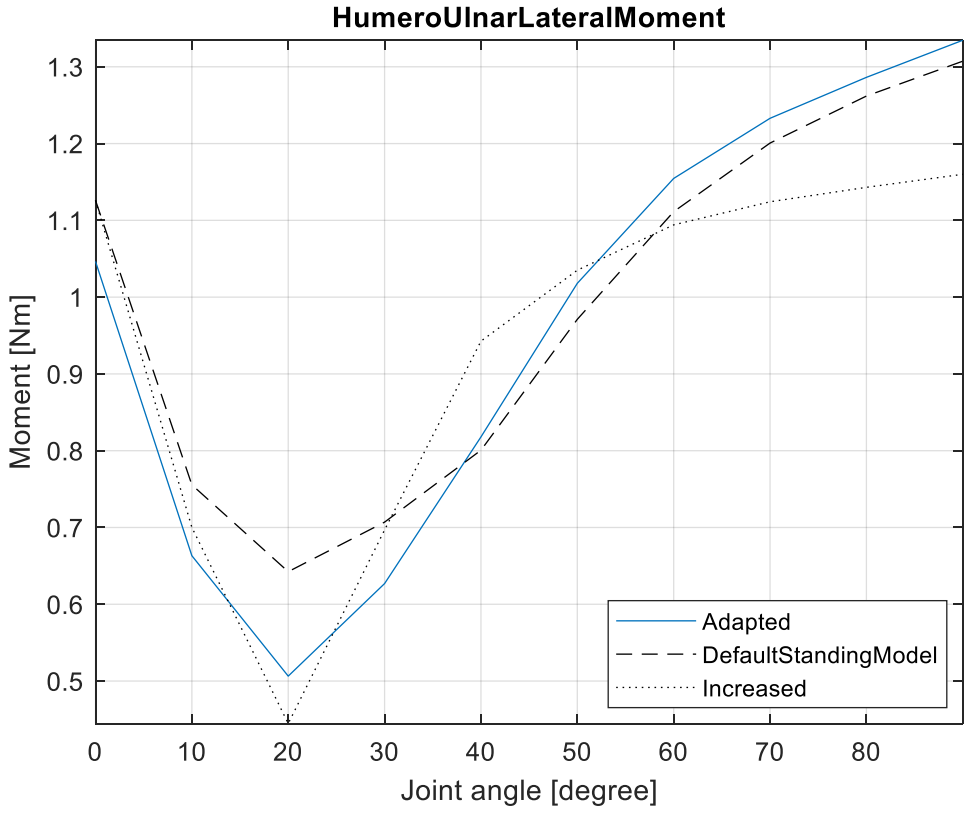


Figure 17: Humeroulnar lateral moment of the three different models

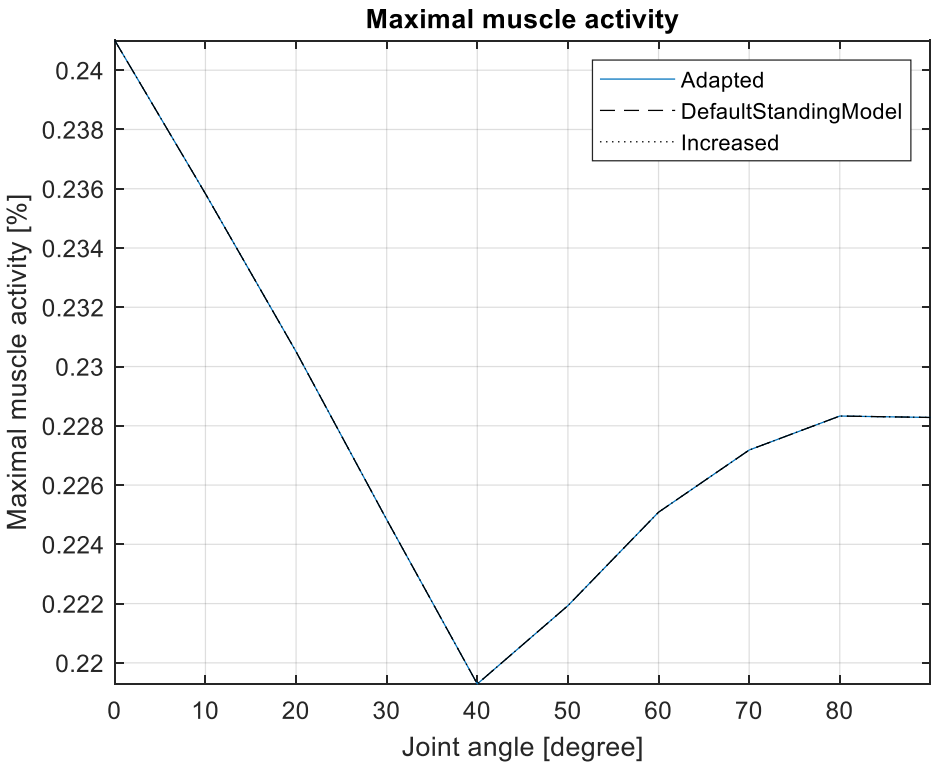


Figure 18: Maximal muscle activity of the three different models

LIST OF FIGURES

LIST OF TABLES

Time-Continuous Simulation of the Field Oriented Control of a Wheel Hub Motor

Daniel Gottschlich
Ostbayerische Technische Hochschule Regensburg
Elektro- und Informationstechnik
Regensburg, Germany
daniel1.gottschlich@st.oth-regensburg.de

Bernhard Hopfensperger
Ostbayerische Technische Hochschule Regensburg
Elektro- und Informationstechnik
Regensburg, Germany
bernhard.hopfensperger@oth-regensburg.de

***Abstract*—Electrical Wheel Hub Motors are considered as a key component for future mobility. Since space inside of the wheel is limited, Wheel Hub Motors are built as permanent-magnet synchronous motors. In industry Field Oriented Control is used to control these machines. The first step in the development process of this control is a time-continuous simulation. When the machine should also work in high speed ranges, the Field weakening operation has to be implemented in the simulation. This paper shows how Field Oriented Control can be simulated with Simulink.**

Keywords—Field Oriented Control, Wheel Hub Motor, Electromobility, Field weakness, Simulink

I. INTRODUCTION

Before an engine control is implemented at the test bench, it is advisable to set up a simulation that is as realistic as possible. On one hand side, tests can be carried out without the risk of destroying mechanical components, and on the other hand, different quantities can be displayed over time, which cannot be recorded with sensors on the test bench.

The core of every engine control is the model of the motor. In field-oriented control, appropriately transformed input and output variables for voltage and current have to be used. In drive technology, it is advisable to set up a cascaded control. The current control loop is cascaded into the speed control loop.

In order to operate the machine at speeds above the nominal speed, the field weakening range must be implemented. In addition, other measures can be incorporated into the model to improve performance. These include an anti wind up measure, a decoupling network and a speed filter.

II. WHEEL HUB MOTOR

Since space is very limited inside of the wheel of a vehicle, wheel hub drives are usually built in the form of permanent-magnet synchronous motors. These have a higher power density than induction machines. The permanent-magnet synchronous machines also have better efficiency in the low speed range, which is particularly advantageous for vehicles of a lower performance class. The motor considered in this paper was developed by E-Motiontech, see Figure 1. It is a symmetrical, permanent-magnet synchronous machine, which means that the direct and quadrature axis inductance are equal. This simplifies the relationship between torque and current.

Figure 1: Wheel hub motor from E-Motiontech

III. MODELING

Modeling the controlled system is the first important step. This is based on the physical equations of the individual components. For the motor, the stator voltage equation,

$$u_S = R_S \cdot i_S + \frac{d\psi_S}{dt} \tag{1}$$

is first taken and divided into direct and quadrature axis components:

$$u_d = R_S \cdot i_d + L_d \frac{di_d}{dt} - \omega L_q i_q \tag{2}$$

$$u_q = R_S \cdot i_q + L_q \frac{di_q}{dt} + \omega(L_d i_d + \psi_{PM}) \tag{3}$$

If the parts of the two equations linked to the angular frequency are omitted, two PT1 terms result, see (4) and (5). Then the transfer function of the motor winding for a symmetrical machine can be defined, see (6).

$$\frac{L_1}{R_S} \cdot \frac{di_d}{dt} + i_d = \frac{1}{R_S} \cdot u_d \tag{4}$$

$$\frac{L_1}{R_S} \cdot \frac{di_q}{dt} + i_q = \frac{1}{R_S} \cdot u_q \tag{5}$$

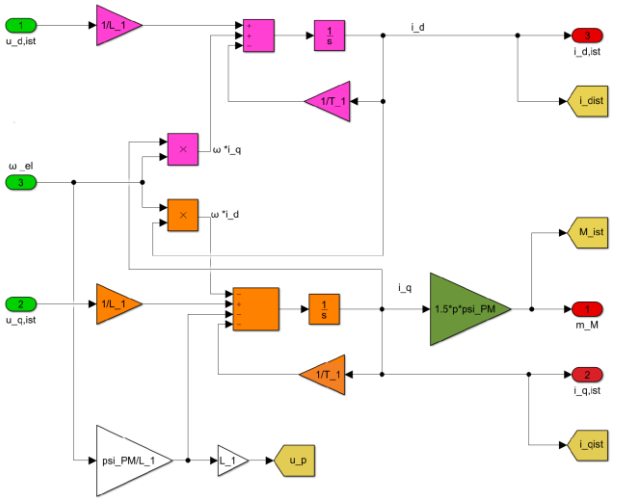


Figure 2: Model of the motor in Simulink

$$G_W(s) = \frac{i_{out}}{u_{in}} = \frac{i_d}{u_d + \omega L_q i_q} = \frac{i_q}{u_q + (\omega_{PM} - \omega L_d i_d)} = \frac{K_W}{1 + s \cdot T_W} \tag{6}$$

$$T_W = \frac{L_1}{R_S} \tag{7}$$

$$K_W = \frac{1}{R_S} \tag{8}$$

With a symmetrical machine, the motor torque is directly proportional to the quadrature axis current component.[1]

$$M_i = \frac{3}{2} p \cdot \psi_{PM} \cdot i_q \tag{9}$$

With the mechanical equation of motion, the speed can be calculated depending on the motor torque, the load torque and the total moment of inertia:

$$M_i - M_L = \frac{J}{p} \cdot \frac{d\omega}{dt} = J \cdot 2\pi \cdot \frac{dn}{dt} \tag{10}$$

The inverter and the measuring device for the currents can approximately be described as a first-order delay element. Since the time constants of the two components are very low compared to the other system components, these can be summarized. [2]

IV. Operation Ranges

In general, two operating ranges occur in electrical machines, the basic speed range and the field weakening range. Both are determined by the current and voltage limits.

A. Current and Voltage Limitation

The stator current can be represented as a phasor. The length can be calculated using the Pytagoras theorem and the direct and quadrature axis current. The same applies to the voltage:

$$|\underline{i_S}| = \sqrt{i_d^2 + i_q^2} \leq i_{S,max} \tag{11}$$

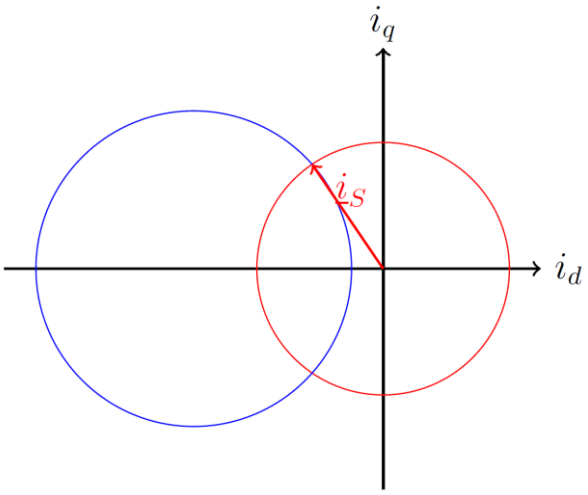


Figure 3: Limitation of current and voltage represented as circles

$$|\underline{u_S}| = \sqrt{u_d^2 + u_q^2} \leq u_{S,max} \tag{12}$$

If you insert (1) and (2) in (12), you get an equation for the voltage limitation depending on the current components. If the voltage drop across the stator resistor and the time-dependent variables are neglected, the following results:

$$\left(\frac{u_{S,max}}{\omega L_1}\right)^2 = \left(i_d + \frac{\psi_{PM}}{L_1}\right)^2 + i_q^2 \tag{13}$$

Equations (11) and (13) can be represented as circles in a d,q-coordinate system, see figure 3. The current phasor has to remain within both circles in order to maintain both the current and the voltage limitation.

B. Rated speed range

The voltage limitation is irrelevant in the nominal speed range. Since the direct axis current component does not matter in a symmetrical machine, only a quadrature axis current component is impressed. How large this proportion of current should be, depending on the desired torque, can be calculated using (9). The limit for the torque is finally determined by the current limit according to (11). This torque control method is called Maximum Torque per Ampere Control (MTPA). [3]

C. Field weakening area

In the field weakening range, both the current and voltage limits have to be taken into account. The angular frequency from which on the field weakening has to be implemented can be calculated with (14). [4] The field must be weakened in order to be able to further increase the speed when the nominal speed is reached. This can be achieved by impressing a negative direct axis current component. However, the maximum possible quadrature axis current component is also reduced, see (11). The equation (15) can be used to calculate how large the required direct axis current component is as a function of the speed. This torque control method is called Maximum Power Control (MP). [3]

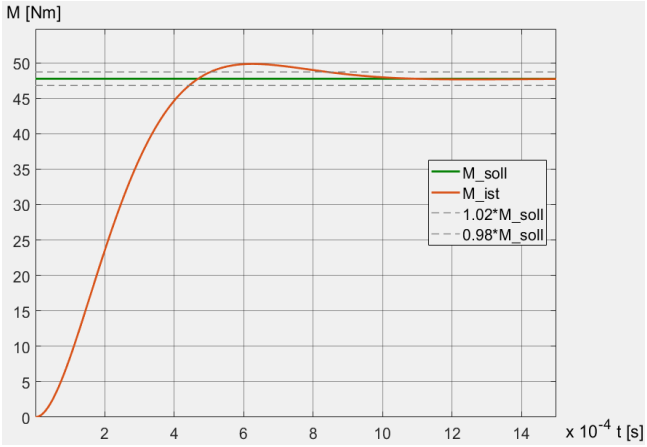


Figure 4: Step response of the torque control loop

$$\omega_n = \frac{u_{S,max}}{\sqrt{(L_d i_d + \psi_{PM})^2 + (L_q |i_q|)^2}} \tag{14}$$

$$i_d = \frac{-i_{S,max}^2 \cdot L_1 - \frac{\psi_{PM}^2}{L_1} + \frac{u_{S,max}^2}{\omega^2 \cdot L_1}}{2\psi_{PM}} \tag{15}$$

In some cases, a third torque control procedure is necessary. This is called Maximum Torque per Voltage Control (MTPV). It is required if the condition in (16) is fulfilled. [3] There are different approaches, from which speed to change from MTPA to MTPV and how the direct axis current component is calculated.

$$\frac{\psi_{PM}}{L_d} = |i_{S,max}| \tag{16}$$

D. Stationary operation

In stationary operation, the machine is fed with three-phase voltage of constant frequency and amplitude and a constant internal torque is generated at a constant speed. Under these circumstances, the voltage equations are simplified. [1] With the help of stationary operation, the simulation values can be compared rather easily with the calculated values.

V. Controller Design

In a cascaded control, the torque control loop is cascaded to the speed control loop.

A. Torque control loop

Due to the proportionality of torque and current, the torque control loop is a current control loop. Since the current in the field-oriented control is divided into direct and quadrature axis current, two current controllers are required. PI controllers are used for this, which are designed according to the avsolute optimum. This enables the best possible command action.

The speed-dependent magnet wheel voltage acts like a disturbance variable in the torque control loop. This can be almost completely compensated for using a decoupling network.

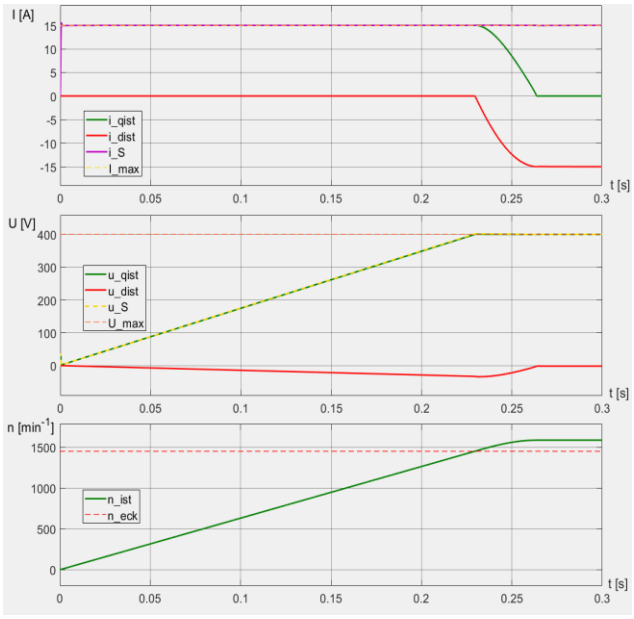


Figure 5: System values by increasing the speed

In addition, a module must be integrated in the chain of effects in front of the controllers, which enables operation in the field weakening area. This must monitor the speed and, if the nominal speed is exceeded, adjust the setpoints for the direct and quadrature components.

A voltage limiter has to be implemented. For this the actual DC voltage has to be known. The voltage limiter can be implemented by limiting the direct and quadrature axis current so that the current pointer remains within the circle, which is described by equation (13). In this case the limiter is in the signal chain in front of the PI controller. Therefore no anti-windup measure is required.

B. Speed control loop

A PI controller is also used to control the speed. In this case, however, this is parameterized according to the symmetrical optimum, which causes the best possible interference behavior.

Due to the limited values for current, an anti-windup measure has to be implemented in the speed controller. This can be achieved by switching off the Integral component of the controller when the limit values are exceeded.

VI. Results of the Simulation

The step response analysis is a popular method of evaluating control systems. A comparison with literature can be made by determining the rise and settling time. You can also check the effects of different measures, such as the decoupling network.

A. Torque control loop

To test the torque control loop, the torque is controlled directly. In the case of a step response, see figure 4, the rise and settling time and the overshoot can be used as comparison values.

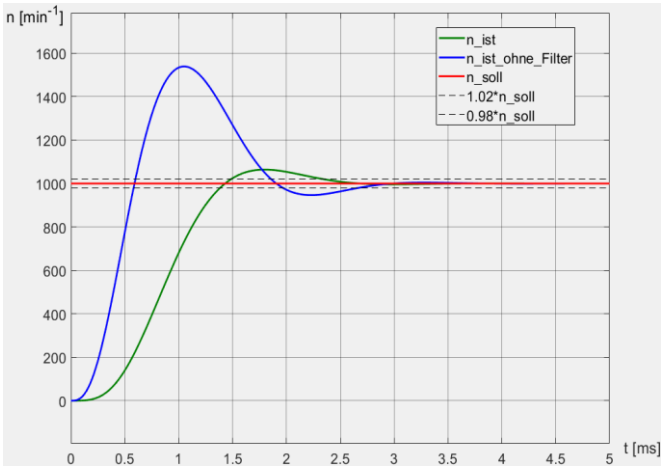


Figure 6: Step response of the speed control loop with and without filter

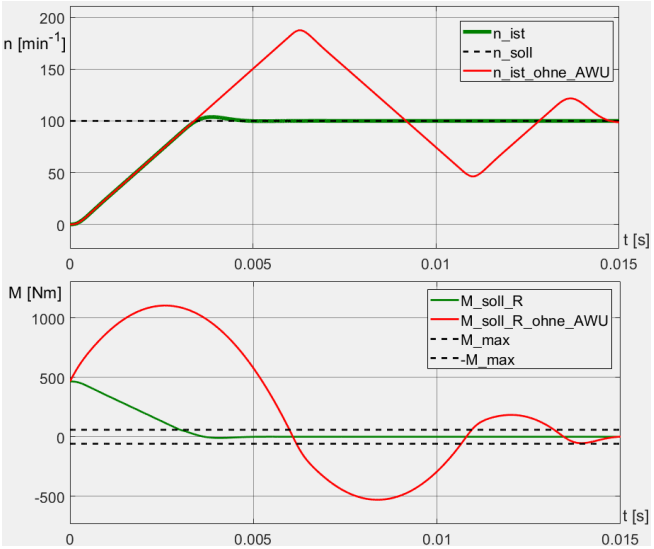


Figure 7: Step response of the speed control with and without anti windup measure

To determine the impact of the decoupling network, it must be simulated over a longer period. This is particularly effective when the speed changes quickly. This is the case if, for example, the inertia in the drive system is low. In this case, there is a high control deviation in the torque control loop without a decoupling network.

The torque control methods can be tested by specifying a maximum torque as a setpoint over a long period of time, see figure 3. While the quadrature axis current in the basic speed range corresponds to the maximum permitted current and the direct current is zero, the direct current according to (15) and the quadrature current according to (11) can be reduced depending on the speed. This also reduces the machine's torque. If the whole works properly, the terminal voltage remains constant in the field weakening mode.

B. Speed control loop

Relatively large time constants are noticeable in the step response of the speed control loop, see figure 6. Also noticeable is the large overshoot, which can be reduced by a PT1 filter.

The effect of the anti windup measure can be seen in Figure 7. Limiting the current means that the maximum possible torque is also limited. If the setpoint for the speed changes suddenly, the PI controller wants to set a high torque in order to quickly compensate for the control difference. If this torque exceeds the limit value, the I component in the controller leads to a further increase in the controller output signal. Ultimately, this leads to an enormous swing. The anti windup measure can prevent this by converting the PI controller into a P controller if the limit is exceeded.

VII. Conclusion and Future Outlook

The development of a continuous-time model is the first important step in the development of a field-oriented control. The model can also be used as the basis for further work. An important factor is to first deal with the various components and understand the models from the literature. Only when it is clear how the system should ideally behave, you can look for sources of error.

In the next step, the model has to be adapted to reality. For this, it must be expanded to a time-discrete control model. On one hand, the sensors only record their values at certain times and, on the other hand, the power electronics and the microcontroller also work according to a certain cycle.

References

[1] J. Teigelkötter, *Energieeffiziente elektrische Antriebe: Grundlagen, Leistungselektronik, Betriebsverhalten und Regelung von Drehstrommotoren; mit 4 Tabellen*. Wiesbaden: Vieweg+Teubner Verlag, 2013. [Online]. Available: http://dx.doi.org/10.1007/978-3-8348-2330-4

[2] D. Schröder, *Elektrische Antriebe – Regelung von Antriebssystemen*, 4th ed. Berlin and Heidelberg: Springer Vieweg, 2015. [Online]. Available: http://dx.doi.org/10.1007/978-3-64

[3] K. H. Nam, *AC Motor Control and Electrical Vehicle Applications*, 2nd ed. Milton: Chapman and Hall/CRC, 2018. [Online]. Available: https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5570774

[4] U. Nuß, *Hochdynamische Regelung elektrischer Antriebe*, 2nd ed. Berlin and Offenbach: VDE VERLAG GmbH, 2017

Secure Software Update of a Secure Module in the Power Grid

Tom Inderwies, Jürgen Mottok
Laboratory for Safe and Secure Systems – LaS^3
Technical University of Applied Sciences Regensburg, Germany
tom.inderwies@st.oth-regensburg.de, juergen.mottok@oth-regensburg.de

***Abstract*—During a product lifecycle the capability to update the software of devices is paramount. Thus, new functionalities can be introduced, adapted and existing errors can be fixed. But by introducing an update capability, new attack vectors regarding the system security are introduced. To benefit from the advantages of software updates, especially from remote, while preventing security vulnerabilities, this paper presents a secure update mechanism of a system comprised of multiple controllers. This concept leverages cryptography and security relevant information to protect the authenticity, integrity and confidentiality of updates, while ensuring a fail-safe update to ensure the availability of the system. Therefore, digital signatures, a secure communication channel and cryptographic hashes are utilized.**

Keywords—software update; power grid; it-security; embedded systems; cryptography; security gateway

I. INTRODUCTION

The national power grid is regarded as critical infrastructure because the availability of energy guarantees public order. Without electricity medical services such as hospitals cannot operate, there will be a short supply of food as well as basic goods and the economy is disrupted due to its digitalization. These are just a few of various effects. In Marc Elsberg's book "Blackout" [1] a scenario is depicted where a hack attack takes down the electricity supply of a country and with it basic supplies. Apart from fiction, governments have crisis scenarios in place on the effect of a power outage and guidelines about how to respond. The responsible government department "Bundesamt für Bevölkerungsschutz und Katastrophenhilfe" gives advice on how to deal with such an event [2] alongside other authorities [3].

Hacking a national electric grid is not a fictional threat but reality. In 2015 the Ukrainian power grid partially went offline due to a hack attack [4], leaving numerous households and institutions without electricity. This incident is not an isolated event but part of the foreign policy of nations. As part of protection and attack scenarios the United States of America have infiltrated foreign power grids in an attempt to hold political and military leverage [5].

This is why the national power supply has to be protected in terms of information security against various threats and attackers such as nations, organized crime and other hostile actors. The Energy Safe and Secure System Module (ES^{3}M) is a system comprised of four microcontrollers with the goal to secure the communication within the power grid by ensuring the authenticity and confidentiality of messages and commands sent between controlling stations and the decentral power infrastructure responsible for the generation and distribution of electricity.

In order to keep the functionality of the Secure Module up to date, its software has to be updatable. This ensures that functionalities can be adapted or added and makes fixing errors and security vulnerabilities possible. An important factor is the scalability of an update process, because applying an update in person is time and cost intensive. Hence, a remote update process over the air is required. However, by introducing the possibility of remotely updating the software, new attack vectors are being created. If an attacker can install a custom software, he gains full control over the Secure Module and thus over the controlled power station.

Therefore, this paper aims to develop a concept to securely update the Energy Safe and Secure System Module with its microcontrollers over the air, which is an important aspect of the overall system security.

II. BACKGROUND AND SYSTEM OVERVIEW

The ES^{3}M is a module with the purpose of securing the communication in the electric grid, where controlling stations interact with the decentral infrastructure, responsible for operating the electric grid. In order to respond to changing power demand, maintenance and status reporting, the controlling stations must be able to send control commands and other information to the power substations. Furthermore, these substations need to send information back to the controlling stations, e.g. status information, on which subsequent commands depend. According to the received commands, power substations adjust their behavior like taking subcomponents off the grid or ramping up production. The electric grid is very sensitive to adjustments, where changes to its behavior have a tremendous impact on the stability of the grid and thus on the national power supply. Therefore, it is of utmost importance to make sure only authorized entities can influence the behavior of the power substations by sending commands. Additional requirements are the confidentiality of exchanged information and the availability of the substations.

The goal of the ES³M is to enforce a secured communication between substations and controlling stations. This is achieved by utilizing the secure communication protocol TLS [6] which encrypts messages sent over a network and enables the authentication of the communication partner. This ensures only authorized and authentic controlling stations can issue commands influencing the power substations. Additionally, data sent over a TLS connection can exclusively be accessed by authorized and authentic substations and controlling stations.

The secure module is located within the physical perimeter of the power substations and acts like a gateway between controlling station and controlled power substations. It routes all traffic between the controlling station and itself through the TLS connection. Thus, no one outside the physical perimeter of the substation can access data send over the network or pretend to be an authentic operator. The system is depicted in Figure 1.

The Secure Module consists of four microcontrollers, to improve the security by separating the cryptographic protocol from the network stack [7] (refer to Figure 2). The crypto controller (Crypto), which will be referred to as primary controller, and three communication controllers: black communication controller (BlackComm), red communication controller (RedComm) and diagnosis controller (Diag). These will be referred to as secondaries. All microcontrollers are from STMicroelectronics and based on the STM32-H7 controller family that offer a dual bank flash.

The BlackComm and diagnosis controller can communicate with the controlling stations and forward all TLS packages to the crypto controller. The crypto controller is responsible for all cryptographic operation, in this case verifying that the communication partner is indeed an authentic controlling station. After the successful authentication the TLS connection is established, and data and commands can be exchanged in an encrypted format, where the BlackComm or Diag forward all traffic to the Crypto. The Crypto then decrypts the data and forwards commands and settings via the RedComm to the controlled power substation.

In an update scenario, every controller must be able to receive and apply a software update. Since only the cryptographic controller can authenticate and decrypt data, it is responsible for distributing software to the secondary controllers and verifying it on their behalf.

III. Threat and Attack Analysis

By introducing a remote software update capability new attack vectors are being introduced which can be exploited by

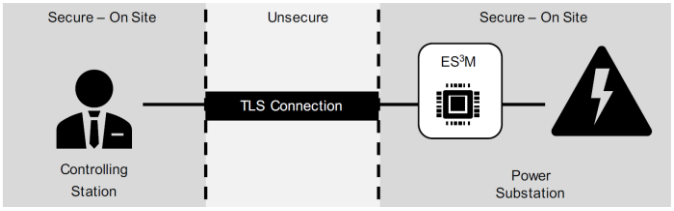


Figure 1. Environment – Controlling stations and power substations

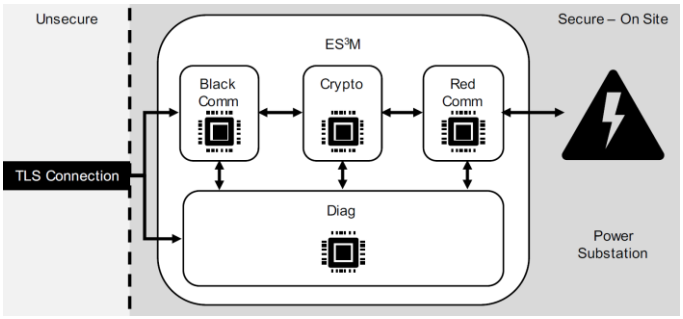


Figure 2. Architecture of Secure Module

hostile actors. Especially the option to alter software permanently is a huge incentive for hostile actors. If a non-authentic firmware is successfully deployed, attackers are able to siphon information over a large period without being discovered and are able to control power substation whenever it suits their timeline. This is particularly interesting from the point of view of national foreign policy. As mentioned, nations have already infiltrated foreign power grids [4] because it might be used as political leverage or physical attack vector.

This chapter discusses assets, potential attacker personas and their motivation, protection goals and various threats.

A. *Motivation of Attackers and their Personas*

There are different motivations, which drive various types of hostile parties to breach secure systems. The four main motivations are curiosity, personal fame, personal gain and national interests [8].

Curiosity brings hostile actors to breach systems. They want to try out attacks and improve their skills for the sake of learning. The goal is not necessarily fame or gain but sharpen their skill set.

Personal fame on the other hand drives people like researchers or activists. The motivation here is to show their community or the public that they are capable of breaching complex systems by hacking and publishing their results, thus gaining reputation.

Personal gain is another important factor. Scenarios are blackmailing authorities and companies to leave important data encrypted [9] unless a ransom is paid. Reverse engineering software to gain access to intellectual property is another example for a potential financial gain.

National interests pose another important motivation to attack systems, especially critical infrastructure. Infiltrating foreign critical infrastructure is part of the foreign policy of various nations as it can be used as leverage in negotiations or to cause physical damage as part of hostile acts [5] [10].

The goals and interests behind the motivation are manifold and a vast number of potential attackers have an incentive to attack systems. The company Intel has derived a list of potential attackers as part of their Threat Agent Risk Assessment (TARA) [11] model. A collection of the attacker types of TARA are listed in Table 1 and show the variety of hostile parties.

TARA – Hostile personas	
Anarchist	Disgruntled Employee
Civil Activist	Government Cyber Warrior
Competitor	Government Spy
International Spy	Terrorist
Irrational Individual	Thief
Legal Adversary	Cyber Vandal
Radical Activist	Vendor
Sensationalist	Organized Crime
Reckless Employee	Corrupt Government Official
Untrained Employee	Data Miner

Table 6: TARA – Threat Agent Risk Assessment

It is not always the usual suspect, e.g. competitors or governments. Untrained personal can do a lot of damage and insiders are often overlooked. Vandals are another category that poses a considerable threat, as their main goal is just to destroy.

B. Protection Goals

To protect systems against potential aggressors and threats, protection goals [12] have to be specified which categorize certain types of attacks. The most important ones for a secure update mechanism are confidentiality, authenticity, availability and integrity, which will be discussed in this chapter.

- Authenticity: Data, such as messages are authentic if it can be proven that they originate from an identifiable source. E.g., a control command is authentic if it really came from an authentic controlling station.
- Availability: This protection goal describes the constant availability of a service or resource despite potential attacks. E.g. if a microcontroller receives a non-authentic software update package it must deal with it in a way that it can still provide its normal service and react to further requests.
- Confidentiality: Data is confidential if only authorized parties are allowed to access data as plaintext. No non-authorized parties must be able to access the plaintext of the confidential data. E.g. only the authorized ES^{3}M can access the software update as plaintext. If someone intercepts the software update during its transportation it would not be of any help since its plaintext is not accessible.
- Integrity: Integrity describes the property that data in transit or in memory cannot be altered without being detected. E.g., data sent over a network connection must provide a mechanism to verify that the data has not been tampered with.

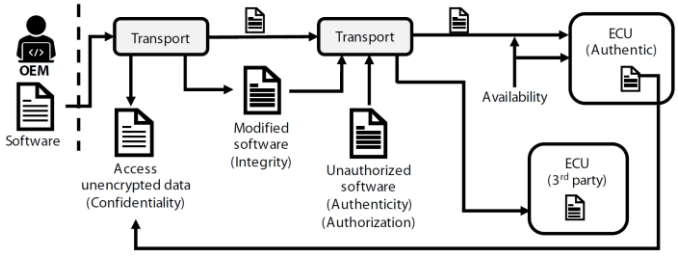


Figure 3. Threats in an update scenario

C. Threats

Attackers have the goal to gain access and control over the system, deny functionality or intercept confidential data. The system introduced in section II is suspect to different attacks without further security measures regarding an update process. During the transportation data can be modified, accessed or non-authentic data can be inserted. Even after the transportation, some threats persist without countermeasures like the installation of authentic software on non-authorized systems or denial of service attacks, endangering the availability of a system.

Figure 3 depicts different threats that must be taken into consideration when building a secure software update process. These attacks threaten the confidentiality, authenticity, integrity and the availability of data and services. Possible attacks are:

- Interception of data in transit or in memory: Violating the confidentiality, attackers can gain access to sensitive data and intellectual property, if the communication channel is unprotected or has weaknesses. The same applies to a microcontroller, if an attacker has local access, where the read-out of flash memory containing sensitive information is possible.
- Installation of non-authentic software: If an attacker can install a custom software that does not originate from the OEM, the attacker can gain full control over the system and its peripherals. This enables the attacker to fully control the power substations, which is a considerable threat to the power grid, belonging to the critical infrastructure.
- Manipulation of data in transit or in memory: If the binary code of a transmitted software image or related metadata is manipulated during its transportation or in the flash memory of a device, errors can occur. These errors can result in wrong parameters, which leads to an unwanted behavior, or even make the system unavailable, if the system cannot boot from the corrupted binary.
- Downgrade Attack: An attacker can try to install a software version that is authentic, but older than the currently installed version. Thus, known vulnerabilities of former versions can be reintroduced.
- Compatibility Attack: An attacker can try to install a software that is authentic but not compatible with the target controller. If the software is installed, the controller malfunctions or becomes unavailable.

- Update Interruption: If the target suffers a power loss during the update process, an incomplete binary can remain in memory. This leads to malfunctioning or unavailability.

The threats towards the update process stem from a combination of cryptographic aspects and missing or invalid information about an update like a version information. These aspects are the foundation for secure update process, covered in the upcoming chapter.

IV. Secure Software Update Concept

This chapter presents a concept about how to secure the software update mechanism within the ES^{3}M environment. It is based on digital signatures, security relevant metadata and encryption. The update process must ensure that only authentic software from an authentic source can be installed and memory content cannot be altered. Furthermore, the confidentiality of the software and the availability of the targets must be ensured. Questions are: Where does data come from? Who is authorized to receive data? Is the authenticity, integrity and confidentiality of data important and at stake? Who is authorized to access a system?

The concept leverages cryptography combined with security relevant metadata and further precautions to ensure the security of the software update. This paper addresses primarily remote threats, while local attacks are out of scope.

A. Controller Interaction

The update scenario affects all four controllers, which can be seen as primary and secondaries, where the crypto controller represents the primary and the others the secondaries (refer to Figure 4). The primary acts as master and receives all software updates. It then distributes the software updates if intended for a secondary or applies the update itself. Furthermore, the primary is equipped with a smart card that offers a secure storage for keys and certificates, is able to de- encrypt data, and verify digital signatures. Therefore, the primary verifies more complex security checks on behalf of secondaries before forwarding an update.

B. Security by Cryptography

This section presents the cryptographic concepts that are leveraged to secure the software update in terms of integrity, authenticity and confidentiality.

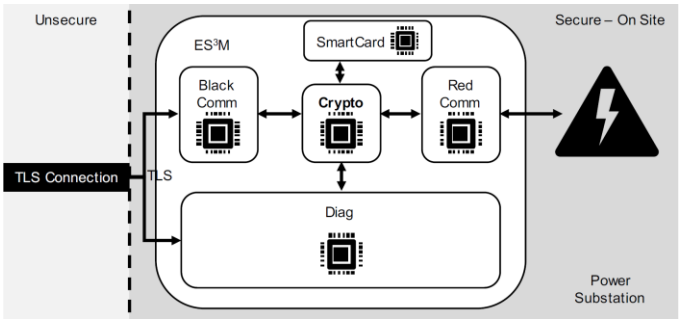


Figure 4. ES^{3}M system with smart card

Secure Channel TLS: To ensure the confidentiality, authenticity and integrity of the distributed software during its transportation, it is sent over an encrypted and authenticated TLS connection [6]. During the initiation the certificate of the connection partner is reviewed, a key exchange takes place and finally the data gets symmetrically encrypted before being sent over the air, together with a tag that authenticates and ensures the integrity of transmitted data. Hence, the sent software update is protected against eavesdropping during its transmission and the identity of the transmitting server can be verified, authenticating the transmitted data.

Digital Signatures: The software update and additional data has to be digitally signed [13] after its creation to ensure its authenticity. Before applying the software update, the signature can then be verified. Thus, only authentic software can be installed. Solely relying on TLS as authenticity mechanism does not protect against tampering with data before the signing process or if data has to be copied locally afterwards. Hence, digital signatures further increase the system security.

Cryptographic Hashes: The controller must be able to verify that data did not change during transportation remotely from server to primary or locally from primary to secondaries. Cryptographic hash functions [14][15] serve the purpose of verifying the integrity of data by processing it in such a way that data of arbitrary length gets scrambled to a value of fixed length called hash or hash value. Requirements for cryptographic hash functions are a strong collision resistance, pseudo-randomness and a one-way property. Strong collision resistance describes the property that it is highly unlikely to find two different messages that result in the same hash value *hash(m1) = hash(m2).* Pseudo-randomness means that a marginal difference of a message leads to entirely different hash value after processing it. The one-way property demands that it is impossible to compute the original message from its hash value.

C. Security-relevant Metadata

Cryptography can only protect data in terms of integrity, authenticity and confidentiality. To secure the system against further attacks, additional information about the software is required in order to decide whether an update should be applied or rejected, even if its authenticity and integrity could be verified.

Version number: In order to prevent downgrade attacks, the software update must come with a version number. Updates with a lower or equal version number must be rejected. Old versions might have vulnerabilities that can be leveraged in an attack chain to compromise the system.

Target controller: It must be specified which software update is intended for which controller. If a software update developed for the RedComm controller is installed on the BlackComm controller, the update fails and the BlackComm remains in a broken state.

Unique identifier: The unique identifier is required to identify a certain software version. It serves the purpose of downloading the correct data and distributing it within the ES^{3}M environment to the intended controller.

Target address: The target controller needs to know where in its memory the update needs to be installed. If it is installed to a wrong location, the device will remain in a broken state.

Expected size: To tackle denial of service attacks it is important to know the expected size of a software update. Hence, a download process must be aborted if the received data exceeds the expected size.

Expiry date: Software updates should have an expiry date to make sure no outdated software can be installed. Even if the pending software update has a higher version number than the currently installed, it is not necessarily the latest release.

D. Additional Security Measures: Fail-Safe Update

Apart from cryptography and information-based security, additional precautions must be taken to defend against further attacks and guarantee a fail-safe operation. In order to recover from a failed update process, a redundant firmware installation must be present. If no backup exists, the controller has to stay in bootloader mode and cannot continue normal operation until a new valid update is presented. This is not an option when used in the power grid environment, as controlling stations must be able to adjust the power grid at any given time. A system that is not operating and not responsive cannot be used in this context. That is why a redundant firmware must be present as a recovery option. If an update process terminates due to errors or a sudden loss of power, it leaves an incomplete update behind. With a backup the controller is able to start again and retry the update. The STM32-H7 controller family offers a dual bank mode, where the flash memory is divided into two separate flash memory units. Thus, a working firmware can be kept on one memory bank, while the update can be written to the second memory bank. On the downside this means that a lot of storage capacity is occupied by an inactive software, but due to the security requirements a redundant installation is more important.

E. Data Organisation

In order for a controller to process and forward an update within the ES^{3}M system the software update itself and other relevant information must be organized and distributed in a suitable fashion. Since the primary has to distribute software updates not intended for itself, it has to know which controller needs which update. Therefore, software updates themselves together with relevant information are clustered in a logical package called *Software Package* (SP) (refer to Table 2). The information which controller has to install which software together with further information is clustered in another logical package named *Update Request* (UR) (refer to Table 2).

Software Package	Information	Abbreviation
Header	Identifier	UUID
	Version	SW-Version
	Hash	SW-Hash
	Compatibility	SW-Recipient
	Binary Address	SW-Address
Update	Software binary	SW-Update
Digital Signature	Digital Signature computed over Header	SP-Signature

Update Request		Information	Abbreviation
Header		Identifier	UUID
		Which ES^{3}M is targeted	UR-Recipient
		Version of Update Request	UR-Version
		Date when UR expires	Expiry
Jobs	Target		
	Crypto	UUID of SP	SP-UUID
		Size of SP	SP-Size
	Comm Black	UUID of SP	SP-UUID
		Size of SP	SP-Size
	Comm Red	UUID of SP	SP-UUID
		Size of SP	SP-Size
	Diagnosis	UUID of SP	SP-UUID
		Size of SP	SP-Size
Digital Signature		Digital Signature computed over Header and Jobs	UR-Signature

Table 2. Software Package and Update Request

The SP is divided into a header, the update and a digital signature. The header holds information about the SP, the update part holds the actual software update as binary and finally a digital signature is computed over header and update to ensure its authenticity.

The UR is divided into three parts, namely the header, jobs and digital signature. The header holds information about the UR itself, a job list that contains update information about each controller and finally a digital signature to ensure the authenticity of the UR.

To defend against endless data attacks, the size of the update request must be known. The software package can be checked because the update request contains this information. However, there is no information about the size of the UR since it is the first package. To overcome this obstacle, a maximum size must be specified.

V. Security Checks

There are two possibilities in an update scenario of an ES^{3}M. The first event is a self-update of the primary controller (crypto controller). The second event is a pending update for a secondary. Since only the primary has the capability to execute certain security checks like verifying digital signatures due to its exclusive smart card access, a distinction between a primary update and a secondary update is necessary.

A. Primary Update

In case of a software update intended for the primary controller, all security checks can be performed on the primary since the cryptographic controller can access the smart card that stores sensitive keys and provides algorithms to verify digital signatures.

#	Security Check	Processed Information (UR)
1	Obtain UR over secure TLS connection	
2	UR smaller than specified max. size	Max-Size
3	Verify digital signature of UR	UR-Signature
4	Is the UR intended for this ES³M	UR-Recipient
5	Is the UR newer than the last seen	UR-Version
6	Has the UR expired	Expiry

Table 3. Security checks performed by primary – Update Request

Every update process starts with the primary retrieving the update request from an update server. It tells the primary which software updates to obtain and for whom an update is pending. In a self-update scenario, an update is pending for the primary. All data is sent over a secure TLS connection. Hence, it is impossible for an unauthorized party to access the plaintext during the transmission. To securely update the primary, all information must be processed in a suitable order. The security checks are described in Table 3.

After the update request has been obtained and verified successfully, the primary must download the correct software package. The required software package can be located by using its UUID, which is provided in the jobs part of the UR.

When the correct software package has been obtained over the secure TLS connection the security checks described in Table 4 must be performed on the software package.

After the SP has been successfully verified, it is written to the memory bank in the flash that is currently not in use. When booting again the banks are swapped and the new software version is running. The bank containing the old software version can now be used for future updates.

#	Security Check	Processed Information
7	Obtain SP over secure TLS connection	UR: SP-UUID
8	SP equals specified size in UR	UR: SP-Size
9	Verify digital signature of SP_Header	SP: SP-Signature
10	Verify hash of software update within the SP	SP: SW-Hash
11	The downloaded software is indeed the software mentioned in the UR	SP: UUID UR: SP-UUID
12	Version of new update is greater than the currently installed version	SP: SW-Version
13	The software update is compatible with the controller	SP: SW-Recipient
14	Write software to specified address on the memory bank that is not currently in use	SP: SW-Address

Table 4. Security checks performed by primary – Software Package

#	Security Check	Processed Information
7	Obtain SP over secure TLS connection	UR: SP-UUID
8	SP equals specified size in UR	UR: SP-Size
9	Verify digital signature of SP_Header	SP: SP-Signature
10	Verify hash of software update within the SP	SP: SW-Hash
11	The downloaded software is indeed the software mentioned in the UR	SP: UUID UR: SP-UUID

Table 5. Security checks performed by primary for secondary – Software Package

B. Secondary Update

When an update is due for a secondary, not all security checks can be performed locally on the secondary, since only the primary has access to the smart card, which contains sensitive keys and is able to verify digital signatures. Therefore, the primary needs to verify digital signatures on behalf of the secondaries.

At first, the primary again retrieves the latest update request, which holds the information for every controller that needs to be updated. Hence, the primary has to validate the update request like in the primary update scenario. All the security checks must be performed as mentioned in Figure 5.

After the successful verification of the UR, the specified software package, intended for a secondary, must be obtained by the primary. This happens again via the secure TLS connection. When validating the software packages there are a few differences compared to the primary update process due to its missing cryptographic capabilities. The primary must verify the signature of the header of the SP and the hash of the update on behalf of the secondary (refer to Table 5).

If successful, the primary sends the software package internally over an unsecure channel to the secondary together with a hash computed over the header of the software package to ensure its integrity. The secondary performs the remaining security checks as described in Figure 9.

After the successful verification, the update is written to the

#	Security Check	Processed Information
12	Verify hash of the header of the SP	SP-Header-Hash
13	Verify hash of the update of the SP	SP: SW-Hash
14	The software update is compatible with the controller	SP: SW-Recipient
15	Version of new update is greater than the currently installed version	SP: SW-Version
16	Write software to specified address on the memory bank that is not currently in use	SP: SW-Address

Table 6. Security checks performed by secondary – Software Package

inactive memory bank and is swapped after a reboot. The now inactive memory bank with the old software can now be used for upcoming updates.

In every case an update is only applied, if all Software Packages could be authenticated and authorized to be applied safely. In a system update, the secondaries are updated first by rebooting and swapping banks, hence booting from the new software. If the secondaries have started successfully, the primary controller is going to apply the update. After this procedure, all controllers have been successfully updated. This prevents only partial system updates.

VI. Conclusion

In this paper a secure update concept was developed that contributes to the overall security of an Energy Safe and Secure System Module. It could be shown how important a secure update mechanism is for the system security due to the profound impact the alteration of software has, which is especially true for critical infrastructure components such as the national power grid. By assessing various motivations for potential attackers as well as their archetypes, it became obvious that hacking the power grid presents a tremendous incentive for hostile parties to obtain personal, professional or national gains. Furthermore, various threats were presented, showing the importance of multiple security precautions at different levels. These threats jeopardize the confidentiality, integrity and authenticity of data. Based on these threats a secure update process was developed.

This secure update process leverages cryptography, security-relevant metadata and further precautions like a redundant software. Cryptography is the fundament, enforcing the confidentiality, integrity and authenticity of software updates. Remaining vulnerabilities could be eliminated by introducing further security-relevant metadata that is divided into an update request containing a job list and the software package, containing the update itself and further information. Based on this metadata, only authorized updates can be installed. Additionally, the software update was designed to be fail-safe by introducing a redundant software installation.

The next step is the implementation of the presented secure update process for all four microcontrollers based on the STM32-H7 series from STMicroelectronics. This includes the development of a bootloader that is capable of verifying information, a verification module that verifies signatures and hashes within the application space as well as the bootloader space and finally a protocol that is suited to transmit update orders and software packages between the different controllers.

With the introduction of this secure update process, the overall security of an ES^{3}M increases drastically since the alteration of software is an important factor. Nevertheless, an update process is only one of many factors affecting the security, and overall security must be assessed from a system-point-of-view that looks at the weakest links.

References

[1] Marc Elsberg, “Blackout”, Blanvalet, 2013

[2] Bundesamt für Bevölkerungsschutz und Katastrophenhilfe, „Ratgeber für Notfallvorsorge und richtiges Handeln in Notsituationen,“ 17.01.2019. [Online]. Available: https://www.bbk.bund.de/DE/Ratgeber/VorsorgefuerdenKat-fall/Pers-Notfallvorsorge/Stromausfalll/Stromausfall.html. [Accessed 28.05.2020].

[3] Regierungspräsidium Karlsruhe, „Musternotfallplan Stromausfall,“ 01.04.2014. [Online]. Available: https://rp.baden-wuerttemberg.de/Themen/Sicherheit/Documents/MusternotfallplanStromausfall.pdf. [Accessed 28.05.2020].

[4] Wired.com, „Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid,“ 03.03.2016. [Online]. Available: https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/. [Accessed 28.05.2020].

[5] The New York Times, „U.S. Escalates Online Attacks on Russia's Power Grid,“ 15.06.2019. [Online]. Available: https://www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html. [Accessed 28.05.2020].

[6] Internet Engineering Taskforce, „The Transport Layer Security (TLS) Protocol Version 1.2,“ 08.2008. [Online]. Available: https://tools.ietf.org/html/rfc5246. [Accessed 28.05.2020].

[7] Tobias Frauenschläger, Sebastian Renner, Jürgen Mottok, “Security Improvement by Separating the Cryptographic Protocol from the Network Stack onto a Multi-MCU Architecture”, 07.2020

[8] A. Shostack, Threat Modeling - Designing for Security, Indianapolis: J. Wiley & Sons, 2014.

[9] Süddeutsche Zeitung, „Wie Hacker Ihre Daten kidnappen können,“ 29.11.2015

[10] K. Zetter, „An Unprecedented Look at Stuxnet, the World's First Digital Weapon,“ 11.03.2014. [Online]. Available: https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/. [Accessed 28.05.2020].

[11] M. Rosenquist, „Whitepaper: Prioritizing Information Security Risks with Threat Agent Risk Assessment,“ 12.2009. [Online]. Available: https://www.researchgate.net/project/Threat-Agent-Risk-Assessment-TARA. [Accessed 28.05.2020]

[12] E. Barker, „Recommendation for Key Management, Part 1: General,“ National Institute for Standards and Technology, Special Publication 800-57 Part 1, 05.2020. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-57-part-1/rev-5/final. [Accessed 28.05.2020].

[13] National Institute of Standards and Technology, „Digital Signature Standard (DSS), FIPS 186,“ 07.2013. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/186/4/final. [Accessed 28.05.2020].

[14] C. Gebotys, Security in Embedded Devices, New York: Springer, 2010.

[15] National Institute of Standards and Technology, „Secure Hash Standard (SHS), FIPS 180-4,“ 08.2015. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/180/4/final. [Accessed 28.05.2020].

Advanced Intrusion Detection Architecture for Smart Home Environments

Julian Graf
Dept. Electrical Engineering and Information Technology
Ostbayerische Technische Hochschule
Regensburg, Germany
julian1.graf@st.oth-regensburg.de

***Abstract*—Due to the increasing number of digitized households worldwide cyber-attacks on Internet of Things (IoT) and Smart Home environments are a growing problem. The purpose of this study is to investigate how an Intrusion Detection System (IDS) can provide more security in IoT and Smart Home networks with a innovative architecture, combining classical and novel machine learning approaches. By combining standard security analysis methods and modern concepts of artificial intelligence and machine learning, we increase the quality of attack / anomaly detection and can therefore conduct dedicated attack suppression. The architectural image of the IDS consists of four different layers, which in parts achieve independent results. The autonomous results of the different modules are calculated by means of statement variables and evaluation techniques adapted for the specific module elements and subsequently combined. The architecture image combines approaches for the analysis and processing of IoT and Smart Home network traffic. From this result it can be determined whether the analyzed data indicates device misuse or attempted break-ins into the IoT / Smart Home network. This study answers the questions whether a connection between classical and modern concepts for monitoring and analyzing IoT and Smart Home network traffic can be implemented meaningfully within a reliable architecture and describes in detail the investigation and preparation modules. In the area of processing modules, the paper is an extended version of the concept architecture presented in the Workshop CosDeo from the PerCom Conference 2020 [1].**

Keywords*—*Artificial Intelligence, Machine Learning, Smart Home, Intrusion Detection System, Architecture, IoT

I. Introduction

In 2018 7 billion Internet of Things (IoT) devices were used worldwide [2]. 14 percent are consumer devices [3]. The more devices are networked together, the more they are vulnerable to attack. This makes the networks very opaque and requires specialists who can distinguish attacks from normal data [4].

Assistants like Google Home Mini [6] are used to control other devices with voice input. The microphones are always active and offer attack surfaces. [1]

To improve the security of the networks, security software like firewalls and intrusion detection systems are indispensable. Current firewalls are enhanced with intelligent algorithms to keep pace with the increasing number of attacks. But there are still new botnets, like Ares [7].

Intrusion detection systems are used to further improve the security level. Network-based IDS can detect attacks on individual devices without additional software. However, these systems cannot detect every attack. With current artificial intelligence (AI) algorithms, detection rates can be improved. [8].

Based on my research work and a recent publication this extended version describes detailed how the data pre-processing was done but if you want to have a more detailed view about the general architecture you can read ”Architecture of an intelligent Intrusion Detection System, 18th Annual IEEE International Conference on Pervasive Computing and Communications, CosDE, in Publication 2020”.

II. Related Work

Machine learning algorithms (ML) are part of many software projects nowadays. Therefore many approaches for IDS can be found with different kinds of AI integration. [9] and [10] are both using ML techniques to improve attack detection. To achieve false positive rates of zero, we need to combine more approaches. Hybrid methods exist, such as the hybrid IDS from [11]. The rule-based component should reduce false positives. Thus we have one algorithm for a low false positive rate and the other for the classification of the attack [1].

There is no similar combination of AI algorithms and rule-based components for our zero-false-positive goal, but there is a lot of work in evaluating individual AI algorithms for IDS, such as [12].

III. Network-based IDS

A network-based IDS uses various evaluation techniques such as protocol stack verification, application protocol verification, advanced protocol creation, etc. Protocol stack verification can be used to identify invalid flags and data packets. Application protocol verification is used to analyze higher order protocols such as HTTP, FTP, TELNET, etc. to investigate and detect unexpected packet behavior. Creating advanced logs can be important to analyze unusual events and monitor advanced network activity [15].

A. Detection methodologies

Three main categories exist, in which the intrusion detection types differ. Signature-based Detection (SD), Anomaly-based

Intrusion detection methodologies			
	Signature-based	**Anomaly-based**	**Stateful protocol analysis**
Strengths	- Simplest method to detect known attacks. - Detail contextual analysis.	- Effective to detect new and unforeseen vulnerabilities. - Less dependent on OS. - Facilitate detections of privilege abuse	- Know and trace the protocol states. - Distinguish unexpected sequences of commands.
Weaknesses	- Ineffective to detect unknown attacks, evasion attacks, and variants of known attacks. - Little understanding to states and protocols. - Hard to keep signatures/patterns up to date. - Time consuming to maintain the knowledge	- Weak profiles accuracy due to observed events being constantly changed. - Unavailable during rebuilding of behavior profiles. - Difficult to trigger alerts in right time.	- Resource consuming to protocol state tracing and examination. - Unable to inspect attacks looking like benign protocol behaviors. - Might incompatible to dedicated OSs or APs

Fig. 1. Intrusion detection methodologies [19] [1]

Classification of network-based intrusion detection approaches			
General	**Specific**	**Detection**	**Performance**
Statistic-based	Distance-based	Unknown	Medium
	Bayesian-based	Unknown & known	High
	Game Theory	Unknown	Low
Pattern-based	Pattern Matching	Known	High
Rule-based	Rule-based	Unknown & known	High
	Data Mining	Unknown & known	Medium
	Model/Profile-based	Unknown	Medium
	Support vector machine SVM	Unknown & known	High
State-based	State-Transition Analysis	Known	High
	User intention Identification	Unknown	High
	Markov Process Model	Unknown	Medium
	Protocol Analysis	AD, SD, SP Combination	Low
Heuristic-based	Neural Networks	Unknown & known	Medium
	Fuzzy Logic	Unknown	High

Fig. 2. Classification of network-based intrusion detection approaches [19] [1]

Detection (AD) and Stateful Protocol Analysis (SPA) [16], [17], [18]. In Figure 1 advantages and disadvantages are shown [1].

B. Fraud Detection Approaches

In the general description of attack detection, two distinguishing features are considered. On the one hand, anomaly detection and on the other hand misuse detection are distinguished. The classification of possible attacks into different categories such as computational, AI approaches and biological concepts is a very general and pragmatic approach and difficult to implement in practice. Therefore a subdivision into the following subcategories is useful: static-based, pattern-based, rule-based, stateful-based and heuristic-based. These subcategories are shown in Figure 2 [1] [16].

IV. Architecture of the Intelligent Intrusion Detection System

The advanced IDS architecture consists of four layers as shown in Figure 3 [1]. The first or base layer is responsible for collecting data and recording network traffic. The second layer prepares the stored data and makes a first pre-analysis. The third layer contains the applied machine learning methods and models. The following layers aggregate the results, summarise them and trigger actions [1]. The results of the different layers are aggregated and processed in the layer above.

A. Layer 1: Data-Collection-Layer

In the base layer, all transmitted data is listened to and stored at the central node of the networks. The base layer implements all necessary components to cut the transmitted data packets and provide them with transmission information. [1]. The software designed for this purpose ensures fast, error-free recording with low computing power, so that it can also be used on devices with moderate hardware equipment.

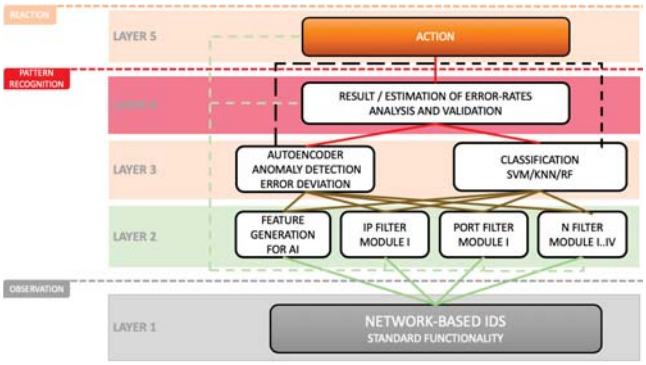


Fig. 3. Architecture of the intelligent Intrusion Detection System [1]

B. Layer 2: Investigation and Preparation Modules

The second layer is divided into two main subgroups, which are used for analysis and for data preparation/generation. Both subgroups combine different data collection and processing functions into a common superordinate layer. The data is analyzed using signature-based, anomaly-based and stateful protocol analysis methods. Subsequently, they are prepared for the use in machine learning procedures [1].

For the preparation of the data so that they can be used in machine learning models, we have to divide and distinguish the data. The data is divided into the following two areas. Network metadata and payload or transmission-specific information. The investigation module and the preparation module handle both data types separately. This subdivision is necessary because unchangeable parameters can be defined for metadata, which need not necessarily apply to specific transmission data.

1) Investigation Modules: The investigation modules analyse the transmission data. The following approaches are used for the analysis. statistic-based, pattern-based, rule-based and state-based methods. This generalization is necessary because only after the explorative data analysis (EDA) it can be determined which modules perform sufficiently well. Modules that combine several different static methods and rules are also used. An example is the snort project [20]. The current IDS architecture uses the following static modules for the detection and analysis of attack/abuse data The IP filter module determines which IP addresses can be used in the network. It analyzes the status of the dynamic host configuration protocol and checks network parameters for violations of threshold values, such as IP range limits. The port filter module determines which ports are open and available in the network. It logs all ports accessed by any member of the network and checks for port policy violations [1].

2) Preparation Modules: The main task of this module is to prepare data for further use in the respective machine learning models and to include features, which have been analysed in the course of EDA, in the data set. Parameters that are not transmitted in the original state of the network packets but can be derived from them are pre-processed in the preparation modules and included in the data set. One module is used to calculate the actual distance between two communicating devices. The calculation is based on the source and destination IP address. By determining the distance between the network participant and the associated cloud server, it is possible to include this value in our ML modules for classification purposes and to check whether a deviation from the default state is an intentional change or an indication of an attempted misuse/attack on the network. The module OneHotEnconding is responsible for data processing / data preparation. For example, text data or data that are not suitable for ML algorithms (such as IP address, protocol names and flags) are revised and stored in a database [1].

For the preparation of the data and the analysis of the variables, different methods are applied to the data set or the database.In the course of the EDA, the following examinations will be carried out at the beginning and also repeatedly during live operation:

- Outlier detection
- Association rule learning
- Clustering
- Classification
- Association analysis
- Regression analysis

The results of the different analyses are then collected, combined and evaluated in a next step. Based on these results, machine learning methods, model structure and architecture, including layer structure, are selected. Data is adapted to the resulting architectural model of selected structures and transmitted to the processing layer.

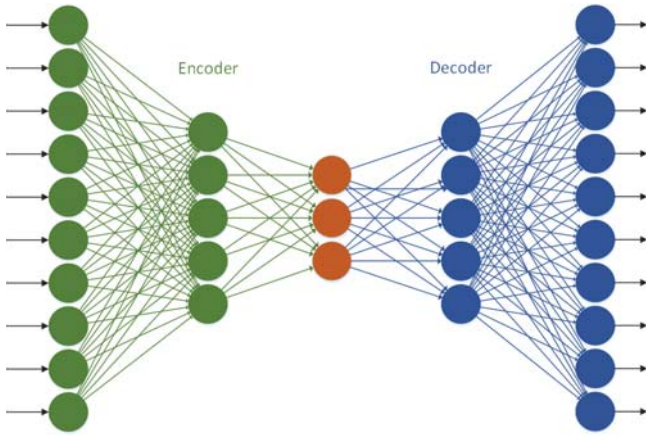


Fig. 4. AutoEncoder [1]

C. Layer 3: Machine-Learning and Deep-Learning Modules

The third layer is divided into two ML modules, which perform different tasks and differ greatly from each other. Both modules are designed to detect attacks, but use different architectures and procedures. The AutoEncoder module is designed to detect anomalies independent of the attack data and to have a high accuracy in detecting whether an attack has occurred or not. On the other hand, the second ML approach is intended to classify the different types of attack, i.e. to provide accurate information about the network threat [1].

1) AutoEncoder Module: Autoencoder are (deep) artificial neural networks with a specific architecture and a unique processing logic. Therefore, AutoEncoders can learn efficient representations of input data without any supervision. This input data is typically lower dimensional than the actual saved data. The specific field of AutoEncoders are powerful feature detection. They can be used for unsupervised pretraining of deep neural networks [21]. However, they can also be used for analysis of the unlabeled network data. AutoEncoders attempt to extract the most important elements from an input set, i.e. reduce the dimension of the input set to a smaller dimension and then extrapolate from this reduced dimension back to the original state (see Figure 4) [1].

Unlike the multi-layer perceptron, which has a very similar architecture to AutoEncoder, the number of neurons in the output layer must be exactly the same as the number of input neurons. AutoEncoder consists of two parts, the recognition network and the decoding or generative network [21].

The recognition network constantly reduces the number of neurons until it reaches the internal representation layer. At this point the generative network tries to restore the input state equivalent to the original one. To restore the initial state, the decoding segment uses only the reduced information stored in the representation layer [1].

AutoEncoders are particularly well suited for the analysis of IoT data for several reasons. When developing the IDS, it is easy to obtain standard transmission data from the IoT devices, but very difficult to simulate attack data to the extent

required for ML. With AutoEncoders we can train the ML model to learn how the network works in everyday situations [1].

2) Attack types classification module: The attack type classification module is different from the auto-encoder part. What we are trying to achieve in this module is not only the detection of an anomaly, but also the classification of the attack type that occurred. Detecting an anomaly gives us information about an incident on the network that was not planned in this way [1]. A suitable approach for this is the classification of attacks by image detection. Based on the network parameters, standardized images are created which can then be distinguished from an image classification procedure and the type of attack can be determined.

D. Layer 4: Estimation Module

Since the network packets are analyzed differently, the appropriate aggregation of the results is an essential component for the reliability of the system. In view of the many different modules, there must be a meaningful evaluation option for the partial results, which summarizes the results of the different modules and combines them to a final score [1]. We are currently investigating this question and already have initial ideas on how such a summary could be implemented sensibly.

E. Layer 5: Action Module

The last layer can take actions according to the results of layer 4. Log entries, notifications, broken connections or shutdown of the entire Internet connection are possible. In our test environment, connections are not broken after an attack is detected. We just want to collect data. In future applications, the IDS can withhold the data for analysis and then decide whether the data should be transferred to the Internet or the network. With this approach, data loss of private data can be avoided. The classification of the attacks carried out in the network is particularly important in order to initiate special security measures for damage limitation / prevention and to ensure the operation of the network as far as possible [1].

V. Conclusion and future work

With two different AI algorithms, one for the detection of anomalies and one for the classification of the attacks carried out, the iIDS should improve the detection rates. We combine these AI results with the static modules and rules to get the best information from all the data and to make the best possible decision.In a further research approach, we are testing a procedure that could detect network anomalies by means of image detection.

References

[1] J. Graf, S. Fischer, K. Neubauer, R. Hackenberg, Architecture of an intelligent Intrusion Detection System, 18th Annual IEEE International Conference on Pervasive Computing and Communications, CosDEO, in Publication (2020).

[2] (2019, Oct.) State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating. IoT Analytics. [Online]. Available: https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

[3] (2019, Oct.) IoT trend watch 2018. IHS Markit. [Online]. Available: https://cdn.ihs.com/www/pdf/IoT-Trend-Watch-eBook.pdf

[4] S. Fischer, K. Neubauer, L. Hinterberger, B. Weber and R. Hackenberg, "IoTAG: An Open Standard for IoT Device IdentificAtion and RecoGnition", The Thirteenth International Conference on Emerging Security Information, Systems and Technologies, 2019, in press.

[5] W. L. Zangler, P. Panek and M. Rauhala, Ambient assisted living systems - the conflicts between technology, acceptance, ethics and privacy. Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2008.

[6] Google Ireland Limited (2019, Oct.) Google Home Mini. [Online]. Available: https://store.google.com/product/google_home_mini

[7] C. Cimpanu. (2019, Oct.) A new IOT botnet is infecting Android-based set-top boxes. ZDNet. [Online]. Available: https://www.zdnet.com/article/a-new-iot-botnet-is-infecting-android-based-set-top-boxes/

[8] N. A. Alrajeh and J. Lloret, "Intrusion detection systems based on artificial intelligence techniques in wireless sensor networks", International Journal of Distributed Sensor Networks 9.10, p. 351047, 2013.

[9] J. Cannady, "Next generation intrusion detection: Autonomous reinforcement learning of network attacks.", 23rd national information systems security conference, pp. 1-12, 2000.

[10] A. Shenfield, D. Day and A. Ayesh, "Intelligent intrusion detection systems using artificial neural networks", ICT Express, 4(2), pp. 95-99, 2018.

[11] S. Koutsoutos, I. T. Christou and S. Efremidis, "An Intrusion Detection System for Network-Initiated Attaclcs Using a Hybrid Neural Network", Artificial Intelligence Applications and Innovations, Springer US, pp. 228-235, 2006.

[12] H. Liu and B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey", Applied Sciences, vol. 9, no. 20, p. 4396, 2019.

[13] DATACOM Buchverlag GmbH. (2019, Sep.) AMI (advanced metering infrastructure), 2013. [Online]. Available: http://www.itwissen.info/AMI-advanced-metering-infrastructure-AMI-System.html

[14] Bundesamt fuer Sicherheit in der Informationstechnik. (2019, Sep.) BSI - Smart Metering Systems - Smart Metering Systems, 2019. [Online]. Available: https://www.bsi.bund.de/DE/Themen/DigitaleGesellschaft/SmartMeter/smartmeter.html

[15] S. Sapiah, Intrusion detection and prevention, Information Systems Department: Course Technology CENGAGE Learning, 2004.

[16] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)", NIST Special Publication 800, p.94, 2007.

[17] P. Stavroulakis and M. Stamp, Handbook of information and communication security, Springer, 2010.

[18] S. Axelsson, "Intrusion detection systems: a survey and taxonomy", Chalmers University of Technology, pp. 1-27, 2000.

[19] H. J. Liao, C. H. R. Lin, T. C. Lin and K. Y. Tung, "Intrusion detection system: A comprehensive review", Journal of Network and Computer Applications 36, pp. 16-24, 2013.

[20] Cisco. (2019, Nov.) Snort - Network Intrusion Detection and Prevention System, 2019. [Online]. Available: https://www.snort.org

[21] A. Geron, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems, O'Reilly Media, Inc., 2017.

[22] K. K. R. Kendall, A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems, Massachusetts Institute of Technology, 1999.

[23] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation.", Springer, pp. 345-359, 2005.

Usage of Image Classification for Detecting Cyber Attacks in Smart Home Environments

Johannes Ostner
Dept. Electrical Engineering and Information Technology
Ostbayerische Technische Hochschule
Regensburg, Germany
johannes.ostner@st.oth-regensburg.de

***Abstract*—The number of private households using Internet of Things (IoT) devices to simplify their daily lives is growing day by day. But the great majority of these devices are barely secured and not even close to meet the lowest thresholds of cyber security standards. The outcome is a worldwide and constant growth of cyber attacks on smart home environments. This study targets the question to what extent an intelligent Intrusion Detection System (iIDS) is capable of increasing the security level of such environments by utilizing a novel and machine learning (ML) based approach that relies on image classification. Whereas the majority of existing Intrusion Detection Systems (IDS) make use of classical security analysis methods we augment the prevailing ways of detecting intrusions by adding artificial intelligence (AI) to an existing IDS. The underlying and crucial goal of this study is to find a fast but unambiguous way of transforming network traffic into images which can be handed over to an classification model afterwards. Therefore various procedures are carried out on one specific dataset that contains pre-labeled normal and attack data. The output images are used for training and testing a machine learning model with unvarying architecture so that the resulting metrics regarding the classification accuracy and precision can be compared in the end. This enables a clear statement to be made as to which procedure has worked best and whether the chosen approach can actually improve the level of security within smart home environments.**

Keywords*—Internet of Things, Artificial Intelligence, Machine Learning, Smart Home, Intrusion Detection System, Image Classification

I. Introduction

Today the amount of technical devices getting connected with each other and the internet is increasing everyday. Voice assistance systems like Alexa, Siri, Google Assistant and Cortana became broadly accepted. Combined with numerous sensors, switches and cameras, they ease everyday activities and add some more comfort to our lives. According to recent studies, the so-called Internet of Things will consist of 20 to 30 billion different devices by the end of 2020 [1].

But even though IoT devices inevitably transmit sensitive personal data, security is often neglected. This counts especially for devices from the consumer sector. For non-technical users who want to enjoy the benefits of a smart home environment but with lack of the appropriate cyber security knowledge it is hardly possible to determine which devices are properly secured and which ones are not. And the number of possible threats is increasing tremendously. The Mirai botnet, for example, exploited publicly visible vulnerabilities of hundreds of thousands IoT devices to control and use them for cyber attacks [2].

IoT devices are not only conquering our private life but also sectors like Ambient Assisted Living (AAL) where sensors and cameras are used to monitor the health status of people in need of care. The collected and transmitted data in such environments is highly sensitive and therefore in need of special protection. In some cases a cyber attack could even constitute a danger of life which makes the protection of AAL networks even more crucial. However, there are some challenges to be met. One problem is for example the great individuality and flexibility of such networks. Every AAL network can consist of different IoT devices from dozens of different manufacturers. Also the number of integrated devices and sensors can vary for each household and over time as well. This makes it almost impossible to secure AAL networks just by applying a standard firewall solution.

Of course firewalls are improving steadily but so are cyber attacks. To add an additional security layer, network based Intrusion Detection Systems can be applied. Such software tools analyze the network traffic and are able to detect attacks according to static rules. Nevertheless, these advanced systems cannot detect all occurring attacks. Therefore a lot of manufacturers extend their IDS by using state-of-the-art artificial intelligence approaches that are capable of detecting special attack patterns and achieve detection rates up to ninety percent [3].

Because it is already proven that machine learning approaches can have a significant impact on the efficiency of IDS but there are almost no research findings when it comes to image classification in combination with an iIDS, we present our results regarding this novel approach in this paper. For this purpose we use an already existing IDS that is integrated into an experimental AAL environment from another research project. The IDS is able to analyze incoming and outgoing network traffic of all connected IoT devices and produces datasets for further analysis and processing. Then we carry out different procedures of transforming this data into images and evaluate the outcome by comparing the different intrusion

detection results of our machine learning model.

The paper is structured as follows. The next section contains information about related work. Section III provides some general information about our AAL environment and the utilized IDS. In Section IV the provided pre-labeled dataset is described. Section V deals with the different procedures for the image transformations, while Section VI comprehends the different results gathered by training and evaluating the machine learning model with the processed datasets. The paper ends with Section VII which contains a conclusion and future work.

II. Related Work

Achieving high attack detection rates by assimilating vast datasets and integrating AI concepts into a customary IDS is not a recently discovered approach. Numerous research papers can already be found that are based on this or other similar ideas. [4] implements a ML model to detect Denial of Service (DoS) attacks of infected IoT devices within private networks by analyzing network traffic on packet level. [5] conducts successful research on detecting network-based attacks by using an IDS involving the rather unconventional approach of reinforcement learning. However, both papers rely on raw network traffic as data input and neglect the opportunity of transforming the data into images.

[6] and [7] actually figured out ways of preprocessing raw data to create images as input for their ML models. But the first paper is focusing on the classification of different malware binaries and not on the resulting network traffic. And although the second research work is indeed dealing with the transformation of network packets into images it is not yet the approach we envision. Their idea is to process raw payloads of concatenated network packets and hand over the outcome into their deep neural network. But we want to take this idea one step further by taking header information of network packets into account. Moreover, we want to focus on single network packets because there is no evidence that concatenated packets are advantageous.

III. AAL Environment and the existing IDS

As already stated, this paper is focusing on intrusion detection within AAL network environments. It is part of a publicly funded research project which is called ”Secure Gateway for Ambient Assisted Living (SEGAL)”. The overarching goal of SEGAL is to enable people in need of care to manage their daily lives independently and remain at their own homes for as long as possible by applying secure, technical services. These services are provided by several IoT devices (e.g. Amazon Alexa, thermostat, heart rate monitor etc.) and sensors (e.g. smoke sensor, fall sensor, etc.) working together to monitor the mental and physical health status of people living in AAL environments (see Figure 1). The gathered data is forwarded to an external control center for further processing through a secure and standardized channel, called ”Smart Meter Gateway”. [8] [9]

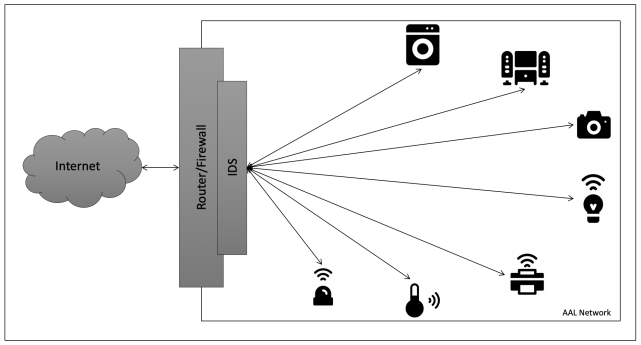


Fig. 1. Exemplary AAL Network Environment

But even though the data transmission itself is properly secured, the participating IoT devices have numerous vulnerabilities that can be exploited by attackers. Possible threats for this highly sensitive and personal data are corruption, manipulation and thievery. Therefore, a system that is capable of detecting such threats and attacks within an inner AAL network is crucial for safety purposes. As part of the SEGAL project another team of our laboratory already developed a special IDS to meet these requirements and set up a AAL test environment with several different IoT devices. Once the IDS is switched on, it permanently monitors the attached network and analyzes incoming and outgoing network packets by utilizing several rule-based modules in parallel. Each module focuses on a different threat and is able to raise an alarm if the currently examined packet contains suspicious data. An evaluation layer is collecting all module results for every single network packet and decides whether the IDS should raise an intrusion alarm or not. All packet information and the final result of the evaluation layer is pushed to a database which is used for further analysis later on (see Figure 2).

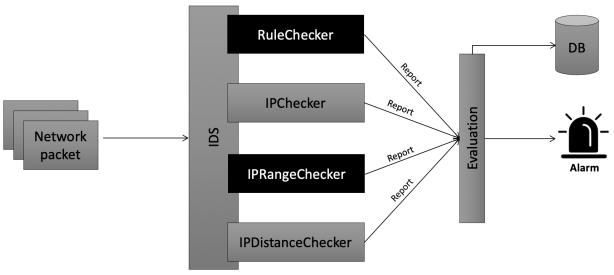


Fig. 2. Current IDS Architecture

However, this modularized and static IDS is not flexible enough to adapt to environmental changes or to learn attack patterns. E.g. the whole ruleset of several modules needs to be updated whenever a previously unknown IoT device is added to the AAL network. Otherwise it would immediately increase the false alarm rate. Hence it is necessary to extend the IDS by several ML based modules including the image classification approach, described in this paper.

IV. Provided Dataset

For the purpose of transforming network traffic data into images and to use the outcome for intrusion detection afterwards a huge amount of pre-labeled data is required that is representing normal traffic as well as attack data. The easiest solution to access high quality datasets would be to use the publicly provided and well maintained data of recent cybersecurity research projects. But because an AAL network is very different compared to common private networks, we decided to make use of the data provided by the IDS from the already described test environment. The process of creating a suitable dataset was already carried out in advance but because it is fundamental for understanding the procedures described in the following section a deeper comprehension of the structure of the dataset and the way of its creation is necessary.

A. Dataset Creation

The collection of the required data consists of two steps. The first step is to gather network traffic that depicts normal cases or in other words no-attack traffic. Therefore, the IDS is switched on after it is assured that all connected IoT devices are clean and properly secured. To guarantee that not one single IoT device is infected by malware they are reset to factory settings beforehand. For a certain time the IDS is now monitoring the traversing network traffic and pushing each packet into the connected database after it was analyzed and labelled as normal packet.

Collecting attack traffic is a little bit more sophisticated. It is necessary to simulate some cyberattacks but in a realistic manner so that there is no bias because of laboratory conditions. Furthermore, it needs to be assured that the evaluation layer of the IDS is labelling the packets correctly because during each attack, normal packets are also sent by unrelated IoT devices. Therefore, packets marked as attack data are validated manually after each simulation by, e.g. checking the IP addresses, protocols and packet sizes. For the provided dataset only one attack type was simulated but in several variations and multiple times. So-called Denial of Service (DoS) attacks are one of the most common attacks in terms of IoT networks and can be carried out both realistically and without great effort [10]. After each attack the labelling of the data was validated before the next simulation started.

Finally, to avoid the pitfall of another bias for machine learning models by having all attack data at the very end of the dataset, it is shuffled in a specific manner. It is crucial to keep every packet belonging to a single attack together because otherwise attack patterns are fragmented. Hence not all packets are shuffled but normal data packets and attack data blocks.

B. Insights regarding Features and other Key Facts

After the analysis of a network packet is completed, it is transmitted to the database for further processing. But not only the payload is stored. Beside that, all header information of network layer 2, 3 and 4 is attached including source and destination IP addresses, TCP or UDP port numbers, header length, and so on. Moreover, the evaluation layer appends an assessment value that determines the likelihood of an attack combining the estimations of the individual IDS modules and the resulting label whether the packet depicts an attack or not. The outcome is a dataset with 51 features which are itemised in Table 1. It consists of almost 250.000 analyzed network packets, 95% of which constituting normal packets. With only 5% attack traffic the dataset is highly imbalanced but this challenge is not addressed in this paper. Nevertheless the fact should be kept in mind because it could harm the machine learning results.

TABLE I. Feature List of Provided Dataset

Feature List		
l2_dstAddr	l2_payload_length	l2_srcAddr
l2_type	l3_dont_fragment_flag	l3_dstAddr
l3_fragment_offset	l3_header_checksum	l3_header_length
l3_id	l3_ipv6_header_flabel	l3_ipv6_header_hoplim
l3_ipv6_header_prot	l3_ipv6_header_tclass	l3_more_fragment_flag
l3_options	l3_padding	l3_payload_length
l3_protocol	l3_reserved_flag	l3_srcAddr
l3_total_length	l3_ttl	l3_type_of_service
l3_version	l4_dstPort	l4_srcPort
packet_length	tcp_ack_flag	tcp_ack_number
tcp_checksum	tcp_data_offset	tcp_fin_flag
tcp_options	tcp_padding	tcp_psh_flag
tcp_reserved	tcp_rst_flag	tcp_seq_number
tcp_syn_flag	tcp_urg_flag	tcp_urgent_pointer
tcp_window	udp_checksum	udp_length
contentraw	contentraw_length	contentclear
contentclear_length	assessment	label

V. Image Transformation Approaches

This section describes three different approaches of transforming network packets into RGB images that can be used for training and testing a machine learning model. As shown in Figure 3 all three approaches are carried out successively by using the same input dataset and before their processed images are fed into one and the same machine learning model. It is crucial to use the same model because otherwise the produced results cannot be compared credibly in the end. Furthermore, the created images are stored on disk so that they can be reused later and to save computational resources for future experiments.

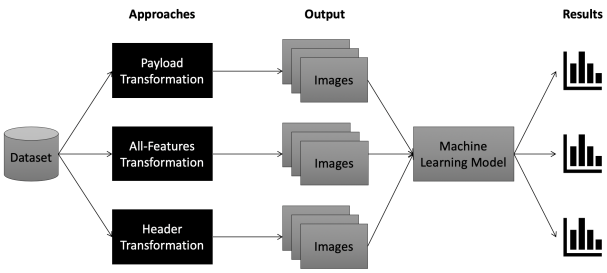


Fig. 3. Transformation Approaches and General Procedure

A. Approach I: Payload Transformation

The first approach is relatively straightforward and the groundwork for the other approaches where some more complexity is added. The idea is to create colored images with a fixed, squared resolution by converting only the payloads of the provided network data. Therefore the feature ”contentraw” is selected which contains the byte stream of the original packet content. There is another feature, called ”contenclear” which contains the decoded and human readable payload but because most of the packets are encrypted anyway it does not constitute additional value. Although this is only an assumption but can easily be proven or disproved when comparing the ML model results at a later stage, the image transformation is continued with the first feature.

At first the dataset is split into smaller chunks of 10.000 packets each to reduce the required memory size during the transformation process. Then each chunk and thus each contained packet is sequentially converted into images (see Figure 4). Every processed data chunk results in an array of images that is stored on disk. Because the transformation procedure is the same for every single packet and its payload, just one fictive operation is explained.

The goal is to create a squared RGB image. For this purpose the payload data must be reshaped to a three-dimensional matrix where each dimension contains values between 0 and 255 for one specific color. Regarding the resolution of the matrix there are two limitations. First, the image has to be squared and second the minimum width defined by the used ML model is 32 pixels [11]. This yields in matrices with a minimum shape of 32x32x3 and thus in 3.072 values. The fact that payloads are already stored as byte streams enables a rather simple and byte-wise transformation into an array containing values between 0 and 255. However, each packet carries payloads with various lengths which would entail images with various resolutions. To overcome this problem the starting point of every transformation is creating a black image with a fixed resolution of 32x32 pixels. Afterwards the required amount of bytes is replaced by payload data. Furthermore it needs to be checked in the beginning if the maximum payload size of the whole dataset is exceeding 3.072 bytes which would result in larger image matrices. But since this is not the case in our dataset we stick to a fixed RGB image size of 32x32. After the entire operation is completed, the transformed payload data together with the corresponding label value is added to an array of images. [7]

Fig. 4. Exemplary Image Transformation Outputs

B. Approach II: All-Features Transformation

As described before, the second approach is very similar compared to the first one beside adding some more complexity on top. The desired image resolution remains the same at a height and width of 32 pixels as well as the way of processing the dataset in several chunks and storing the resulting images on disk. However, the input for the transformation process differs. Not only the payload data of each network packet is transformed but also the other available features which primarily contain information that was extracted from the IP-headers.

Whereas payload data can be easily reshaped into values between 0 and 255, it needs more effort to process header information. The reason is that the majority of relevant features contain string values or integers with a value range that exceeds 0 to 255 by far. Therefore, two solutions for dealing with these issues are applied. First, the string values are encoded by utilizing a so-called Label Encoder. All corresponding features and the contained strings are passed to this tool one after the other and for each run it automatically determines the amount of different values and thus also the possible value range. Subsequently, every string value gets encoded as integer starting with 0 where similar values get the same label code. Since none of the affected features contain more than 255 different values, no further action is required. [12] The second solution targets features that contain integers or byte values beyond zero, above 255 or both. Such values cannot be converted into the RGB scheme and thus need to be rescaled under the precondition that relative differences between them are preserved. For this purpose a so-called Min-Max-Scaler is used which is capable of transforming passed feature values to our desired value range of 0 to 255. After setting the range to values between 0 and 255 and passing the whole dataset it translates each integer feature individually and consecutively. [13] After executing these two additional preprocessing steps on the given dataset, the procedure for creating images can be restarted. Again, for every single network packet a black and squared image is created. Afterwards the payload data together with the encoded or rescaled header information is reshaped and replacing a specific amount of bytes of the black image. The resulting images look very similar to those in Figure 4, except that they contain a little more colored pixels.

C. Approach III: Header Transformation

Even the third approach of transforming the input dataset into images is reusing some techniques of the previous approaches. What distinguishes this one is that all features are processed except both payload data features and of course the label feature, i.e. only the header information of each network packet. Therefore the same preprocessing steps as in approach II are applied to transform strings and integers into the correct format and the required value range. The image size and the way of processing the dataset in several chunks stay the same but there is another difference compared to the previous procedures. The downside of ignoring the payload data is that there are only 48 features and thus only 48 data points left

to be transformed into colored pixels. But because the aspired images require three layers of 32x32 pixels, resulting in 3.072 values, the major area of each image would be displayed black. We assume that this would cripple the capabilities of the ML model to detect intrusions later on because there are almost no clues on the images that could indicate attacks. To tackle this problem in the first place, the header data of each network packet is repeated 64 times (3.072 divided by 48) so that the whole image is filled with data from the original dataset. As shown in Figure 5, the resulting images look quite different compared to those of the previous approaches.

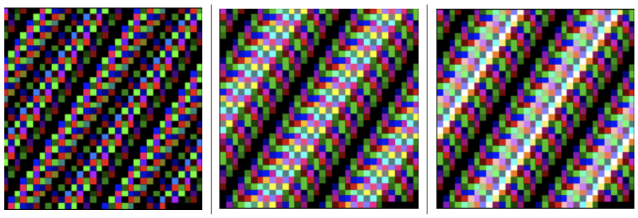


Fig. 5. Exemplary Image Transformation Output (Approach III)

VI. Machine Learning and Result Comparison

Executing all image transformation approaches on the original data results in three different image datasets ready for being passed into a ML model. To determine which approach reveals the best and most reliable performance in detecting intrusions on AAL networks, each dataset is utilized by one and the same ML model. The architecture as well as the reasons for choosing this specific model type are explained in the following section. Afterwards the prediction process is described, followed by a conclusion that compares the achieved results for each approach.

A. Machine Learning Model - Adjusted VGG-19

For this research work we decided to use transfer learning combined with a pre-trained and well-known Convolutional Neural Network (CNN), called VGG-19 which was invented in 2014 and is now part of the Keras library. As shown in Figure 6, it consists of 16 convolutional and three fully-connected (FC) layers and is trained on more than a million images with 1.000 different object categories. [14] This fact makes the VGG-19 capable of extracting features from a huge variety of distinct images and therefore very suitable for transfer learning [15]. As already proven in several research projects ([7], [14], and [15]), this rather simple model is a good choice when it comes to image classification tasks with very diverse input data.

After downloading the model including the pre-trained weights, some minor adjustments must be made so that it is applicable for this specific image classification task. At first, the input layer has to be replaced to match the size of the created images, i.e. by a layer with 32x32x3 input neurons. Moreover it needs to be assured that the pre-trained weights are not changed during the training phase later on. For this purpose the ”Trainable”-Flag of each already existing layer is

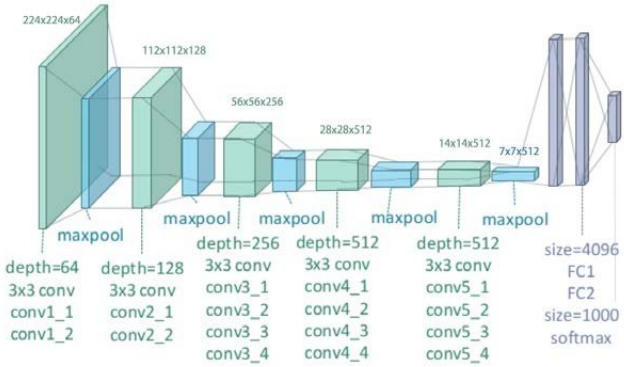


Fig. 6. VGG-19 Model Architecture [14]

set to false. Finally, the fully-connected layers at the end of the architecture are removed. On the one hand, the original model is designed to classify 1.000 different object types whereas this research work rests on a binary classification task for distinguishing attack data from normal network traffic. Therefore, a new output layer with just one neuron instead of 1.000 is created. On the other hand, a new fully-connected layer consisting of 128 neurons together with another fully-connected layer which flattens the output of the last max-pooling layer are inserted right before the output layer to enable transfer learning. These new layers have no pre-trained weights yet and thus the weights can be adjusted to the input data during the following training phases. [7]

B. Prediction Process

As soon as the model is properly adjusted and prepared, the prediction process can be started which is carried out three times, once for each image dataset. Because only the input data is varying but the process itself remains the same, it is described only once. At first, the dataset is split into training data and test data by a commonly used ratio of 3:1. During this step it is crucial to take care of an equal distribution of normal and attack data within each dataset. Afterwards the so-called hyperparameters for the ML model must be set. We decided to use cross entropy as loss function because it is the best choice for binary classification tasks [18]. RMSprop is used as optimization algorithm because it is fast and supports adaptive learning rates [19]. Moreover, batch learning is applied to reduce computational costs and the number of epochs is set to 20 but in combination with early stopping to avoid overfitting from the start. Now that all preparations have been made, the training of the model instance can be initialized by passing the training dataset. After completion, the performance of the trained instance can be tested by evaluating its predictions on unknown data of the test dataset.

C. Evaluation and Result Comparison

To figure out which image transformation approach works best for detecting intrusions in AAL networks the three resulting model instances and their ability to distinguish between normal and attack data needs to be evaluated and compared.

Therefore certain metrics are taken into account. These metrics are calculated from the contents of a so-called confusion matrix, which in turn provides information about the accuracy of made predictions by specific model instances. As shown in Figure 7, it gives an overview over properly classified intrusions (label = 1, prediction = 1; True Positive (TP)), false alarms (label = 0, prediction = 1; False Positive (FP)) as well as overseen intrusions (label = 1, prediction = 0; False Negative (FN)).

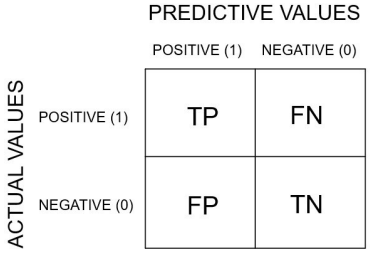


Fig. 7. Confusion Matrix [16]

The precision metric (p) depicts the ratio of how often network packets that were marked as intrusions by the ML model instance were actually intrusions, whereas the recall metric (r) tells how many intrusions were detected at all. Calculating the harmonic mean of recall and precision puts both metrics in relation and results in the commonly used F1 score [17]:

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{p * r}{p + r}$$

Precision, recall and f1 score are now calculated for each model instance and thus for each approach of transforming network traffic into images. Afterwards they are compared, as shown in Table II.

TABLE II. RESULT COMPARISON BETWEEN APPROACHES

	Approach I	Approach II	Apporach III
Precision	99%	99%	100%
Recall	95%	96%	92%
F1 Score	97%	97%	96%

VII. CONCLUSION AND FUTURE WORK

Surprisingly, the overall results shown in Table II are by far better than expected and all model instances are almost equally good in classifying network traffic or in other words: in detecting intrusions. There are almost no false alarms, just the amount of detected intrusions varies slightly but still on a high level. Approach II, where payloads in combination with header information were transformed into images seems to work best. But the gathered results also raise some doubts because they are rather too good to be true. It needs further investigations on what kind of features were extracted by the ML model during training and what leads the model to its decision whether a specific network traffic image is normal or malicious. Maybe the dataset is not really representative, the attack data is too similar or there are too few attack types included as well. Before a valid statement can be made whether image classification can extend an IDS in a valuable way or not, these issues need to be examined. Nevertheless, the results are more than promising and research on the approach of transforming network traffic into images to detect intrusions within AAL networks should be continued.

REFERENCES

[1] Costa, Luís and Barros, Joao Paulo and Tavares, Miguel, "Vulnerabilities in IoT devices for smart home environment", ICISSP 2019 - Proceedings of the 5th International Conference on Information Systems Security and Privacy, pp. 615-622, 2019

[2] O. Von Westernhagen. (2020, Jan.) Mirai: Die Entwickler des IoT-Botnetzes arbeiten jetzt für das FBI. [Online]. Available: https://www.heise.de/security/meldung/Mirai-Die-Entwickler-des-IoT-Botnetzes-arbeiten-jetzt-fuer-das-FBI-4169926.html

[3] N. A. Alrajeh and J. Lloret, "Intrusion detection systems based on artificial intelligence techniques in wireless sensor networks", International Journal of Distributed Sensor Networks, no. 351047, p. 3, 2013.

[4] R. Doshi, N. Apthorpe, N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices", IEEE Symposium on Security and Privacy Workshops, pp. 29-34, 2018.

[5] J. Cannady, "Next generation intrusion detection: Autonomous reinforcement learning of network attacks.", 23rd national information systems security conference, pp. 1-12, 2000.

[6] Q. Le, O. Boydell, B. Mac Namee, M. Scanlon, "Deep learning at the shallow end: Malware classification for non-domain experts", Proceedings of the Eighteenth Annual DFRWS USA, pp. 118-126, 2018.

[7] T. Mirza. (2018, Sep.) Building an Intrusion Detection System using Deep Learning. [Online]. Available: https://towardsdatascience.com/building-an-intrusion-detection-system-using-deep-learning-b9488332b321

[8] K. Schwendner (2019, Aug.) Secure Gateway Service for Ambient Assisted Living. [Online]. Available: https://www.konsensplan.de/segal/

[9] S. Fischer, K. Neubauer, L. Hinterberger, B. Weber and R. Hackenberg, "IoTAG: An Open Standard for IoT Device IdentificAtion and RecoGnition", The Thirteenth International Conference on Emerging Security Information, Systems and Technologies, 2019, pp. 107-113.

[10] E. Hodo et al., "Threat analysis of IoT networks using artificial neural network intrusion detection system", 2016 International Symposium on Networks, Computers and Communications (ISNCC), Yasmine Hammamet, 2016, pp. 1-6, doi: 10.1109/ISNCC.2016.7746067.

[11] keras.io (2020, June) VGG16 and VGG19. [Online]. Available: https://keras.io/api/applications/vgg/

[12] S. Raschka, V. Mirjalili, "Machine Learning mit Python und Scikit-Learn und TensorFlow: Das umfassende Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning", mitp Verlag, pp. 132-133, 2017, isbn: 9783958457355.

[13] J. Brownlee, "Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions", Machine Learning Mastery, pp. 123-124, 2018.

[14] Y. Zheng, C. Yang, A. Merkulov, "Breast Cancer Screening Using Convolutional Neural Network and Follow-up Digital Mammography", Researchgate.net, 2018, doi: 10.1117/12.2304564.

[15] M. Mateen, J. Wen, Nasrullah, S. Song, Z. Huang, "Fundus Image Classification Using VGG-19 Architecture with PCA and SVD", Symmetry 11, 2019, no. 1: 1.

[16] P. Prateek Sharma (2019, Jul.) Decoding the Confusion Matrix. [Online]. Available: https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

[17] A. Geron, "Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems", O'Reilly Media, Inc., pp. 84-90, 2017.

[18] J. Brownlee (2019, Jan.) How to Choose Loss Functions When Training Deep Learning Neural Networks. [Online]. Available: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/

[19] V. Bushaev (2018, Sep.) Understanding RMSprop — faster neural network learning. [Online]. Available: https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a

SESSION A2

Building up a Development Environment for Fans
Analytical Tool for Technical Predesign

Robert P. Keegan
Ostbayerische Technische Hochschule Regensburg
Turbomachinery Laboratory
Regensburg, Germany
Email: robert.keegan@st.oth-regensburg.de

***Abstract*—The Turbomachinery Laboratory of the OTH Regensburg is mostly used for students education and research in the field of turbomachinery. One of the current research projects in this facility is about building up a development environment for the technical design of fans.**

The current project phase is the first sub-project. Therefore the first step is to develop a sequential path of analytical equations for technical predesign. Starting from customers requirement the aim is to transmit variables to the following CAD and CFD-programs. Due to their universal use in the industry, Microsoft Excel and VBA will be used for the calculations. In the end the hand over variables are defined and exported to the follow up programs.

In the second sub-project the geometry data will be used to automatically generate a 3D model in order to perform a CFD simulation to optimise the blade geometry.

The final sub-project contains a mechanical simulation with FEM about the strength of the fan as well as a verification with real prototypes.

***Index Terms*—analytical tool, CFD, MS Excel, fan, predesign**

I. Introduction

Fans like most technical products have to be predesigned and properly analysed and calculated for maximum efficiency. The state of the art fan design requires an iterative development process including analytical predesign and numerical simulation, Computational Fluid Dynamics (CFD), for optimisation. However, without a careful predesign optimal results are not to be expected.

Before you can build a CFD-simulation, you first have to calculate some parameters and general shape dimensions analytically. Therefore a reliable predesign tool is needed. Furthermore with increasing importance of hydrogen for burner applications a second goal of a predesign to is to estimate the impact of the different fuel gas mixtures for the burner blower performances.

After the predesign it is possible to generate a 3D-geometry with Computer Aided Design (CAD). That geometry can then be loaded and calculated in a CFD-Environment. The geometry can then iteratively be optimised based on the results of the CFD-simulation.

The goal of this project is to build a development environment where fans are calculated analytically, a 3D-geometry is built and a CFD-simulation is performed with one tool. The first sub-project is about the technical predesign of the fans. Therefore the first step is to develop a sequential path of analytical equations for technical predesign. Starting from customers requirement the aim is to hand over variables for the CAD and CFD-programs. Because of their universal use in the industry, Microsoft Excel and Visual Basic for Applications (VBA) will be used for the calculations. At last the hand over variables are defined and exported to the follow up programs.

II. Analytical Foundation

A. Input parameter

For developing the sequential analytical path all input parameters have to be defined at first (see Figure 1). For this development environment the input parameters are:

- environmental pressure p_∞,
- environmental temperature T_∞,
- motor power P_{mot},
- pressure difference Δp,
- volume flow $\dot{V}$,
- specific gas constant R,
- heat capacity ratio κ and
- form of limitation:
  - Space limitation: number of revolutions per second of the motor $n_{mot,ini}$ and maximal external diameter of the fan $D_{a/2,max}$ or
  - Motor limitation: maximal number of revolutions per second of the motor $n_{mot,max}$ and external diameter of the fan $D_{a/2,ini}$.

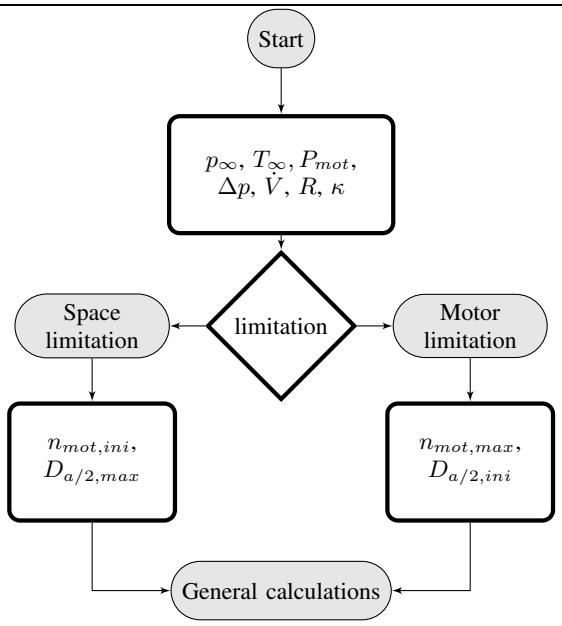


Figure 1. Sequential flow path: Input

These parameters have been chosen, because they are commonly used for fan design in the industry. What has to be known is the environment, where the fan will be operated, the motor, which is powering the fan, pressure difference and volume flow, the fan should achieve.

Almost every fan design is driven by one of two major constraints either by space or the motor. This will be factored in by the decision "form of limitation" in the Input. By deciding for either ones of these constraints the parameter with the index $_{max}$ will be the master parameter and the other one labelled with the index $_{ini}$ will be figured out later on with the help of the CORDIER-diagram (see Figure 7).

B. General calculations

The *General calculations* follow after the input definition. The sequential path is shown as a flow chart in Figure 3. The *General calculations* are defining shape, missing environmental and dimensionless properties of the fan.

A velocity triangle is used to help visualize the different velocity directions of the medium entering and exiting the blades. For the definition of this triangle see Figure 2. Variables with the index $_1$ are marking entrance and index $_2$ marks the exit of the fan.

For the decision of the fan shape the barrier of the specific speed $\sigma = 0.6$ was assumed. This decision is based on the diagram "Menny-Design-Dimension"-Diagram (σ) [6, p. 255, fig. 6.2]. In this diagram between $0.5 < \sigma < 0.7$ it is possible to design either an axial or a radial fan. Therefore the middle between these two borders $\sigma = 0.6$ was assumed to be the divider for the fan shape.

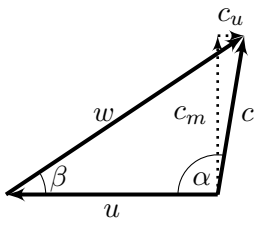


Figure 2. Velocity triangle

C. Fan dimension

If the shape of the fan is defined, the calculation of its dimensions can begin. This path is displayed in Figure 4. These calculations are also mostly based on the "Menny-Design-Dimension"-Diagram [6, p. 255, fig. 6.2]. The diagram is displays the relative dimensions of the fan to the outer or second diameter $D_{a,2}$. The dimensions of each shape are schematically displayed in Figure 5.

Because the width of an axial fan is mostly based on experience, there are no reference points for that in the literature. For this reason an assumption based on real fans of the laboratory was made. The simplfied width of an axial fan is calculated like shown in Equation 1.

$$b_{ax} = \frac{D_a - D_i}{2} * K \tag{1}$$

The width scale factor K is assumed to be $K = 0.75$. This is based on sample fans in the turbomachinery laboratory of the OTH Regensburg.

D. Blade design

Following the general calculations, the shape and dimension design, the blade design is the next part of the sequential calculations in fan design.

The procedure for these calculations is the same as before. Developing a sequential path and then program it in MICROSOFT EXCEL.[1] The sequential path for the blade design will be mostly based on CAROLUS [2, p. 15ff].

Another approach in blade design is the method of AUNGIER [1]. But his approach is focused on gas compressibility, which is usually not that important for fans. Therefore his approach is not the one, the further

[1]Because this paper is only an overview of the current project state, this part is not fully completed yet and therefore not presented here.

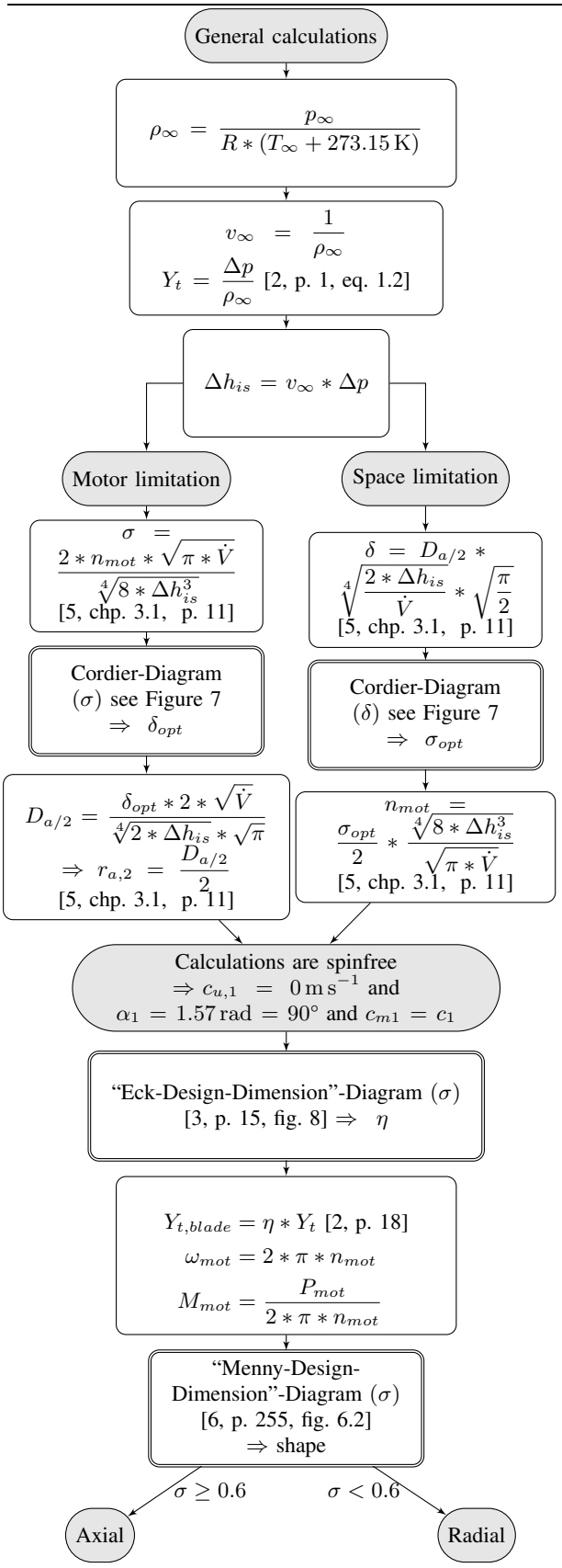


Figure 3. Sequential flow path: General calculations

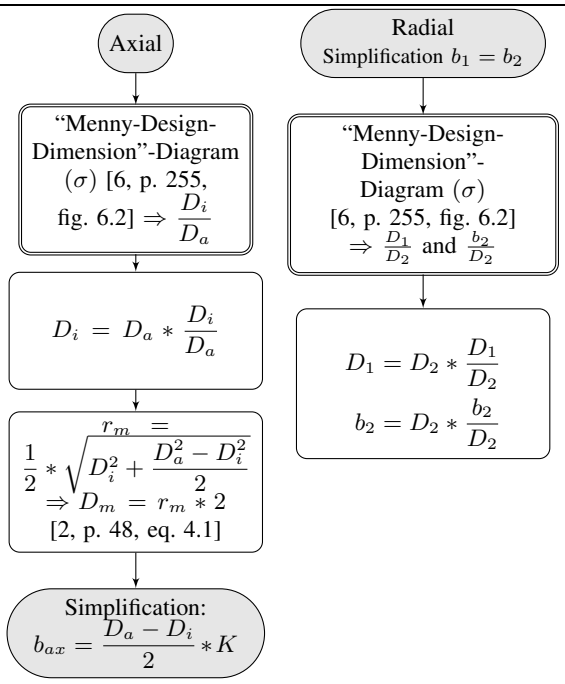


Figure 4. Sequential flow path: Fan dimension

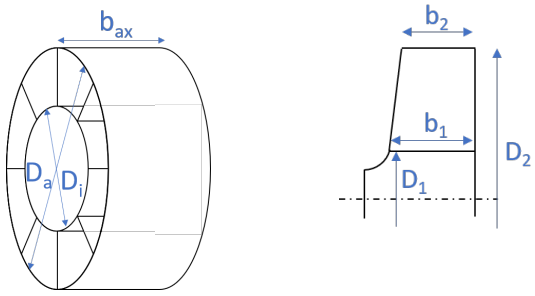


Figure 5. Schematic dimensions axial (left) and radial fan (right)

calculations in the blade design will be following. To consider the influence of compressibility in the predesign tool, a quick estimation of the percentage of the rate of density change will be present. This estimation is according to [3, p. 44]:

$$\frac{\Delta\rho}{\rho} = \frac{1}{2} * \left(\frac{c}{a}\right)^2. \tag{2}$$

The amount of blades will be determined with the ZWEIFEL-number for the axial fan and with the approach of ECK for the radial fan.

III. REALISATION IN MICROSOFT EXCEL

The calculations from section II are implemented in MS EXCEL mostly using standard formulas of the software. Exceptions are performed with VBA.

The MS EXCEL document is split into five sheets: Main, Fan dimension, Blade dimension, Trend curves and VBA related.

A. Main

The sheet **Main** contains the sequential flow path of the *Input* and the *General calculations*.

The parameters of Figure 1 have to be put in the "Input form" in MS EXCEL (see Figure 6). The yellow colored cells are indicating cells, where a input variable has to be filled in. Grey colored cells are indicating cells where a formula calculates the shown value. Exceptions here are the input parameters for R and κ, these cells are colored grey aswell, because the default medium for the calculations is air.[2] But it is possible to change these values. The decision of the limitation is done with a VBA-"List Box". As described in section II the initial value of either $n_{mot,ini}$ or $D_{a/2,ini}$ will be mostly ignored and are calculated from the CORDIER-diagram to an optimal value. If the optimal value exceeds the initial value a warning is displayed for the user.[3]

Next to the *Input* are the *General calculations*. The decisions are performed with "If"-functions. To realize the reading from diagrams it was necessary to digitize them, more to this topic in subsection III-D.

B. Fan dimension

In the sheet **Fan dimension** the calculations are performed, as shown in Figure 4. The dimensions of the fan are calculated for both shapes axial as well as radial. For this case the "Menny-Design-Dimension"-Diagram [6, p. 255, fig. 6.2] was digitized (see subsection III-D).

C. Blade dimension

The blade dimensions as well as the calculation of the amount of blades shall be calculated in this sheet. This step is at the current state still missing and therefore not shown in this paper.

D. Trend curves

A few diagrams have to be digitized first to automatize the decisions in the calculation and open up the possibility to calculate further on.

As an example CORDIER-diagram (see Figure 7), first measuring points (MP) from the three curves (minimal, optimal and maximal) have been taken.[4] The MPs were then exported as a csv-file and loaded in a MATLAB script. This script calculates trendcurves through the MP's using the "fit"-command. The type of curve has to be predefined, for the example of the CORDIER-diagram the curve type was assumed to be a broken rational function with an squared numerator. The output for the fitted δ_{opt}-curve in this example is:

$$\delta_{opt} = \frac{p_1 * \sigma^2 + p_2 * \sigma + p_3}{\sigma + q_1} \tag{3}$$

$$= \frac{0.0205 * \sigma^2 + 0.9237 * \sigma + 0.7093}{\sigma + (-0.0113)}. \tag{4}$$

The same procedure as described previously was also done for "Menny-Design-Dimension"-Diagram [6, p. 255, fig. 6.2] and "Eck-Design-Dimension"-Diagram [3, p. 15, fig. 8].

E. VBA related

In this sheet the data for the VBA-"List Box" in the **Main**-sheet is stored. In the finished Excel document this sheet will be hidden as it should not be modified by any user.

IV. PERSPECTIVE

This section is a short perspective of the further actions. At first the sequential path of the blade design subsection II-D has to be completed. The programming to MS Excel of these equations has to be done in the sheet **Blade dimension**. Also the calculations on how many blades are needed and which profile is used needs to be defined.

In addition, the final project work will include a brief analysis of influencing factors.

ACKNOWLEDGMENT

This paper is written as part of the RARC Regensburger Applied Research Conference 2020. Also it has to be noted that this paper is a summary of the current state of the project and therefore cannot fully represent the final results of this project.

NOMENCLATURE

For nomenclature, abbreviations and indices see tables I to III.

[2]The heat capacity ratio κ is necessary for further calculations of the speed of sound. This is necessary to calculate the percentage ofdensity change of the medium, while it is flowing through the fan. But the sequential path is not completed yet to this stage, therefore κ is not used in the sequential path of the *General calculations*.

[3]In a later revision of this tool it is possible to implement an automatized iteration for an optimal calculation of n_{mot} and $D_{a/2}$, that no boundary condition will be exceeded.

[4]In this case the software WebPlotDigitizer was very useful.

Input								
Description	**Variable**	**Value**	**Unit**	**Value**	**Unit**	**Comment**		
Environmental pressure	p_{∞}	101300	Pa	1.013	bar	Motor limitation: D a/2 ini and n mot max Space limitation: n mot ini and D_max		
Environmental temperature	T_{∞}	20	°C	293.15	K			
Number of revolutions (motor)	n mot max	20	s^-1	1200	min^-1			
Power of the motor	P mot	200	W					
Max outer diameter (fan)	D a/2 ini	450	mm	0.45	m			
Pressure difference	Δ p	1000	Pa	0.01	bar			
Flow rate	V dot	2	m^3*s^-1	7200000	l*h^-1	7200	m^3*h^-1	
Specific gas constant	R	287.058	J*kg^-1*K^-1			# R_air= 287.058 J*kg^-1*K^-1		
Heat capacity ratio	κ	1.402	-			# dry air (20°C) κ=1.402		

Figure 6. Input form in MS Excel

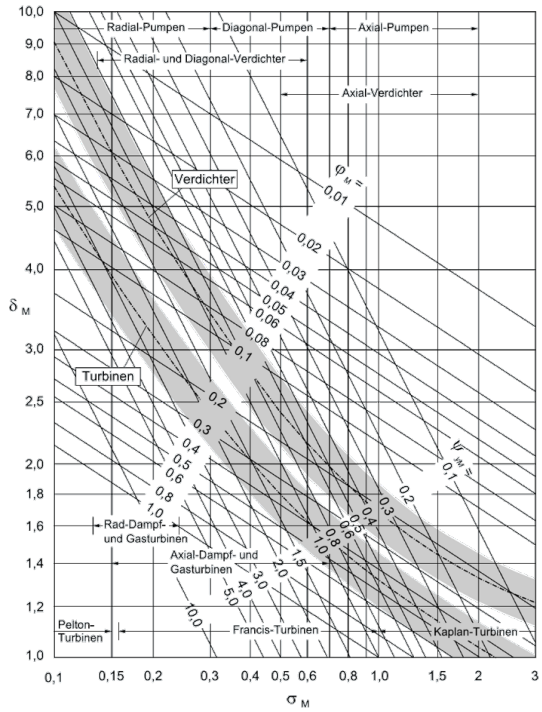


Figure 7. Cordier-diagram [4, p. 1309, fig. 31]

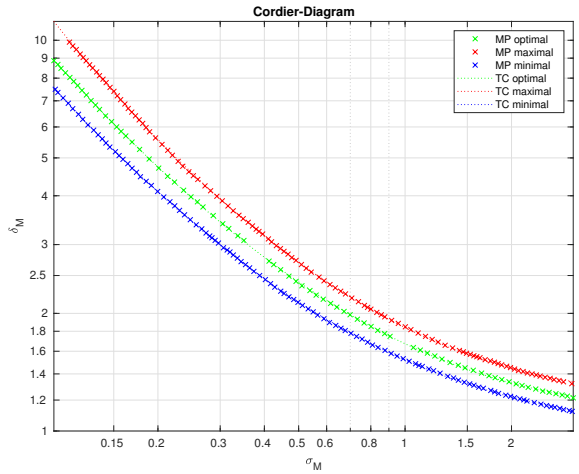


Figure 8. Cordier-diagram digitized with MATLAB

References

[1] R. H. Aungier, *Centrifugal compressors: A strategy for aerodynamic design and analysis*. New York: ASME Press, 2000, ISBN: 0-7918-0093-8.

[2] T. Carolus, *Ventilatoren: Aerodynamischer Entwurf – Konstruktive Lärmminderung – Optimierung*, 4th ed. 2020. 2020, ISBN: 978-3-658-29258-4. DOI: 10.1007/978-3-658-29258-4.

[3] B. Eck, *Ventilatoren: Entwurf und Betrieb der Radial-, Axial- und Querstromventilatoren*, 5., neubearb. Aufl. Berlin: Springer, 1972, ISBN: 3540056009.

[4] K.-H. Grote, J. Feldhusen, and H. Dubbel, *Dubbel: Taschenbuch für den Maschinenbau*, 24., aktualisierte Aufl. Berlin: Springer Vieweg, 2014, ISBN: 978-3-642-38890-3. DOI: 10.1007/978-3-642-38891-0. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38891-0.

[5] A. Lesser, "Vorlesung Strömungsmaschinen," Skript, Ostbayerische Technische Hochschule Regensburg, Regensburg, 2019.

[6] K. Menny, *Strömungsmaschinen: Hydraulische und thermische Kraft- und Arbeitsmaschinen ; mit 36 Tabellen und 47 Beispielen*, 5., überarb. Aufl., unveränderter Nachdr, ser. Lehrbuch Maschinenbau. Wiesbaden: Teubner, 2011, ISBN: 9783519463177.

Table I
FORMULA SYMBOLS

Symbol	Description	Unit
α	flow angle	rad
β	flow angle	rad
δ	specific diameter	$[-]$
Δ	difference	$[-]$
κ	heat capacity ratio	$[-]$
K	width scale factor	$[-]$
ϕ	flow coefficient	$[-]$
ρ	density	$\mathrm{kg\,m^{-3}}$
σ	specific speed	$[-]$
ω	angle velocity	$\mathrm{rad\,s^{-1}}$
a	speed of sound	$\mathrm{m\,s^{-1}}$
b	width	m
c	absolute velocity	$\mathrm{m\,s^{-1}}$
D	diameter	m
h	enthalpy	$\mathrm{m^2\,s^{-2}}$
M	torque	N m
n	number of revolutions	$\mathrm{s^{-1}}$
p	pressure	Pa
P	power	W
r	radius	m
R	specific gas constant	$\mathrm{J\,kg^{-1}\,K^{-1}}$
T	temperature	°C
u	circumferential velocity	$\mathrm{m\,s^{-1}}$
v	specific volume	$\mathrm{m^3\,kg^{-1}}$
$\dot{V}$	volume flow	$\mathrm{m^3\,s^{-1}}$
w	relative velocity	$\mathrm{m\,s^{-1}}$
Y	Work	J
z	amount of blades	$[-]$
Z	ZWEIFEL-Factor	$[-]$

Table II
INDICES

Index	Description
$_1$	entrance
$_2$	exit
$_\infty$	environmental
$_{ax}$	axial
$_{blade}$	blade
$_{calc}$	calculated
$_{ini}$	initial
$_m$	meridian
$_{max}$	maximal
$_{mot}$	motor
$_{rad}$	radial
$_t$	total
$_u$	circumferential

Table III
ABBREVIATIONS

Abbreviation	Decription
CAD	Computer Aided Design
CFD	Computational Fluid Dynamics
MP	Measuring point

Impact of the electrolyte chloride ion concentration and the substrate crystal orientation on the surface morphology of electroplated copper films

Tobias Schwarz[1], Alfred Lechner[1]
[1]Kompetenzzentrum Nanochem, University of Applied Sciences Regensburg, 93053 Germany
Tobias.Schwarz@oth-regensburg.de

***Abstract*— The use of electroplated copper films in semiconductor technology enabled further miniaturization of semiconductor components with a simultaneous increase in performance. In order to manipulate both the electrochemical deposition and the properties of the copper layers, additives are added to the electrolyte. Despite the levelling properties of these additives, increased surface roughness can be observed depending on the crystal orientation of the substrate. The impact of the electrolyte chloride ion concentration and the substrate crystal orientation on the surface morphology was investigated to gain a deeper understanding of this roughness phenomenon in electroplated copper films. The presented results allow a more detailed description and qualitative modelling of the roughness phenomenon.**

***Index Terms*— electrochemical copper deposition, surface roughness, crystal orientation, specific anion adsorption, materials and manufacturing technologies**

I. INTRODUCTION

The electroplating of copper in the semiconductor industry involves the application of additives to the electrolyte to control the deposition process [1, 2]. Such additive systems primarily consist of three components influencing the deposition. During electroplating, a suppressor additive is locally inhibiting the deposition and an accelerator additive is enhancing the deposition. The third component, a leveler additive, also inhibits the deposition but does not interact with the accelerator. In order for these additives to function correctly, chloride ions must be present in the electrolyte. Having the additive package and the chloride ions in the electrolyte a defined deposition behavior can be achieved [3, 4].

The interaction between the accelerator and suppressor is well understood for polyethylene glycol (PEG) as a suppressor and Bis-sodiumsulfopropyl-disulfide (SPS) as an accelerator [4, 8]. The inhibitory effect of the PEG is caused by the formation of a complex consisting of PEG, on the copper surface, adsorbed chloride ions and the copper atoms from the copper metal surface. The SPS accelerates the deposition by decomposing this PEG complex or by slowing down the PEG complex formation. More precisely, the SPS must decompose into Mercaptopropane sulfonic acid (MPS) on the metallic surface to remove the inhibitory effect of the PEG [8]. Corresponding to literature [7, 10], the adsorbed anions on the copper surface counteract the accelerating effect of the SPS by acting as a barrier and preventing the decomposition of the SPS on the metal surface into MPS.

For copper electroplating a copper seedlayer as starting layer is necessary. The grain orientation of this layer depends on the underlying material. A dielectric layer causes a random grain orientation of the copper seed layer, while a metallic layer causes a strong (111) grain orientation. Based on the crystal orientation of the copper seed layer, a roughness phenomenon does occur after copper electroplating.

Depending on the crystal orientation of the copper seedlayer the phenomenon shows a rough or smooth electroplated surface. This results in a rough electrochemical copper deposition on a strong (111) oriented copper seedlayer and a smooth copper surface on a random textured copper seedlayer after the electrochemical deposition. Figure 1 shows the

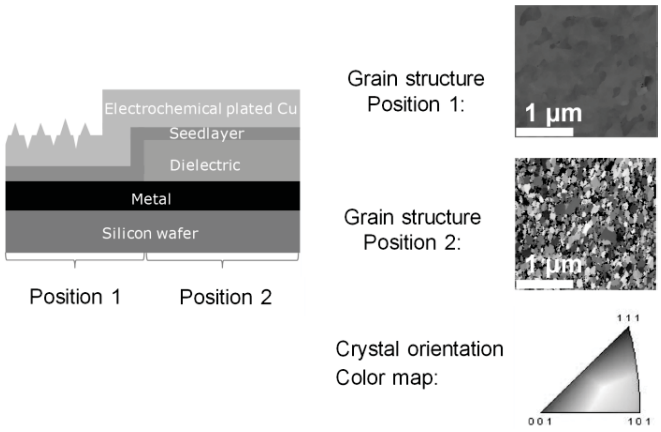


Fig. 1. The roughness phenomenon and the grain structure revealed by electron-backscatter-diffraction (EBSD) images of the copper seedlayer located above a metallic (position 1) and a dielectric layer (position 2). The grain structure on position 1 shows a strong (111) oriented copper seedlayer with individual grains deviating from the (111) orientation. The individual grains appear darker. The grain structure on position 2 is randomly oriented.

electron-backscatter-diffraction (EBSD) images of the copper seed layers above a dielectric or metallic layer, which cause different grain orientations.

This observed behavior is based on the surface interaction of the chloride ions, which are forming a direct chemical bond with the metal surface. This behavior is called specific adsorption and is influenced by the crystal orientation of the metal surface [5]. It is common that specific adsorbed anions manipulate continuous faradaic reactions. Moreover, they are able to alter the potential distribution in the double layer and can completely block reaction sites on the metal surface [5, 6].

Since the specific adsorption of chloride anions is dependent on the substrate crystal orientation [5] and does effect the additive mechanism [5-7] the impact of chloride concentration and the orientation of the substrate on the surface morphology was studied, in order to get further insight into the specific roughness phenomenon.

II. Experimental

The electrochemical copper deposition was carried out in two steps under galvanostatic conditions in a commercial laboratory plating cell (A-56-W SMART CELL, Yamamoto MS). A potentiostat (PGSTAT302N, Metrohm) was used as a power supply. In the first step of electrochemical deposition a current density of 1 A/dm² was applied for 60 s and in the subsequent step a current density of 6 A/dm² for 331 s. The electrolyte system used for copper deposition consists of sulfuric acid (H_2SO_4 Merck, ACS standard, degree of purity >99%), hydrochloric acid, (HCl Merck, ACS standard, degree of purity >99%), copper (II) sulfate pentahydrate ($CuSO_4*5H_2O$ Merck, ACS standard, degree of purity >99%) and a commercial additive package which includes a suppressor, an accelerator and a leveler agent. The substrate used for electrochemical deposition was a layered structure with dielectric and metallic layers below the copper seedlayer as shown in the schematic in Fig. 1.

Electron-backscattering-diffraction (EBSD) was employed with the Hikari-XP-EBSD-Camera attached to the Zeiss-Ultra-55-FE-Scanning-Electron-Microscope to examine the grain orientation and grain size distribution of the copper seedlayer.

The surface morphology of the deposited copper layer was characterized using a Laser Confocal Scanning Microscope (LEXT OLS4100 3D, Olympus). The root mean square height, further abbreviated as Sq, [8] was chosen for the evaluation of the areal surface roughness.

III. Results

The formation of the surface roughness during electrochemical deposition above a strong (111) oriented layer is based on a locally varying growth rate. To determine the local differences on the substrate, both the morphology and the Sq roughness were determined every 30s during the electroplating of the copper layer in order to get an in-sight into the deposition time dependent change of surface morphology. The time dependency of the roughness Sq for a strong (111) textured seedlayer is illustrated in Fig. 2. The linear increase of Sq is based on the raising layer thickness during the electrochemical deposition.

The corresponding change in surface morphology is depicted in Fig. 3 at predefined deposition times. The change in surface morphology reveals that the grain boundaries as well as the individual grains cause a local increased deposition rate and

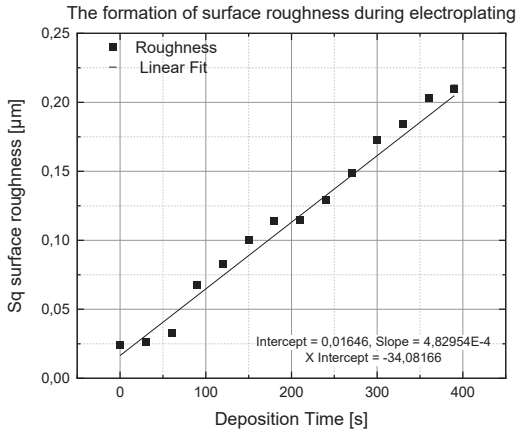


Fig. 2. Development of the Sq surface roughness during electroplating on a strong (111) oriented copper seedlayer above a metallic layer.

thus the roughness phenomenon for a (111) oriented seedlayer. The EBSD image of the (111) oriented copper seedlayer in Fig. 1 shows also individual dark areas. These dark areas correspond to grains with a slight deviation from the (111) orientation. Consequently, it can be concluded that the increased deposition

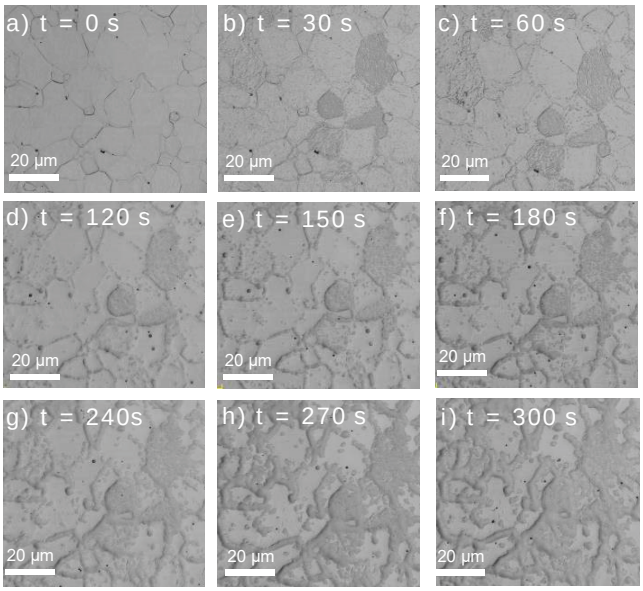


Fig. 3. Surface morphology dependent on the deposition time. At the time step t = 0s the grains and grain boundaries of the seedlayer is visible. The surface morphology at deposition time a) 0 s, b) 30 s, c) 60 s, d) 120 s, e) 150 s, f) 180 s, g) 240 s, h) 270 s, i) 300 s is shown.

rate on those individual grains is linked to their deviation from the (111) grain orientation.

In order to investigate the impact of the specific absorption of chloride based on the crystal orientation as well as on crystal defects (e.g. grain boundaries), the chloride ion concentration of the electrolyte was varied from 2,5 mg/l to 65 mg/l. The Sq

roughness after deposition is illustrated in Fig.4. In dependence of chloride concentration. It can be seen that the roughness profile of the electroplated copper layer on a randomly oriented copper seedlayer, initially decreases strongly with increasing chloride ion concentration and then remains approximately constant. The roughness curve for a strongly (111) oriented copper seed layer above a metallic layer is clearly different. The Sq value initially also decreases strongly with increasing chloride ion concentration. However, the surface roughness raises linearly as the chloride ion concentration in the electrolyte increases further.

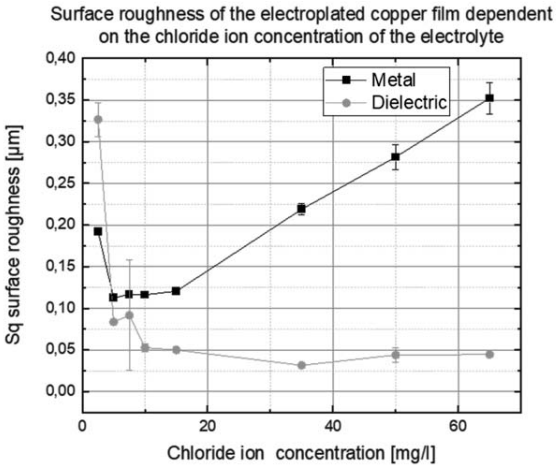


Fig. 4. The surface roughness of the electroplated copper film on a metal and a dielectric layer dependent on the chloride ion concentration of the electrolyte.

The utilized additive package in this studies are designed to address smoothing properties on deposition. Therefore, the initial strong decrease in roughness of both profiles is explained by the fact that chloride ions are essential for the activity of the additives to obtain smooth surface morphology. The Sq value for a copper layer above a dielectric material shows no dependence of larger chloride concentration up to 65 mg/l. It is evident, that the constantly low surface roughness is based on the active surface smoothing capabilities of the additive system. Contrary, the linear increase in the Sq value in the range of 15 mg/l to 65 mg/l for the electroplated copper layer above a metallic layer can be explained with the change of surface morphology as it can be seen in Fig. 5. At a chloride ion concentration of 15 mg/l, the electrodeposited copper layer shows a fine-grained homogeneous structure, whereas with increasing chloride ion concentration, the surface morphology is coarsening with increasing areal cavities.

Consequently, it can be followed that the crystal orientation of the grains as well as the crystal defects of the copper seedlayer affects the specific adsorption of the chloride ions and therefore the functioning of the additives [6-8]. Thus the roughness phenomenon is caused by the locally disturbed activity of the additive system.

IV. Discussion

A qualitative model for the observed partial linear roughness progress and the associated change in surface morphology of the electrochemically deposited copper layer on a strongly (111) oriented copper seedlayer was derived. It was presupposed that an increase of the chloride ion concentration in the electrolyte results in an increase of chloride ions at the surface. Additionally, the model bases on the precondition that the functionality of the commercial additive package has the same surface interaction principle as the PEG and SPS system [8, 9]. As already described, the specific adsorbed anions counteract the accelerating effect of the SPS. Here, the anions are acting as a barriers and are consequently preventing the decomposition of the SPS on the metal surface into MPS. Since the grain orientation and crystal defects of the substrate influence the barrier effect of the specifically adsorbed chloride ions, a local increase in deposition rate can be obtained in accordance with other studies [10].

Thus, a (111) oriented copper surface shows a high barrier effect. This results in a lower deposition rate of electrochemical deposition. Crystal defects on a (111) oriented copper surface reduce this barrier effect and therefore increase locally the deposition rate. In contrast, punctual inhibition is only obtainable at low concentrations, as not all areas can be covered by chloride. This results in the observed fine-grained homogeneous surface structure at an electrolyte chloride ion concentration of 15 mg/l for the (111) textured seedlayer see Fig. 6 a). A further increase in the chloride ion concentration causes a denser specifically adsorbed chloride ion layer, thus increasing the barrier effect. For a chloride ion concentration in the electrolyte of 35 mg/l a local inhibition of electrochemical deposition, as shown in Fig. 6 b), was observed. If the chloride ion concentration is increased even further, the electrochemical deposition is inhibited over a larger area due to the increasing barrier effect of the specifically adsorbed chloride ions.

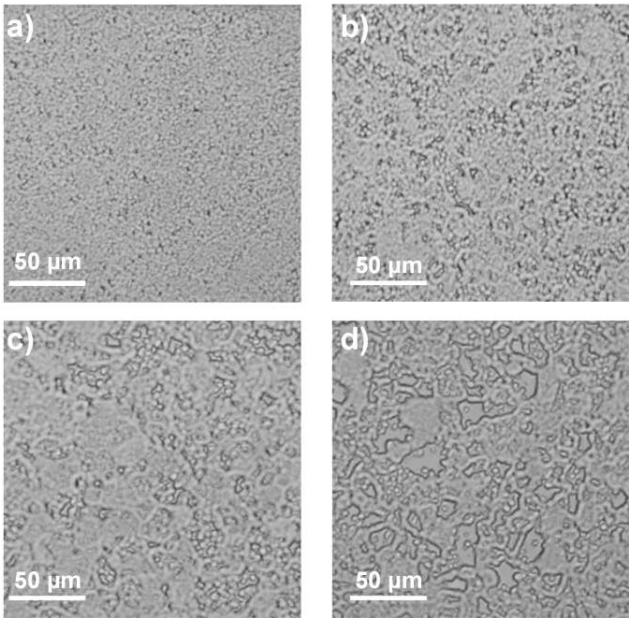


Fig. 5. The surface morphology of the electroplated copper film on a strong (111) oriented seedlayer dependent on the chloride ion concentration a) 15 mg/l, b) 35 mg/l, c) 50 mg/l and d) 65 mg/l is shown.

As it is evident from Fig. 3, the adsorbed chloride ions do not show a barrier effect on crystal defects and grain orientations

besides the (111) texture. Hence, the areal inhibition of the electrochemical copper plating enhances the deposition rate of the crystal defects and grains with a different crystal orientation (see Fig. 6 c). The linear increase in surface roughness with higher chloride ion concentration is due to the growing surface inhibition of the deposition. Since the amount of copper deposited is constant, more copper is deposited locally due to the increasing areal inhibition of the deposition. This locally enhanced deposition with rising chloride ion concentration, is the reason for a higher surface roughness.

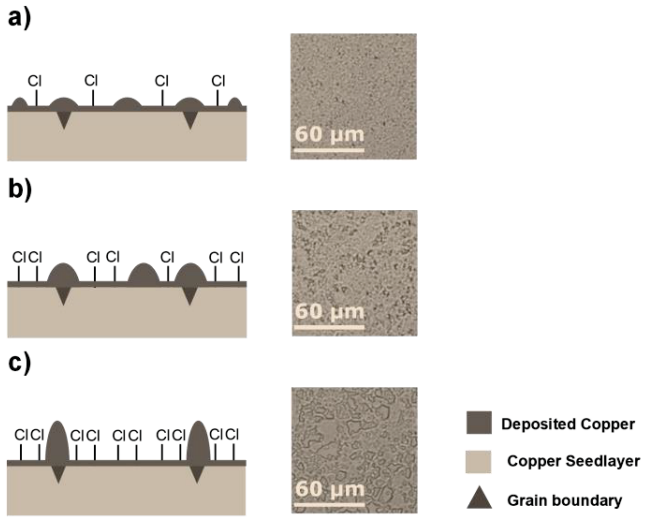


Fig. 6. Schematic model of the electrochemical deposition (left) with the corresponding surface morphology of the electroplated copper film dependent of the electrolyte chloride ion concentration (right). a) electrolyte chloride ion concentration of 15 mg/l, b) electrolyte chloride ion concentration of 35 mg/l, c) electrolyte chloride ion concentration of 65 mg/l. With increasing chloride ion concentration the deposition is increasingly suppressed. The surface morphology is coarsening as the deposition rate of the grain boundaries are enhanced due to the areal suppression. .

V. CONCLUSION

The impact of the electrolyte chloride ion concentration as well as the substrate crystal orientation on the surface morphology of electroplated copper films was investigated in order to get further insight into a well known roughness phenomenon.

The investigation of the time-dependent roughness development revealed that the roughness phenomenon is caused by crystal defects such as grain boundaries and grains with a grain orientation of the copper seedlayer deviating from the (111) crystal orientation.

Furthermore, in accordance with literature [9, 10], the roughness phenomenon could be attributed to the interaction between specifically adsorbed chloride ions, the crystallographic properties of the seedlayer and the plating additives.

A model was developed which qualitatively describes the linear increase in surface roughness and the associated morphology as a function of chloride ion concentration.

IV. REFERENCES

[1] Hooper, R. C., Roane, B. A., & Verret, D. P. (1988). U.S. Patent No. 4,742,014. Washington, DC: U.S. Patent and Trademark Office.

[2] Schmidt, R., Beck, T., Rooney, R., & Gewirth, A. (2018, May). Optimization of electrodeposited copper for sub 5 µm L/S redistribution layer lines by plating additives. In 2018 IEEE 68th Electronic Components and Technology Conference (ECTC) (pp. 1220-1225). IEEE.

[3] Kanani, N. (2004). Electroplating: basic principles, processes and practice. Elsevier.

[4] Vereecken, P. M., Binstead, R. A., Deligianni, H., & Andricacos, P. C. (2005). The chemistry of additives in damascene copper plating. IBM Journal of Research and Development, 49(1), 3-18.

[5] Magnussen, O. M. (2002). Ordered anion adlayers on metal electrode surfaces. Chemical reviews, 102(3), 679-726.

[6] Brisard, G., Bertrand, N., Ross, P. N., & Marković, N. M. (2000). Oxygen reduction and hydrogen evolution–oxidation reactions on Cu (hkl) surfaces. Journal of Electroanalytical Chemistry, 480(1-2), 219-224.

[7] Braun, T. M., Josell, D., Silva, M., Kildon, J., & Moffat, T. P. (2019). Effect of Chloride Concentration on Copper Deposition in Through Silicon Vias. Journal of The Electrochemical Society, 166(1), D3259-D3271.

[8] Hai, N. T., Krämer, K. W., Fluegel, A., Arnold, M., Mayer, D., & Broekmann, P. (2012). Beyond interfacial anion/cation pairing: The role of Cu (I) coordination chemistry in additive-controlled copper plating. Electrochimica acta, 83, 367-375.

[9] Tan, M., & Harb, J. N. (2003). Additive behavior during copper electrodeposition in solutions containing Cl−, PEG, and SPS. Journal of The Electrochemical Society, 150(6), C420-C425.

[10] Nguyen, H., Huynh, T. M. T., Flügel, A., Arnold, M., Mayer, D., & Broekmann, P. (2013, October). Towards An Atomistic Understanding of the Activation of Plating Additives At the Copper/Electrolyte Interface. In Meeting Abstracts (No. 38, pp. 2398-2398). The Electrochemical Society.

Investigation of Continuous Fiber Filling Methods in Additively Manufactured Composites

Leopold Grimm
Laboratory for Composite Technology
Department of Mechanical Engineering
OTH Regensburg
Germany
Email: leopold.grimm@st.oth-regensburg.de

***Abstract*—Fused deposition modeling (FDM) is an additive manufacturing method which allows layer-by-layer build-up of a part by the deposition of thermoplastic material through a heated nozzle. The technique allows for complex geometries to be made with a degree of design freedom unachievable with conventionally manufacturing methods. However, the mechanical properties of the themoplastic materials used are low, compared to typical engineering materials. In this work, additively manufactured fiber-reinforced composites with glass fibers are investigated, wherein fibers are embedded into a thermoplastic matrix. The specimens are tested for tensile property following DIN EN ISO 527. The results are presented and the conclusions are given about the mechanical effects of the modification of the fill algorithm. The benchmark of two fill algorithms demonstrates that the tensile strength of a 3D printed specimen depends on it. The verification by microscopy analysis is used to find out the impact of unknown print quality on the mechanical properties of 3D printed structures. Specimens evaluated in this study were produced by a *Mark One* 3D printer of MARKFORGED by varying the fill algorithm. The experimentally determined stiffness was found to be 19600 ± 1050 MPa and 18900 ± 2100 MPa for algorithm *Concentric* and *FullFiber*, respectively. The determined strength were found to be 589 ± 31.9 MPa and 670 ± 12.4 MPa. The mechanical examination shows a failure behavior with a strong dependency on the manufacturing process. The track-dependent damage can be seen as a limiting factor for the mechanical properties.**

***Index Terms* – additive manufacturing, 3D printing, fiber-reinforced composites, mechanical properties**

July 14, 2020

I. INTRODUCTION

The industry faces the challenge of continuously producing lighter structures with the same strength for energy and cost savings. Currently, steel and aluminium constructions are used, as well as shot fiber-reinforced injection moulding and occasionally fiber-reinforced plastics (FRP) [16]. Due to the expensive and complex autoclave and resin transfer moulding process, the material is less distributed by continuous fiber-reinforced plastics [17]. The use of FPR is often implemented in flat panel designs and is therefore very limited in constructional variations.

Three dimensional (3D) printing or Rapid Prototyping (RP) is another description of additive manufacturing. Additive Manufacturing can be divided into several categories: Fused Deposition Modelling (FDM), Selective Laser Melting (SLM), Stereolithography (STL) or Laminated Object Manufacturing (LOM) [4]. This process produces components from computer-aided design (CAD) programs. Additive manufacturing is a technology in wich open source software is increasingly used. This can be attributed due to the widespread use of low cost 3D printers.

Some low-cost desktop 3D printers utilize FDM as the manufacturing process. FDM forms a 3D geometry by assembling individual layers of extruded thermoplastic filament. The FDM manufacturing process is useful for producing prototypes and in some cases, it can be used to produce low loaded components. FDM components are formed by an additive manufacturing process combining successive layers of molten thermoplastics. Due to this process delamination of the component layers can occur resulting in premature failure.

In addition to that, new manufacturing methods, like additively manufactured fiber-reinforced plastics, are emerging as potentially promising new systems. This process has become highly popular with researchers for the design and manufacturing of complex 3D components. These structures can be used in complex lightweight constructions. The force-oriented manufacturing makes it possible to avoid unnecessary material on the part. New designs can be created due to the geometrically design of freedom, compared to construction of conventional fiber-reinforced plastics. Compared to conventional FRP, additively manufactured composites differ strongly in their method, so that the implementation of a high fiber volume content to increase stiffness and strength is difficult due to lower process pressures. In the same way, a track-dependent structure is introduced into the part due to the influence of manufacturing processes.

To determine, if additive manufactured fiber reinforced plastics can be used for functional component structures, the material characteristic has to be investigated and tested in the context of the research and development project “FIBER-PRINT”. The subject of the “FIBER-PRINT” project is the examination and further development of the behavior of additive manufactured composites [2], [14].

In this paper, the influence of additive manufactured 3D printing structures is investigated. Two different fiber-filling methods were assessed and comparisons were made in terms of mechanical properties and part quality. The continuous glass fiber-reinforced specimens are produced using the MARKONE

3D printer. Production-related cross influences can have an effect on the component quality. The aim is to identify and evaluate these.

II. Materials

In order to investigate the effect of possible production influences, test specimens of different types are required. These should have different properties in order to assess the influence of manufacturing. The glass fiber-reinforced test specimens are manufactured with two different filling algorithms. In addition to the identification of the fibre volume content, microscopy examinations are carried out to determine the quality of the specimens. The mechanical load capacity is determined quantitatively by destructive material testing.

A. Material extrusion processes

Currently, new thermoplastic materials are becoming available. These include thermoplastic filaments with embedded metallic particles or reinforced with short carbon fibers [13]. Additionally, there are 3D printer commercially available which reinforce 3D printed parts with continuous carbon fiber, glass fiber or aramid (Kevlar) fiber filaments. This 3D printer of MarkForged [9] called Mark One is designed to produce FDM printed components. The 3D printer reinforces FDM printed parts by embedding tracks of fibers into the components geometry. Specifically, this new FDM printing methods aim to increase the strength of 3D printed parts. Such components can be used for functional products rather than producing non-functional scale models.

To achieve an increase in mechanical properties, the fiber content can be increased to a limited range. Beside the mechanical properties, the processing properties of fiber and matrix are also relevant. From a processing point of view, the filament diameter, surface preparation and drapability of the fibers are of interest [16]. The rheological and morphological behaviors of the matrix are important. The impregnation behavior of the reinforcing structure and the flow ability of the matrix are thus properties relevant to processing. The process pressure, the process temperature and the process time have a direct influence on the composite properties.

B. Cross-Influence

In addition to the above-mentioned process-determining parameters such as temperature, process pressure and time, the additive manufactur process offers a wide range of variation options that influence component quality. In the FDM process, nozzle diameter, mass flow, cooling time and the filling algorithm have a strong influence on component quality. If these parameters are changed, they have an effect on the height of the component, the production time and the composition of the filling structure.

C. Fill algorithm

In the production of prototype components made of pure plastics, there is the option of optimizing the infill structure. In the case of components that are not subject to high loads, such as a building model, the user is only interested in the surface structure. Subsequent interest is here how the building

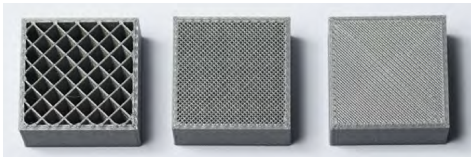


Fig. 1. Fill structure from left to right in percent: 20, 50 and 75 [11].

is constructed from the inside. Often, resources are saved by saving material and the printing time is significantly reduced if the interior is not completely filled. Structural patterns are used for this purpose and are shown in Figure 1 in different filling densities. For components with load-bearing structures the variation of the infill quantity is not useful. Therefore they are always printed with 100 percent infill mass.

The filling algorithm can be chosen between two variants for the Mark One of the company Markforged. These are *Concentric* and *FullFiber* which are shown in Figure 2. Compared to the press type manufacturing methods, additive

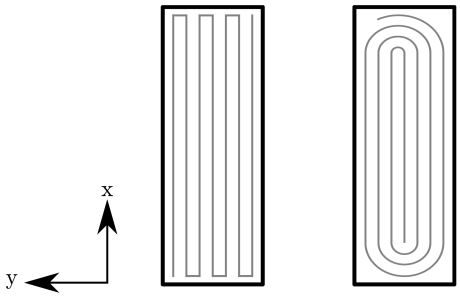


Fig. 2. Fill algorithm *FullFiber* and *Concentric*.

manufacturing can only be performed under ambient pressure (in some manufacturing processes pressures of more than 50 bar are applied). The bonding interface between the single tracks can be responsible for a possible early material failure. Such tracks have the character of ropes, the left-hand variation in Figure 2 probably fails earlier than the right-hand variant. The arrangement similar to a loop connection could prevent the early separation of the outer strand.

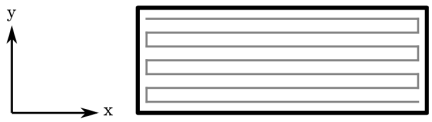


Fig. 3. Continous fiber-reinforcement orientation in a layer.

D. Specimen manufacturing

For the following experiments, test specimens were produced additively. In this case, an important assumption is violated: Due to the additive production, the specimen is no

longer homogeneous over the cross-sectional area. The homogeneity of the specimen geometry is disrupted by the extrusion manufacturing process. In the Y-direction no homogeneous structure is possible. The depositing of the tracks – under atmospheric pressure – forces weak points between the printed tracks in the plane. In the thickness direction, the homogeneity is violated by the layered structure. The test specimens are produced with the printer MARK ONE of the manufacturer MARKFORGED.

The multi-layer composites produced from the fiber-reinforced composites have different layer orientations of their unidirectional (UD) individual layers depending on the type of load. UD-like individual layers can only be produced with glass fibres due to the restrictions of the filling algorithm method of the MARKFORGED slicer software. A discontinuously material deflection of the fiber tracks is not possible. The component is manufactured like Figure 3.

A further restriction is the necessary use of base, top and wall layers. This means that no homogeneous surfaces over the specimen cross-section can be manufactured. Supporting structures at the edge are made exclusively of Nylon. These structures cannot be printed with reinforcing fibers in the program. Figure 4 shows a schematic structure of a specimen geometry and is intended to illustrate the limitation of the unsteady distribution of the fiber content. The excessive increase of the cross-section of the specimen implies that the proportion of the boundary layer becomes smaller and smaller. However, this type of design is not effective, fibre composite components are mainly used in a flat structural design. This is because of the specific stiffness and strengths to the weight ratio of composites.

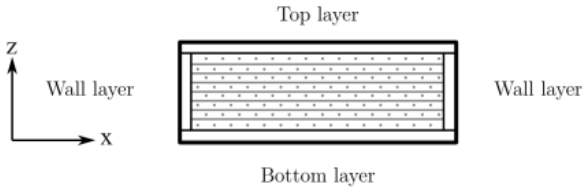


Fig. 4. Standard structure of a 3D printed fiber-reinforced composite.

III. METHODS

The mechanical performance and quality of 3D printed composite parts manufactured using two different printing algorithms are investigated here. Various experiments are performed to quantify key mechanical properties, which depend on the fill algorithm, and optical microscopy is used to examine the quality of the parts.

In the following, the cross influences forced in the production process by the choice of the filling algorithm are considered. The tensile test specimens have been manufactured using two different filling algorithms. A distinction is made between the CONCENTRIC and FULLFIBER algorithm.

A. Optical microscopy

The realization regarding a homogeneous cross-sectional area is limited by the additive manufacturing process. For a qualitative assessment, micro section specimens are prepared. The microscopic images are generated on the cross-sectional surface of the unidirectional specimen. A digital microscope of the VHX-5000 series from KEYENCE CORPORATION is used. For the microscopy examination, the cross-sectional area is examined on the sample with the FULLFIBER filling algorithm.

B. Fiber volume

The determination of the fiber volume content should first be implemented using the geometric data from the slicer software. This is because the number of tracks and layers is known exactly. The dimensional accuracy requires a larger tolerance ($\pm\,0.2\,\mathrm{mm}$) of the additive manufacturing, they can be differ to the real model. The total fiber volume content can be determined from the number of strands deposited and the knowledge of the fiber volume content from one strand. The optical calculation via the area ratio of fiber and matrix provides more feasible results. Due to the almost identical refractive index of glass and plastic, optical differentiation is much more difficult than, for example, when using carbon fibers. For the determination of the individual surfaces, the geometric data is known and refer purely to the data from the CAD model and the specifications in the slicer software. The schematic structure can be seen in Figure5. The eight layers

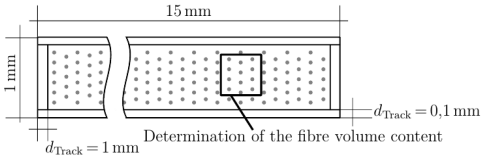


Fig. 5. Schematic test specimen structure of an additive manufactured composite.

with fiber-reinforcement are shown as one total area simplified. A total of 10 layers with a layer thickness of 0.1 mm are in the composite. The strand width is assumed to be 1 mm. To determine the fiber volume content $\varphi_{\mathrm{F,\,tot}}$ the fiber volume content is calculated using an area segment via the area ratio. The selection of the area element can be seen in Figure 5. By using automated colour comparison functions, fiber volume content of $\varphi_{\mathrm{Finfill}} = 33.7\,\%$ can be determined. To calculate the fiber volume content for the entire cross-sectional area, the area percentage of the fibers in the infill is important. The fiber area A_{F} is calculated by converting and adjusting the reference area from Figure 5 with

$$A_{\mathrm{infill}} = 8 \cdot A_{\mathrm{infill,lay}} = 8 \cdot (13\,\mathrm{mm} \cdot 0.1\,\mathrm{mm}) = 10.4\,\mathrm{mm}^2 \quad (1)$$

to

$$A_{\mathrm{F}} = \varphi_{\mathrm{F}} \cdot A_{\mathrm{infill}} = 0.337 \cdot 10.4\,\mathrm{mm}^2 = 3.051\,\mathrm{mm}^2. \quad (2)$$

The fibre volume content of the entire sample is calculated from the area ratio of fibre area fraction to a total area fraction and is

$$\varphi_{\mathrm{F,\,tot}} = \frac{A_{\mathrm{F}}}{A_{\mathrm{total}}} = \frac{3.051\,\mathrm{mm}^2}{(15 \cdot 1)\,\mathrm{mm}^2} \approx 23.4\,\%. \tag{3}$$

C. Uniaxial tensile experiment methodology

Samples for mechanical testing were fabricated using a MARK ONE desktop 3D printer. The sample geometry was created according the DIN EN ISO 527-5 [8] using the Type A geometry. The geometry used in this study and critical dimensions are shown in Figure 6. The test specimen geometry

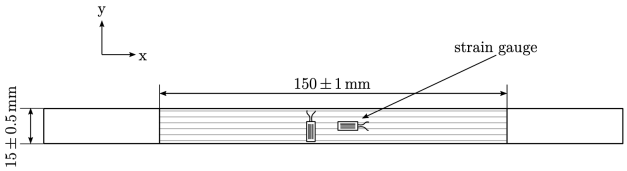


Fig. 6. Tensile test specimens according to DIN EN ISO 527-5 [8] in Typ A orientation with 0 grad fiber orientation.

was crated using a CAD software. The specimen geometry was exported as a stereolithography file (STL) and loaded into the 3D printer slicing software package eiger.io from MARKFORGED [10]. The EIGER software is required in conjunction with the MARKFORGED 3D printer, as it uses its own encrypted file type (.mfp). All specimen were printed with a Nylon filament with glass fiber-reinforcement. Table I summarizes the printing parameters used to manufacture the test specimen.

TABLE I
TEST SPECIMEN PRINT PARAMETERS.

Print Parameters	Value
Layer Height (mm)	0.1
Infill Percentage (%)	100
Infill Orientation (degrees)	0
Number of infill layers	8
Number of floor layers	1
Number of wall layers	1
Number of top layers	1
Total number of layers	10

The 3D printed specimen are reinforced with two different types of fill algorithm. The first are printed with a *FullFiber* pattern like demonstrated in Figure 3. The second variant is printed with *Concentric* fiber pattern. The number of reinforcement tracks in the cross-sectional area are equal. The different types of fill algorithm used in this study was selected to characterize the effect of different fiber track filling types on 3D printed specimen. The reinforcement of the test specimens with glass fiber is shown in Figure 7.

1) Experiment testing parameters: The fiber reinforced 3D printed specimens were evaluated by performing tensile tests. The test setup used to evaluate the 3D printed specimen is shown in Figure 8. The tests to determine the tensile stiffness

Fig. 7. the investigated tensile test specimens according to DIN EN ISO 527-5 [8] in Typ A orientation with different fill algorithm. Top: *FullFiber*, Bottom: *Contentric*.

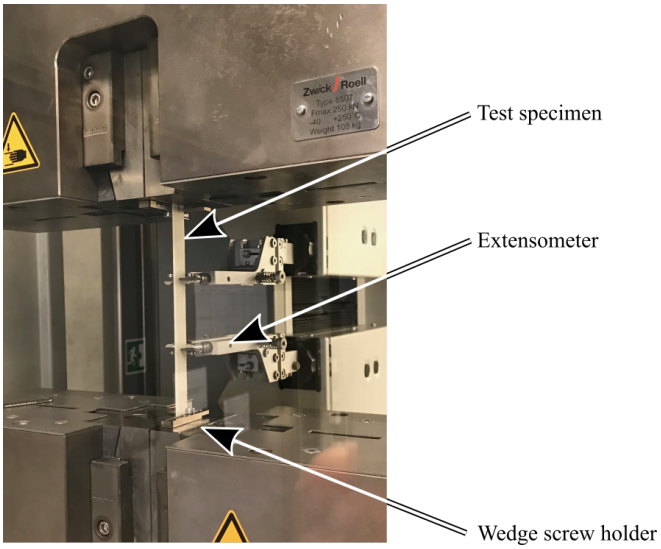


Fig. 8. Mechanical testing setup to evaluate the tensile properties of the glass fiber-reinforced 3D printed specimen.

and strengths are carried out on a universal force testing machine from ZWICKROELL with the designation Z250. The 250 kN load cell was used to measure the loads on the test specimen. A wedge screw sample holder of type 8406 with a maximum force of 30 kN is used to fix the flat specimen. Strain of the test specimen was measured using a 50 mm gauge length extensometer. The samples were loaded at a rate of 2 mm/min. The specimen is tested until it fails due to breakage.

The technical stress σ in the layer plane is calculated according to:

$$\sigma = \frac{F}{A} \tag{4}$$

with:

- σ the stress, in MPa,
- F the measured force in N,
- A the beginning cross-section area in mm^2.

The technical strain determined by means of an extensometer is calculated by:

$$\varepsilon = \frac{\Delta L_0}{L_0} \tag{5}$$

with:

- ε the strain value, as dimension 1 or in percent,
- ΔL_0 the strain of the test piece in mm,
- L_0 the gauge length in its initial state in mm.

IV. Results

The following sections presents the results of the investigations.

A. Microstructure analysis

Figure 9 shows the cross-sectional area of a tensile test specimen, which was extracted in the centre of the specimen. The differences between the outer layers and the infill are

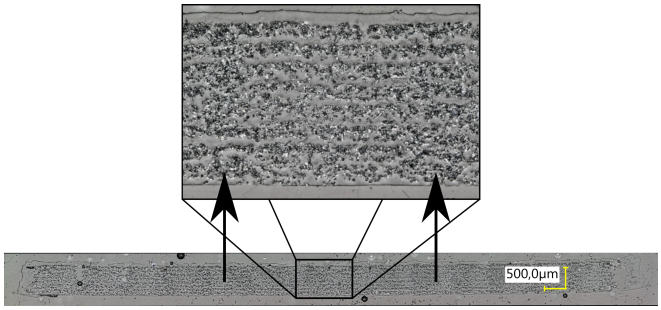


Fig. 9. Microscope image of the fiber-reinforced specimen, showing the cross section area.

clearly visible in the bottom image. The sample width consists of 16 strands, the outer layers are not reinforced. The glass fiber-reinforced infill has a visible contrast to the transparent Nylon. The layer structure is also clearly visible in the image and is composed as follows: The Bottom layer is partly strongly melted and the fibre reinforcement forces a mixing from the first reinforcement fibre layer to the bottom layer. The eight successive infill layers are clearly visible in the enlarged image detail. The test specimen is finished with a top layer. Between the individual tracks, fibre accumulations can be seen in the cross-sectional area (indicated by the arrows).

B. Mechanical testing results

Mechanical testing was performed on two sample configurations *FullFiber* and *Concentric* to examine the effect of different fiber-reinforcement fill algorithm on the mechanical properties. The stress-strain curve shown in Figure 10 and

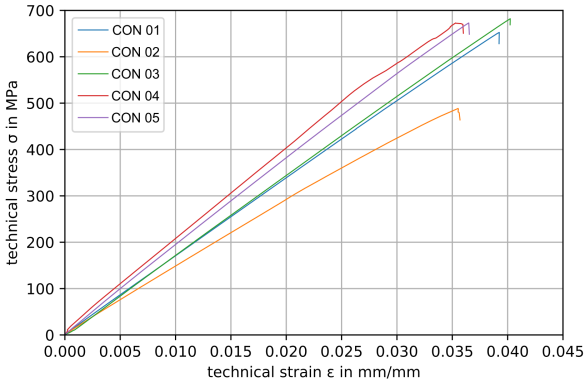


Fig. 10. Stress-strain curves for the specimens with the *Concentric* fill algorithm.

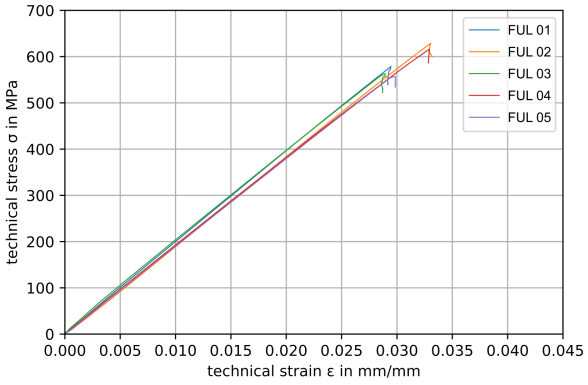


Fig. 11. Stress-strain curves for the specimens with the *FullFiber* fill algorithm.

11 demonstrate the possible effect of variations of the fill algorithm on 3D printed specimen on the stress-strain behavior. The technical strain ε in % is shown on the abscissa and refers to the start length. The ordinate indicates the technical stress σ in MPa. Figure 10 shows the results of the 5 specimen with the *Concentric* fill algorithm in the stress/strain behavior. Four specimen failure just below 700 MPa. One specimen with the label CON 02 fails at a significantly lower stress of less than 500 MPa. This specimen is to be seen as an outlier, no abnormalities were detected during the test. In the tensile tests according to the *FullFiber* fill algorithm, the stress/strain behavior is almost identical (see figure 11). No abnormalities were found in this tensile behaviour on additive-manufactured longitudinal test specimens. All specimens show fracture failure at a stress of about 600 MPa. In both figures the failure is not the complete failure it is the first failure of one or more tracks. All specimens show a linear strain behavior until failure of the specimen. The stiffness and strength of each sample configuration was determined from the stress-strain curves shown in Figure 10 and 11. The average stiffness and strength for the two specimen configurations are shown in Figure 12. This Figure shows the resulting stiffness and strength with the standard value deviation for each specimen configuration. Figure 12 demonstrates that a change of the fiber-reinforcement fill algorithm does not significantly change the stiffness in comparison. The stiffness with fill algorithm *FullFiber* is 19600 ± 1050 MPa. The reinforcement with the fill algorithm *Concentric* has a stiffness of 18900 ± 2100 MPa.

The ultimate tensile strength of the fiber reinforced test specimen was also examined. Figure 12 shows the resulting average ultimate tensile strength and standard deviation for the two fiber-reinforced specimen configurations. This Figure indicates that the fill algorithm *Concentric* increases the ultimate tensile strength of the fiber reinforced 3D printed components. All specimens in this comparison failed during testing after reaching there maximum of strength. The specimen configuration *FullFiber* failed at a strength of 589 ± 31.9 MPa. The stronger configuration *Concentric* failed at a strength of 670 ± 12.4 MPa.

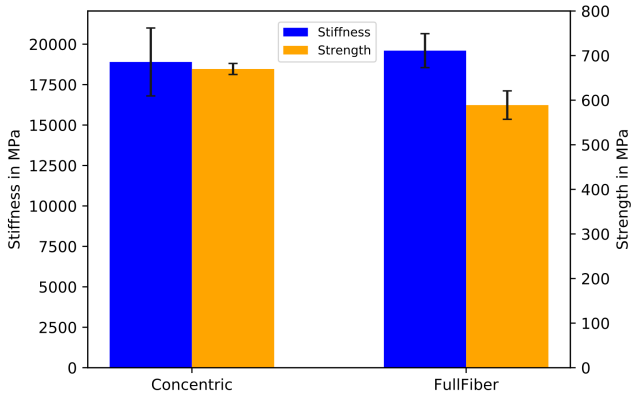


Fig. 12. Comparison of stiffness and strength for the two specimen configurations.

The damage behavior of additively manufactured fiber-reinforced plastics can be considered as special. The distribution of damage depends on the printed tracks. Damage often starts at the outer tracks. All further tracks fail sequentially due to the increasing stress in the cross section until the specimen is completely destroyed as is shown in Figure 13.

Fig. 13. Destroyed glass fiber-reinforced specimen with fill algorithm *FullFiber*.

V. Discussion

Fused deposition modelling has been investigated as a low-cost manufacturing method for fiber-reinforced composites materials. An important aspect of composite materials is the consolidation of fiber and matrix as well as between the single layer. Traditional industrial grade composite manufacturing techniques such as autoclave process and automated tape placement (ATP) use additional consolidation rollers to improve the final part quality [5]. Compared to this, 3D printers are simple in design and lack the ability of apply additional pressure and heat to the part.

The fiber volume content is calculated to 23 percent. It is substantially lower than the fiber volume contents of over 50 percent used as a common literature value for autoclave production [15]. The fiber volume content can be increased by removing the outer nylon layers. The fiber volume content is then increased to 33.7 percent.

Multiple studies report on an increase in mechanical properties from unreinforced to reinforced filaments, but to be used as a structural material the absolute strength and stiffness must increase as well as the consistency and quality of manufactured parts. The test results of the Mark One continuous fiber printed parts presented here indicate good mechanical properties which are an order of magnitude higher than typical Nylon materials, although still significantly lower than unidirectional composites made with traditional manufacturing methods with a strength/stiffness of 1500 MPa/135 GPa [3]. Placement of continuous fiber filament is limited by a number of geometric and processing constraints, such as a minimal deposition length and minimal corner radii.

The image produced from the optical microscopy analysis are of sufficient quality for qualitative analysis. Some error was noted in the alignment of specimen plane to the ground layer, as can be seen in Figure 9. This can be a result of the manual polishing approach that was used. Between the individual layers of the fiber-reinforcement, there are gaps in which no fibers are present. At the borders of a track, fibre accumulations which can force a lack of consolidation. For this reason, the specimen fail in a strand-dependent way without delamination between the single layers, which is the characteristic of conventional composites.

The failure start location for the fiber-reinforced 3D printed specimen was consistent for all tested specimens. The failure started in the location where the fiber track begins for the specimen in the schematical application in Figure 2. All manufactured specimens using the Mark One 3D printer have a start location of failure. Figure 2 demonstrates that understanding the start location of the fiber-reinforcement is critical for manufacturing functional components. Manufacturing functional structure components using this 3D printing method, the start location of the fiber-reinforcement should be placed in a position of low loading.

The stress-strain plots in Figure 10 and 11 demonstrate the effect of fiber-reinforcement on the behavior of fiber reinforced 3D printed components. The stiffness does not depend on the fill algorithm. The ultimate strength and ultimate strain depends on the fill algorithm. The significantly higher values are detected by the *Concentric* fill algorithm. A reason for that can be the start location for the continuous fiber-reinforcement as discussed previously.

VI. Conclusion

From the current body of work on additive manufactured composite structures it must be concluded that the quality of a 3D printed part is still low compared to classical aerospace grade composite materials.

In this work the cross-section influence of 3D printed composite parts has been presented, and the performance of two different fill algorithms of the fiber reinforced structure have been benchmarked with mechanical testing and optical microscopy. Both series of glass fiber-reinforced composite specimens were produced using the Mark One printer. The tensile strength of the *Continuous* fiber printed parts is significantly higher than the *FullFiber* specimens, whereas the tensile stiffness is nearly equivalent.

The modified filling algorithm confirmed that track adhesion has an effect on strength. Presumably, it was not possible to determine the real permissible strength with the existing equipment, as deviating specimen geometries in the clamping area had an effect on the strength. The fracture behavior is

track-dominant. Delamination in the individual layers could not be detected.

Compared to conventionally produced test specimens, it can be stated that the effects of the additive manufacturing process has an enormous influence on the material behavior. The manufacturing process at ambient pressure also forced a problem with regard to consolidation. Under this limitation, additive fiber composites cannot be directly compared with autoclave components made of prepregs. With regard to the limitations imposed by the manufacturing process,the stiffness and strength is one magnitude lower.

This study provides a basis for predicting the tensile properties of fiber reinforced 3D printed structures. Further research is required to fully characterize the mechanical behavior of these fiber reinforced 3D printed structures.

ACKNOWLEDGMENT

This research work and its results are part of the project FIBER-PRINT, which is funded by the Federal Ministry of Defence of the Federal Republic of Germany. In cooperation with the Bundeswehr Research Institute for Materials, Fuels and Lubricants (WIWeB). Furthermore, I would like to thank Professor Dr.-Ing. Ingo Ehrlich, the head of the LABORATORY COMPOSITE TECHNOLOGY and Christian Pongratz for their initial ideas, their guidance and their support in this investigation.

REFERENCES

[1] BLOK, L. G.; LONGANA, M. L.;YU, H.; WOODS, B.K.S.: *Evaluation and prediction of the tensile properties of continuous fiber-reinforced 3D printed structures.*, Additive Manufacturing, Vol. 22, pp. 176-186, 2018.

[2] EHRLICH, I.: *Personal Communications.* Ostbayerische Technische Hochschule (OTH) Regensburg, Faculty of Mechanical Engineering, Laboratory Composite Technology, Regensburg, 2019

[3] GEBHARDT, A.: *Additive Fertigungsverfahren.* Edition 5, Carl Hanser Verlag, Munich, 2016

[4] GIBSON, I.; ROSEN, D.; STUCKER, B.: *Additive manufacturing technologies: 3D printing, paid prototyping, and direct digital manufacturing.*, 2nd ed. Sringer, New York 2015

[5] LUKASZEWICZ, C.; POTTER, K.D.: *The engineering aspects of automated prepreg layup: history, present and future.* Composite Part B Eng., Vol. 43, 2012.

[6] MELENKA, G.W.; CHEUNG, B.K.O.; SCHOFIELD, JS.; DAWSON, M.R.;CAREY, J.P.: *Evaluation and prediction of the tensile properties of continuous fiber-reinforced 3D printed structures.*, Composite Structures, Vol. 153, pp. 866-875, 2016

[7] MORI, K,; MAENO, T.; NAKAGAWA, Y.: *Dieless forming of carbon fibre reinforced plastic parts using 3D printer.* Procedia Eng, Vol. 81, pp. 1595-600, 2014.

[8] N. N.: *Plastics - Determination of tensile properties – Part 5: Test conditions for unidirectional fibre-reinforced plastic composites (ISO 527-5:2009).* Normenstelle Technische Grundlagen (NATG) im DIN Deutsches Institut fr Normung e.V., Beuth Verlag, Berlin, 2009

[9] N. N.: *3D printer Mark One* https://www.mark3d.com/de/mark-one/#mark-one, Februar 2019

[10] N. N.: *Markforged slicing software* https://www.eiger.io, Februar 2019

[11] N. N.: *Selecting the optimal shell and infill parameteres for FDM 3D Printing.* https://www.3dhubs.com/knowledge-base/selecting-optimal-shell-and-infill- parameters-fdm-3d-printing, Juli 2019

[12] N. N.: *Selecting the optimal shell and infill parameters for FDM 3D Printing.* https://www.3dhubs.com/knowledge-base/selecting-optimal-shell-and-infill-parameters-fdm-3d-printing, Juli 2019

[13] N. N.: *ColorFab BrassFill 3D Printing Filament.* https://www.colorfabb.com/bassfill, Juli 2019

[14] PONGRATZ, C.: *Personal Communications.* Ostbayerische Technische Hochschule (OTH) Regensburg, Faculty of Mechanical Engineering, Laboratory Composite Technology, Regensburg, 2019

[15] SCHLIMBACH, J.; NEITZL, M.: *Der industrielle Einsatz von Faser-Kunststoff-Verbunden.* Carl Hanser Verlag, Munich, 2004

[16] SCHUERMANN, H.: *Konstruieren mit Faser-Kunststoff-Verbunden.* Edition 2, Springer-Verlag, Berlin/Heidelberg 2007

[17] WITTEN, E.: *Handbuch Faserverbundkunststoffe/Composites.* Springer Vieweg, Wiesbaden, 2014

Evaluation of a Novel Design of Venous Valve Prostheses via Computational Fluid Simulation

Felix Klinger, Lars Krenkel
Department of Biofluidmechanics
University of Applied Sciences
Regensburg, Germany
Email: felix.klinger@oth-regensburg.de

***Abstract*— Chronic venous insufficiency (CVI) of the lower legs is a common medical problem and occurs widely in the general population in the Western world. The implantation of artificial valves is a promising approach for the treatment of venous incompetence. Geometry of an artificial valve was inserted into a cylindrical vein segment in a CAD environment and flow characteristics in the immediate surroundings of the valves were investigated using Computational Fluid Dynamics. Flow characteristics were acquired for both forward and retrograde flow directions. Static pressure drop from before and after the valve increased by 9.34 % when flow was in in blocking direction (retrograde flow) when compared to passing direction (forward flow). For two valves, placed in series in a relative distance of 50 mm to each other, static pressure drop did not increase significantly compared to single valve setup. Shear rates along the valve surface was below 3,500 s^{-1}, suggesting there be no hemolytic effects on blood cells. Strain rates obtained from bidirectional flow analysis were in a range of 1000 - 1500 s^{-1}within the fluid domain. Maximum strain rates of ~ 4000 s^{-1} and 19,000 s^{-1} were observed at the valves‘ inner surface for brief time periods. Strain rates within the fluid domain and at the valves surfaces were not subject to change for dual valve setup.**

Keywords - CVI, Artificial Venous Valves, Computational Fluid Dynamics, Venous Incompetence, Prostheses

I. Introduction

A. Motivation

Venous disease has been a scourge of humanity for a very long time, with reports dating back more than 2000 years [1]. Today, chronic venous insufficiency (CVI) of the lower legs is a common medical problem and occurs widely in the general population in the western world [2]. In CVI, the return of venous blood to the heart is impaired. As a result, blood pools in the veins, straining the walls of the vein. A major issue in CVI is superficial venous retrograde flow, which is caused by gravitational forces in the upright position. Together with other biological adaptions, venous valves play a crucial role in preventing pathologic retrograde flow. Venous valves are densely distributed in the region of the lower leg and reduce the full pressure of the fluid column on the distal veins by dividing the hydrostatic column of blood into segments. Consequently, incompetent valves can lead to a variety of complications within the venous system, such as superficial varicosities, stasis dermatitis and venous ulcers [3].

B. Aims & Objectives

The implantation of artificial valves is a promising approach for the treatment of venous incompetence. For the construction and design process of the valve, it is important to understand the flow characteristics in the immediate surroundings of the valves. Computational Fluid Dynamics (CFD) is a powerful tool for getting a detailed overview of complex flow phenomena. It can also deal as a reference for experimental testing. In this project, a leafless artificial venous valve was designed (see figure 1) and its near field flow characteristics were examined with CFD. Meshes of the geometry were obtained using *ICEM CFD 19.2* software and computational fluid simulation was carried out with *ANSYS Fluent v 19.1* (both ANSYS, Inc.). The main objective was to simulate blood flow in various artificial valve setups and evaluate pathological retrograde flow. Intuitively, one would suggest that volumetric flow at the outlet is the key state variable in determining valve performance. However, taking this approach is not feasible due to the *ANSYS Fluent* framework, which is based on continuity considerations. Instead, pressure characteristics along the artificial venous valves were quantified. Pressure is reduced with an increased resistance along the fluid's flow path. Hence, differentials obtained from the valves' boundaries are used to evaluate valve performance. Flow resistance should be high when the direction of flow is towards the large valve orifice

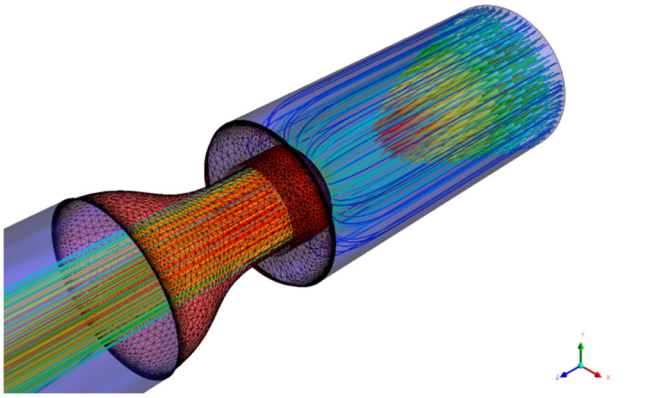


Figure 1: Fluid domain of an artificial venous valve placed within a vein segment with streamlines and parabolic velocity profile at the inlet of the segment.

(blocking scenario) while it should be low for reversed flow direction to allow uninhibited flow. Analyses were undertaken for both scenarios in a single and dual valve setup and pressure results were compared.

Yet another objective was the investigation of peak flow velocity magnitudes, vortex creation and shear stress on the fluid in a transient bidirectional flow scenario. High wall shear rates have been shown to have a negative effect on human blood cells, by increasing platelet activation and binding [4]. Shear induced thrombosis is avoided with a physiologic shear rate below $3{,}500\ s^{-1}$ [5].

II. Methology

A general assessment of velocity magnitudes prior to simulation is an important factor for a reliable fluid simulation. Among other factors, velocity magnitude directly affects the underlying Reynolds number, which is crucial for the selection of the simulation model. Common velocities in the human saphenous vein are between $0.08 - 0.26\ ms^{-1}$ [6, 7]. The velocity magnitude at the inlet is the only variable directly affecting the Reynolds number, as all of the other parameters are assumed to remain constant. The Reynolds number was calculated using

$$Re = \frac{\rho * v * d}{\mu}$$

with

$\rho = fluid\ density = 1{,}060\ kg\ m^{-3}$
$v_{in} = inlet\ vel.magnitude = 0.26\ ms^{-1}$
$d = vein\ diameter = 0.01\ m$
$\mu = dyn.viscosity = \ 0.0027 Ns\ m^{-2}$

The density of blood was taken to be $1{,}060\ kg\ m^{-3}$ and the blood viscosity to be $0.0027\ Ns\ m^{-2}$ [8]. For a flow velocity magnitude of $v_{in} = 0.26\ ms^{-1}$ the Reynolds number was $Re = 1{,}020$ for the overall flow domain. Within the valve orifice region the velocity magnitude was expected to increase to a value up to $v_{max} = 0.7\ ms^{-1}$. Together with a geometry diameter of $d = 0.005\ m$, the Reynolds number increased to $Re = 1374$.

The average diameter of an incompetent Greater Saphenous Vein (GSV) has been reported to be wider than that of competent saphenous veins, measuring 4.4 mm to 14.8 mm [9–11]. CAD software (*Creo Parametrics 6*) was used to generate a cylindrical geometry with a diameter of 10 mm, representing a human GSV. Geometry was modeled and simulated without symmetry assumptions. The CAD model of the venous valve was inserted concentrically into and then substracted from the GSV segment with a Boolean operation to obtain the fluid domain for mesh generation. The fluid domain extended 25 mm before and 100 mm after the venous valve. The mesh contained tetrahedral and hexahedral elements with a global maximum size of 0.5 mm. 8 Prism layers were added to the surfaces of the venous valve and to the inner layer of the venous segment, with a first layer height of 0.02 mm and a height ratio of 1.2. Curvature – Proximity based refinement of delicate features was applied with a minimum size of 0.01 mm. For both, the passing scenario and the blocking scenario, an individual mesh was created. Cell sizes before and after the valve were set equally for both meshes. The unidirectional meshes contained 626,350 nodes and 1,946,393 solid elements. For bidirectional simulations, the inlet length was extended to a value of 100 mm. Mesh element sizes and prism parameter settings were set up as in the unidirectional case. The bidirectional mesh contained 835,503 nodes and 2,751,128 solid elements. Double valve setup increased the number of elements to 1,086, 302 nodes and 3,459,248 solid elements for the unidirectional case and 1,322,895 nodes and 4,379,201 solid elements for the bidirectional case.

For internal flow domains, a flow is considered laminar for $Re < 2300$. Thus, for the simulation a laminar viscous model was chosen. Ambient inlet and outlet regions were defined at the extremes of the fluid domain with zero pressure applied to the outlet and a parabolic velocity profile defined at the inlet of the domain. Fluid elements were defined using material characteristics of human blood, as discussed above. Analyses were undertaken for passing scenario and blocking scenario using the same boundary conditions respectively. Static pressure was obtained over the whole fluid domain and compared for two points (P1 and P2) in the center plane of the fluid domain, with P1 being 15 mm proximal and P2 being 15mm distal to the valve's center point. In dual valve setup (P2) was set to be 15 mm distal to the second valve's center point downstream. In the bidirectional analysis, strain rates were obtained for the whole fluid domain. Analyses were run in transient solver setup for a time of 0.7 seconds to allow an equilibrium to develop, with a solution timestep of 0.001 seconds for unidirectional and 0.005 seconds for bidirectional flow. Results of unidirectional flow were saved for each time step and data quantities were compared at 0.7 seconds simulation time. For bidirectional analysis, a sine-function was applied to the parabolic velocity profile at the inlet with an amplitude of $0.26\ m^{-1}$. Cycle duration was set to 2 seconds, accounting for pathologic reflux conditions > 0.5 seconds. 800 steps were calculated, resulting in a flow simulation time of 4 seconds.

An animation of the 800 time steps was created, so that the strain rate over the whole cycle could be easily evaluated. All simulations were run on 4 cores of an Intel® Core™ i7-6700 processor at 3.40 GHz, with typical run times of the order 20 hours per analysis. Pressure changes and velocity magnitudes through the valve region were assessed by exporting nodal velocity, shear stress and pressure results on the center plane of the fluid domain. For the dual valve setup, valves were placed in a relative distance of 50 mm from each other, representing physiologic conditions in the GSV [12].

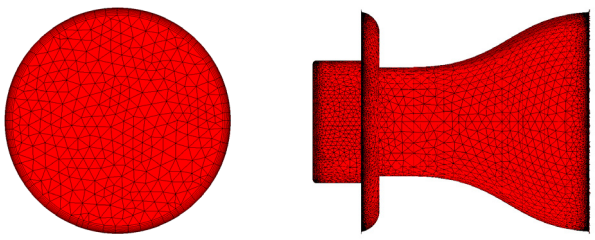


Figure 2: Left: Tetra surface mesh of the inlet with prism layers on the wall regions. Right: fluid domain within the artificial venous valve.

III. Results

A. Pressure Differentials for Unidirectional Flow Scenario

In the single valve setup, static pressure decreased from 180.22 Pa before (P1) to -43.37 Pa (P2) behind the valve for blocking scenario. Pressure values for passing direction were 165.56 Pa (P1) and -38.72 Pa (P2) (see figure 3). In the dual valve setup, overall static pressure was higher, with values reaching 275.32 Pa (P1) in blocking scenario and 249.64 Pa (P2) in passing scenario. Pressure behind the valves were -40.31 Pa (blocking) and -52.93 Pa (passing).

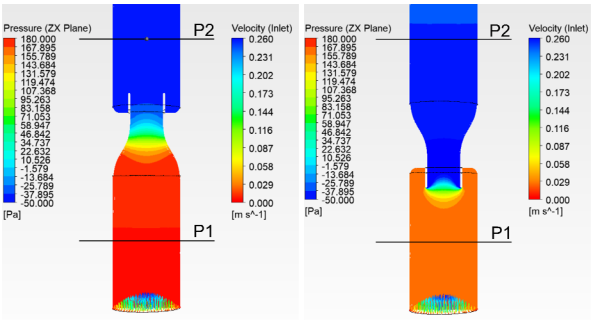


Figure 3: Static pressure obtained for unidirectional flow single valve setup. The parabolic velocity profile indicates flow direction. Left: Blocking scenario. Right: Passing scenario.

B. Strain Rates for Bidirectional Flow Scenario

Strain rates within the fluid domain were in a range of 1,000 – 1,500 $s^{-1}$for the whole bidirectional flow sequence. Maximum strain rates of ~ 4,000 s^{-1}occured at the valves‘ inner surface in the mid-section in blocking direction, with values reaching 19,762 s^{-1} at the orifices in passing direction (see figure 4). These high strain rates were observed only for brief periods of time, namely at the timesteps where there is maximum velocity magnitude at the inlet. Strain rates did not change significantly in dual valve setup compared to single valve setup.

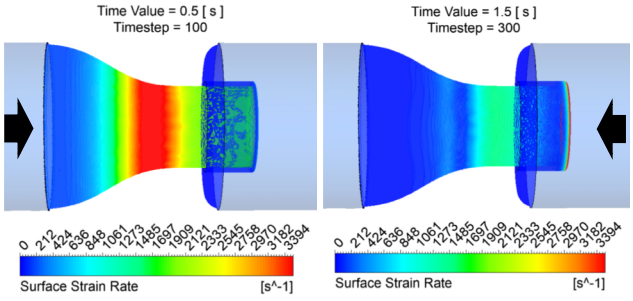


Figure 4: Strain rates at the valve’s inner surface for blocking scenario (left) and passing scenario (right). Black arrows indicate flow directions.

IV. Discussion & Outlook

For the single valve setup, static pressure increased by 9.34 % when the inlet flow was pointing in blocking direction, compared to the passing scenario, suggesting an increased resistance to the flow induced by the artificial venous valve. Flow resistance is expected to increase even more in a dual valve setup. However, the results show an increased flow resistance of only 4.32 % in dual valve blocking scenario, compared to dual valve passing scenario. This might be due to boundary layer separation of the flow, which was observed in the fluid domain proximal to the second valve, having a potential effect on the static pressure at position P2. Therefore, the significance of pressure values obtained in the dual valve setup might be limited. Relative distance between the two valves could have an effect on the generation of such flow separation phenomena. An intensive parameter study will be conducted in order to find the optimal distance between the valves.

Simplifications had to be made in the experimental design of this study. This applies mostly to the physiologic depiction of venous compliance. In the human body, pressure differentials have a significant effect on venous vessel diameter, directly affecting flow characteristics. In this study, no solid-fluid interaction was modeled, thus venous compliance could not be taken into account. However, this should affect the depiction of strain rates at the valve’s inner surfaces. The influence of venous compliance on pressure characteristics is obvious, but not relevant for the findings of the study, because blocking and passing scenario were directly compared within the same setup.

In bidirectional flow scenario, the direction of flow through the veins was assumed to change periodically in the shape of a sine-curve, which is also a simplification of physiological conditions. Apart from physiologic characteristics there were limitations regarding the resolution of the mesh as well as the fluid domains’ outlet length. Mesh resolution was held as high as possible and outlet length as long as possible, without increasing element count too much. As both aforementioned factors are known to have an impact on numerical computations, the solution is subject to slight variations. An extensive mesh parameter study will be conducted in the future, to decrease those uncertainties. Additionally, the relative distance of the valves in the double valve setup will be varied in yet another parameter study and the possible effects on velocity magnitudes and shear rates will be investigated. Furthermore, results from CFD will be used as a baseline for experimental testing.

References

References

[1] C. Kügler, *Venenkrankheiten*. Berlin: ABW Wissenschaftsverlagsgesellschaft mbH, 2011.

[2] L. Robertson, C. Evans, and F. G. R. Fowkes, “Epidemiology of chronic venous disease,” *Phlebology*, vol. 23, no. 3, pp. 103–111, 2008, doi: 10.1258/phleb.2007.007061.

[3] V. Baliyan, S. Tajmir, S. S. Hedgire, S. Ganguli, and A. M. Prabhakar, “Lower extremity venous reflux,” *Cardiovascular Diagnosis and Therapy*, vol. 6, no. 6, pp. 533–543, 2016, doi: 10.21037/cdt.2016.11.14.

[4] B. Savage, E. Saldívar, and Z. M. Ruggeri, "Initiation of Platelet Adhesion by Arrest Onto Fibrinogen or Translocation on Von Willebrand Factor," *Cell*, vol. 84, no. 2, 1996, doi: 10.1016/s0092-8674(00)80983-6.

[5] D. E. Tanner, "Design, analysis, testing, and evaluation of a prosthetic venous valve: Design, analysis, testing, and evaluation of a prosthetic venous valveUR - https://smartech.gatech.edu/handle/1853/51758," Georgia Institute of Technology. [Online]. Available: https://smartech.gatech.edu/bitstream/1853/51758/1/tanner_daniel_e_201305_mast.pdf, date of access: 11.06.2020

[6] W. A. Marston, V. W. Brabham, R. Mendes, D. Berndt, M. Weiner, and B. Keagy, "The importance of deep venous reflux velocity as a determinant of outcome in patients with combined superficial and deep venous reflux treated with endovenous saphenous ablation," *Journal of Vascular Surgery*, vol. 48, no. 2, pp. 400–406, 2008, doi: 10.1016/j.jvs.2008.03.039.

[7] P. Abraham, G. Leftheriotis, B. Desvaux, M. Saumet, and L. Saumet, "Diameter and blood velocity changes in the saphenous vein during thermal stress,", *Eur J Appl Physiol*, vol. 69, no. 4, pp. 305–308, 1994, doi: 10.1007/BF00392035.

[8] A. Garvin, B. and N. Clarke, "Computational Phlebology: The Simulation of a Vein Valve," *undefined*, 2006. [Online]. Available: https://doi.org/10.1007/s10867-007-9033-4, date of access: 11.06.2020

[9] C. and A. Engelhorn, Salles-Cunha, Sergio, F. Picheth, C. Gomes, "Relationship Between Reflux and Greater Saphenous Vein Diameter," *Journal of Vascular Technology*, 1997.

[10] E. Mendoza, W. Blättler, and F. Amsler, "Great Saphenous Vein Diameter at the Saphenofemoral Junction and Proximal Thigh as Parameters of Venous Disease Class," *European Journal of Vascular and Endovascular Surgery*, vol. 45, no. 1, pp. 76–83, 2013, doi: 10.1016/j.ejvs.2012.10.014.

[11] J. H. Joh and H.-C. Park, "The cutoff value of saphenous vein diameter to predict reflux," *Journal of the Korean Surgical Society*, vol. 85, no. 4, pp. 169–174, 2013, doi: 10.4174/jkss.2013.85.4.169.

[12] G. Schweighofer, D. Mühlberger, and E. Brenner, "The anatomy of the small saphenous vein: Fascial and neural relations, saphenofemoral junction, and valves," *Journal of Vascular Surgery*, vol. 51, no. 4, pp. 982–989, 2010, doi: 10.1016/j.jvs.2009.08.094.

Evaluation of Surface Plasmonic Effects in Glass Fibers

1st Sophie Emperhoff
Applied Natural Sciences and Cultural Studies
Sensorik-ApplikationsZentrum
Regensburg, Germany
sophie.emperhoff@st.oth-regensburg.de

2nd Johannes Fischer
Applied Natural Sciences and Cultural Studies
Sensorik-ApplikationsZentrum
Regensburg, Germany
johannes.fischer@oth-regensburg.de

***Abstract*—Surface Plasmon Resonance has gained a great deal of attraction and has faced many advancements in sensing physical, chemical and biochemical parameters. In particular, optical fibers offer advantages in size and flexibility of use. In this paper, we evaluate the use of a cylindrical glass stick as an optical waveguide. This waveguide is coated with metal layers to achieve a surface plasmonic effect. The presence of surface plasmons is proven by a simulation and an experimental validation. The simulation has shown that the excitation of surface plasmons in waveguides has a stronger impact on the transmitted intensity than a shift of the critical angle. The presented setup in the experiment already achieves remarkable resolutions.**

***Index Terms*—surface plasmon resonance, fiber, optical sensor, biosensor**

I. Introduction

Surface Plasmon Resonance (SPR) is a highly sensitive and label-free optical method which has become a tool for characterizing and quantifying biomolecular interactions as well as for detecting very small concentrations of molecules. There are several different setups to excite those surface plasmons in a metal-dielectric interface. Using optical fibers as a way of excitation is gaining more and more importance since the use of fibers offers numerous advantages such as their small size and robustness [1].
In this paper a simple glass stick is used as an optical fiber. It is evaluated whether it is possible to measure SPR effects in a setup like this. This is done in two ways. First, through a simulation of the occuring effects and second, through an experimental validation. It is also investigated what resolution can be reached with this setup.
This paper begins with a short explanation of the principle of surface plasmon resonance in optical fibers. After this, the method used in the simulation and in the experiment is demonstrated. Following, results of both parts are presented and finally the results are discussed.

II. Principle

Surface plasmon oscillations are charge density oscillations located at a metal-dielectric interface. These oscillations can be excited through total internal reflection of p-polarized light. The oscillation causes an electric field that decays exponentially to both sides of the interface which is also referred to as an evanescent wave. The electric field has its maximum at the metal-dielectric interface which makes surface plasmon resonance a highly surface sensitive principle [2].

A very common configuration to excite surface plasmons

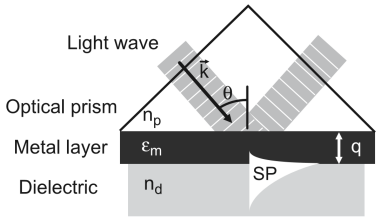


Fig. 1. Attenuated total reflection through excitation of surface plasmons with the Kretschmann configuration [3].

is the Kretschmann configuration as shown in figure 1. This configuration uses a prism coupler and the attenuated total reflection method (ATR). The prism has a high refractive index n_p and is interfaced with a sufficiently thin metal layer of thickness q and a permittivity ϵ_m. A light wave is propagating through the prism and when impinging on the metal layer one part of the light is reflected back into the prism and the other part is propagating as an evanescent wave. The evanescent wave couples with the surface plasmons on the metal dielectric interface. The propagation constant of surface plasmons is influenced by the dielectric constant of the adjacent medium and therefore the excitation of surface plasmons is dependent on the refractive index of the dielectric n_d. Whether the entering light ray is exciting surface plasmons is also dependent on its wavelength and angle of incidence [3].

Surface plasmons can also be excited using optical fibers. Figure 2 shows a typical setup for fiber-based SPR sensing. The fiber consisting of a core and the surrounding cladding guides light rays through total internal reflection. The cladding is partly removed from the fiber and the bare core is coated with a metal layer (usually gold or silver). As the rays pass through the fiber, light impinges on those sensing areas and excite surface plasmon waves (SPW) at the metal-dielectric interface [4].
A change of the refractive index of the dielectric can be

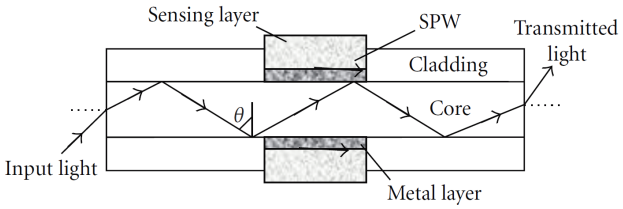


Fig. 2. A possible probe for fiber SPR sensing [2].

detected by a dip in the spectrum of the transmitted light or as a change in intensity of the transmitted light [5].

III. Method

In this paper we use a simple glass stick as a light guiding structure. The refractive index of the glass is higher than the refractive index of air and water. Therefore, total internal reflection takes place within the glass stick even though there is no cladding around it. The stick is coated with a 4 nm chromium layer and a 44 nm gold layer, has a length of 15 cm and a diameter of 1 mm.
The evaluation of surface plasmonic effects is separated into two sections. First, a simulation is conducted to compare coated with uncoated glass guiding structures in order to see if an intensity change can be attributed to SPR effects or to a shift of the critical angle of the total internal reflection. Secondly, the results of the simulation are validated through an experimental approach.

A. Simulation

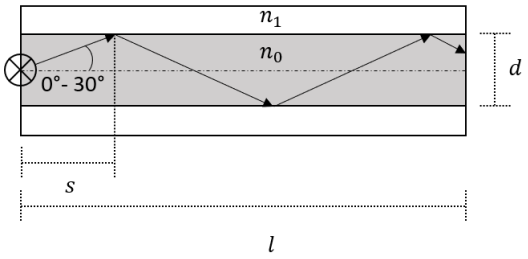


Fig. 3. Schematic drawing of a light guiding structure ($n_1 < n_0$) with thickness d and length l. A light source on the left side radiates light into the structure with angles between $0° - 30°$.

We are very familiar with the reflectivity of different SPR systems thanks to software tools like "Winspall" that help simulating their behaviour in a Kretschmann configuration. For the new approach using a system with several reflections the following shows an approach to estimate the resulting reflectivity after passing through a planar wave guide system. We can simulate the result for different systems but instead of just one reflection the following simulation calculates the reflectivity with multiple reflections dependent on the length and the diameter of the structure. Figure 3 shows a schematic drawing of the examined system. Light enters on the left side and radiates through the glass with a refractive index of $n_0 = 1.472$. The refractive index n_1 is dependent on the medium that surrounds the glass.
We assume the entrance cone of a light source entering the light guiding structure has an angle of $60°$. The simulation takes half of the entrance cone ($0° - 30°$). The center of the light source is situated right at the border of the structure. The model does not include any coupling into the glass. First, the single reflectivity of a SPR system is calculated for this range using a SPR system. The model uses both s- and p-polarized light. The number of reflections is calculated using a trigonomic approach. If we know the length of the guiding structure, we know how many times the ray reflects at the glass-medium interface until it exits the light guiding structure.

$$s = \frac{d/2}{\sin(\alpha \cdot \frac{\pi}{180})} \tag{1}$$

$$N = Integer\left(\frac{l-s}{(2 \cdot s)} + 1\right) \tag{2}$$

With equations 1 and 2 we can calculate the number of reflections N a ray with the angle α needs to pass through a guiding structure of a thickness d and a length l. The distance a ray travels until it reaches the glass-medium interface is called s. The simulation defines two exceptions. If α is zero the number of reflections is zero and the total reflectivity is set to 1. If α is $90°$ the number of reflections is also set to zero and the total reflectivity is set to zero. For all other cases the total reflectivity R_{total} is calculated with:

$$R_{total} = R^N \tag{3}$$

The simulation uses a thickness of 1 mm, a length of 15 cm and a wavelength of 660 nm. The simulation is conducted for two different systems. One system is the combination of several layers that build the sensor surface. System GCGD is a glass-chromium-gold-dielectric system. Thicknesses for chromium and gold are assumed to be 4 nm and 44 nm respectively. Whereas system GD consists of a glass-dielectric interface. That means we are comparing a system which is able to excite surface plasmon waves with a system in which the amount of transmitted light depends mostly on the refractive index of adjacent medium and therefore the shift of the critical angle. Table I shows the refractive indices and their respective thickness used in the simulation.

TABLE I
Refractive indices and layer thicknesses used in this simulation

Material	Refractive Index	Layer Thickness
Glass	1.472	-
Chromium	3.48 + 4.36i	4 nm
Gold	0.13 + 3.16i	44 nm
Water	1.33327	-

B. Experimental Validation

In order to validate the results of the simulation, a glass stick with a length of 15 cm and a diameter of 1 mm is used. It is coated with chromium and gold and represents system GCGD. A red LED with a wavelength of 660 nm is used for the experiment. The LED is held directly against the glass stick. The transmitted light hits a photodiode detector right after it exits the glass. A circular tank with a diameter of 5 cm is used as the sample container which means 5 cm of the coated area are used as the sensor surface. The whole setup is standing in a light shielding box to prevent outside light from influencing the results. A tube leads from outside the box into the tank. The tube is used to add analyte medium to the tank.

A glucose solution with a concentration of $1\ \frac{\text{mol}}{\text{l}}$ is used as the analyte. At the beginning of the experiment 15 ml deionised water is added to the tank. Every 120 s 0.5 ml are added to the analyte through the tube. This is repeated 10 times. The transmitted intensity is measured with a photodiode and recorded in ADC digits.

The resolution Δn of the measured refractive index n is calculated as follows:

$$\Delta n = \frac{3 \cdot \Delta I_{max}}{S_n} \tag{4}$$

with ΔI_{max} being the maximum standard deviation of the measured mean intensity at different concentration steps and $S_n = \frac{dI}{dn}$ being the calculated sensitivity of the sensor.

IV. Results

A. Simulation

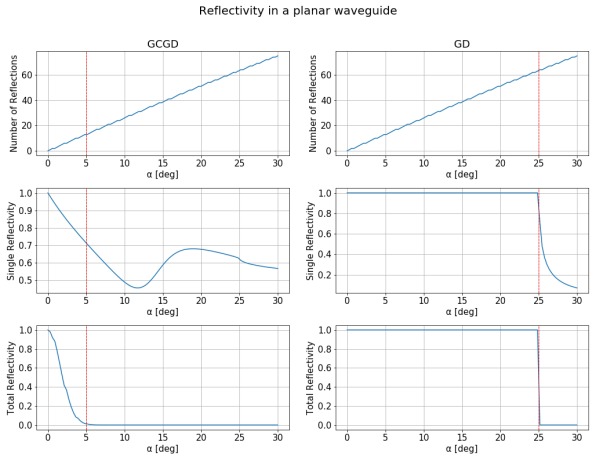


Fig. 4. Simulation of reflections within a planar waveguiding structure of a certain thickness and length: the top row shows the number of reflections of light rays within the waveguide depending on their angle to the optical axis; the middle row shows the reflectivity profile for a single reflection; the last row shows the resulting reflectivity after multiple reflections

This simulation compares the behaviour of reflectivity for two different layer systems. One of them is capable of exciting surface plasmons and consists of the layers glass, gold, chromium and a dielectric (GCGD). The other one consisting of only glass and a dielectric does not have a metal layer which means there are no surface plasmons that can be excited (GD). The aim of the simulation is to check whether there is a significant difference between a SPR system and a non SPR system.

In a first step, a general behaviour of those two systems is examined for $n_1 = 1.33327$. This includes the number of reflections, the reflectivity for just one reflection and the total intensity at the end of the waveguide after several reflections. The first row of figure 4 shows the number of reflections dependent on the angle to the optical axis as indicated in figure 3. The larger this angle gets the more reflections are needed for the ray to pass through the waveguide.

In the second row we can see the reflectivity after just one reflection on the interface of the corresponding system. System GCGD shows a so-called SPR dip at $\approx 12°$. Left of this dip we can see nearly linear behaviour of the reflectivity. The critical angle for system GD is $25°$. For angles larger than this reflectivity decreases rapidly.

The third row shows the total reflectivity according to equation 3. For system GCGD total reflectivity drops to nearly zero for angles larger than $5°$. Whereas, for system GD total reflectivity looks similar to a single reflectivity except that there is a sharp drop to zero at $25°$. The red line highlights this turning point.

Next, this simulation is conducted for four different refractive

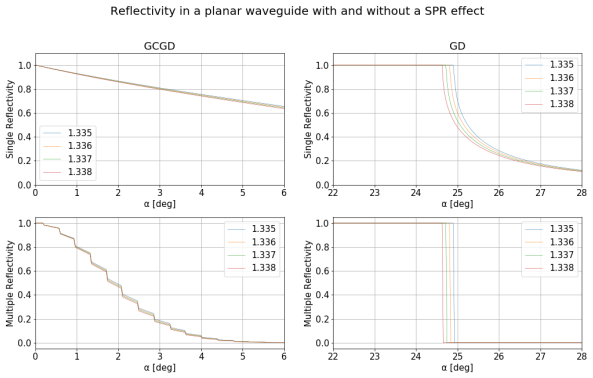


Fig. 5. Comparison of the resulting reflectivity with SPR effect and without the effect.

indices n_1 of the adjacent dielectric. Figure 5 zooms into areas with a high impact on the resulting reflectivity. The first row again shows a single reflection and the second row shows the total reflectivity. We can see a shift to the left for increasing refractive indices in both systems.

Finally, an integral is taken of each reflectivity graph. Figure 6 shows the relative integral dependent on the refractive index of the dielectric. Over the range from $n_1 = 1.335$ to $n_1 = 1.338$ the area underneath the reflectivity for system GCGD shows a change of more than 3%. The same range causes a change of 1% for system GD.

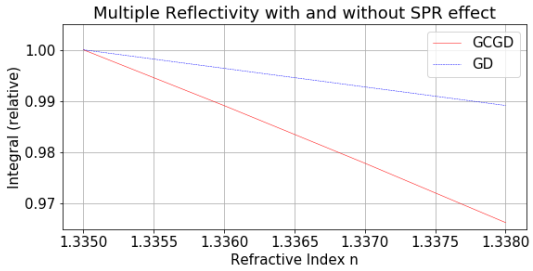


Fig. 6. Integral over multiple reflectivity depending on the refractive index of the analyte for two different interfaces.

B. Experimental Validation

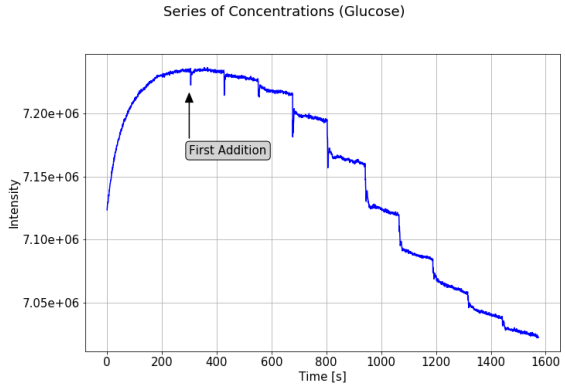


Fig. 7. Intensity measured over time. The first addition of medium to DI water is highlighted. Each further addition can be seen as a drop in intensity.

This experiment validates whether surface plasmonic changes of the transmitted intensity are measurable in a glass stick used as a simple waveguide. Figure 7 shows the measured intensity in ADC digits after conducting the experiment as described in III-B. We can clearly see each step in the concentration series as a drop in intensity. We don't see equidistant changes in intensity but the concentration is not changed linearly as well. Therefore, the refractive indices of each step is measured with a refractometer. The results are shown in table II. The refractive index of a medium is a dimensionless number used to describe how fast light travels through a material. For small changes of the refractive index or quantities that are normalized by the refractive index, Refractive Index Units(RIU) are used. We narrow the results down to the range which was used in IV-A. The mean intensity is calculated over 100 values for each step. A linear regression is calculated for the intensity values over the range from 1.3350 to 1.3385. This leads to a sensitivity of $-5.54 \cdot 10^{+7}$ RIU^{-1} for this setup.

TABLE II
REFRACTIVE INDICES OF CONCENTRATION STEPS

Step	Refractive Index
0	1.33296
1	1.33330
2	1.33444
3	1.33513
4	1.33580
5	1.33642
6	1.33691
7	1.33754
8	1.33795
9	1.33844
10	1.33879

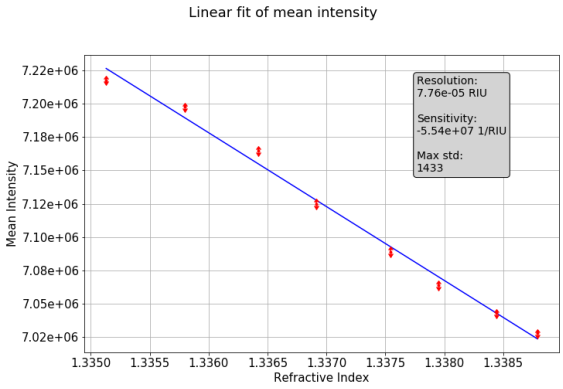


Fig. 8. Linear fit over the range from 1.3350 to 1.3385. This leads to a resolution of $7.76 \cdot 10^{-5}$ RIU and a sensitivity of $-5.54 \cdot 10^{+7}$ RIU^{-1}. The maximum standard deviation is 1433 ADC digits.

With the maximum standard deviation of 1433 ADC digits we receive a resolution of $7.76 \cdot 10^{-5}$ RIU.

V. Discussion

Based on the simulation we can draw several conclusions:

1) The total intensity we measure at the end of the waveguide is larger in system GD than in system GCGD.
2) The total intensity we measure differs between different adjacent refractive indices for both systems.
3) For the range from $n_1 = 1.335$ to $n_1 = 1.338$ system GD shows a change in intensity of about 1% whereas system GCGD shows a change of 3%. That means the relative change of intensity is larger for systems with surface plasmonic effects.
4) Certain angles show a larger sensitivity than others. Therefore, we can reach higher sensitivities if we allow only a certain range of angles in the waveguide.

The experiment clearly shows that changes of the refractive index of the analyte can be observed with the introduced setup. This experiment reaches a resolution of $7.76 \cdot 10^{-5}$ RIU. To further enhance this resolution several possibilities open up. As we can see from the simulation small angles result in a large amount of transmitted light. If we exclude small angles from entering the waveguide, the relative change of intensity

will increase. This can, for example, be achieved through tilting the LED when coupling into the waveguide. Large angles result in a large number of reflections. Each reflection increases the SPR effect. The number of reflections can also be increased through a larger sensor surface. The glass stick in this experiment has been coated with chromium and gold on two opposing sides of the stick. The sensor surface can be significantly increased by evenly coating the whole cylinder surface.

The examined setup has proven capable of measuring intensity changes caused by surface plasmon resonance and already has shown good results. The resolution can be improved through further measures.

REFERENCES

[1] Radan Slavík, Jirí Homola, Jirí Ctyroký, "Single-mode optical fiber surface plasmon resonance sensor" in Sensors and Actuators B, vol. 54, 1999, pp.74-79.

[2] B. D. Gupta, R. K. Verma, "Surface Plasmon Resonance-Based Fiber Optic Sensors: Principle, Probe Designs, and Some Applications" in Hindawi Publishing Corporation, Journal of Sensors, vol. 2009.

[3] J. Homola and O.S. Wolfbeis, "Surface Plasmon Resonance Based Sensors", Ser. Springer series on chemical sensors and biosensors, Berlin: Springer, 2006, vol. 4.

[4] A. Leung, K. Rijal, P. M. Shankar, and R. Mutharasan, "Effects of geometry on trans-mission and sensing potential of tapered fiber sensors",Biosensors & bioelectronics, vol. 21, no. 12, pp. 2202–2209, 2006, DOI:10.1016/j.bios.2005.11.022.

[5] B. Lee, S. Roh, and J. Park, "Current status of micro- and nano-structured opticalfiber sensors", Optical Fiber Technology, vol. 15, no. 3, pp. 209–221, 2009, DOI:10.1016/j.yofte.2009.02.006.

SESSION B1

Absorber free laser transmission welding of COC with pyrometer-based process monitoring

Stephan Englmaier*1, Frederick Maiwald*1, Stefan Hierl*1

*1 Ostbayerische Technische Hochschule Regensburg
Labor Lasermaterialbearbeitung
Technologie Campus Parsberg-Lupburg
Am Campus 1, 92331 Parsberg

stephan1.englmaier@st.oth-regensburg.de

***Abstract*—Optical and medical devices made of transparent polymers like cyclic-olefin-copolymers have high demands in precision, visual appearance and reliability. Laser transmission welding is an interesting joining technique for these devices, with advantages in contactless energy input, high precision and no particle formation. For absorber free laser transmission welding, laser sources emitting in the polymers' intrinsic absorption spectrum between 1.6 µm and 2 µm are used to obtain absorption without additives. Additionally, the laser beam is focused inside the specimen using high numerical aperture, enabling selective fusing of the joining zone. However, to meet the strict quality requirements in optical and medical industries, the process lacks stability. Thus, online process monitoring is needed. Aim of this work is the identification of process parameters resulting in a good weld, as well as development and application of pyrometer-based process monitoring. In a first step, a simulation model is set up to forecast parameters for welding. Then, using a fixed focus setup, these parameters are verified. Last, a pyrometer for online temperature measurement is integrated. Comparing the pyrometer signal to thin-cuts from processed parts shows, that it is possible to localize the weld seam inside the material using a pyrometer and identify process irregularities.**

Keywords—laser transmission welding, pyrometry, absorber free, process monitoring, cyclic olefin copolymer

I. Introduction

Laser transmission welding is an established technique for joining polymers, especially in automotive industries. The two joining partners are placed in an overlap and fixed with a clamping device. The laser beam is guided at least one time along the weld trajectory.

Since medical or optical devices have high demands in visual appearance and cleanliness, usage of additives to ensure laser absorption is prohibited. Moreover, additives can cause problems with the admission procedure for medical devices, particularly with regard to physiological tolerances. To obtain laser absorption without additives, laser sources emitting in the polymers' intrinsic absorption spectrum between 1.6 µm and 2 µm, are used. Additionally, focusing with high numerical aperture ensures selective fusing of the desired joining zone. Thus, it is possible to process hermetically sealed seams with just the size of some tenth of millimeters without affecting surface damage or warpage [1, 2]. Fig. 1 shows the process principle of absorber free laser transmission welding.

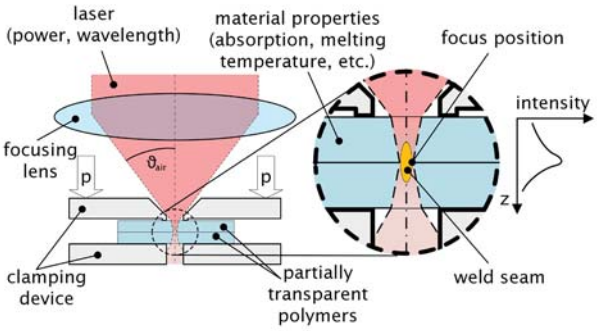


Figure 1: Processing principle of absorber free laser transmission welding using high NA optics for focusing.

Although the basic usability of absorber free laser transmission welding is already proven, the process is still unstable, resulting in loose joints or damaged surface. In order to fulfill the extreme demands on quality and reliability for medical and optical industries, a stable process with online process monitoring is needed. In conventional transparent-absorbent laser transmission welding of polymers, pyrometry is already established for online temperature measurement, enabling gap detection and process monitoring during welding [3-5].

However, in absorber free laser transmission welding conditions are more challenging for pyrometry. The weld seams and thus the heat emitting volume is smaller. Additionally, the processing laser wavelength is inside the spectral detectivity range of the pyrometer.

Goal of this work is to identify and verify welding parameters leading to a tight joint without surface defects and analyze the correlation between weld result and pyrometer signal. First, a simulation model based on finite elements method is built to find promising parameters for welding success. These parameters are tested using a fixed focus test stand for welding COC Topas 8007-04 and polyamide 6 Ultramid B3s. Finally, a pyrometer is integrated to prove its capability to detect the weld seam position inside specimen material.

II. Simulation based process layout

To predict the impact of process and material parameters on the welding process, the temperature distribution and seam formation is computed using finite element method.

A. Thermal Simulation

Fourier's differential equation describes the temperature field of a polymer heated by radiation (1). The variables are density ρ, specific heat capacity c and thermal conductivity k. Lambert-Beer's law of absorption describes the amount of power deposited in the polymer, characterized by absorption coefficient α. Equation (2) describes the power supply $\dot{Q}$ for feed rate v of a Gaussian beam with Power P and beam diameter d.

$$\rho \cdot c \cdot \frac{\partial T}{\partial t} - \nabla \cdot (k \cdot \Delta T) = \dot{Q} \quad (1)$$

$$\dot{Q}(x,y,z,t) = \frac{8 \cdot \alpha \cdot P}{\pi \cdot d^2} \cdot e^{-\alpha \cdot z - 8 \cdot \frac{(-x_0 - v \cdot t)^2 + y^2}{d^2}} \quad (2)$$

For process simulation, a two-dimensional, thermal finite elements analysis is set up, since the temperature gradient in feed direction is insignificant because of the polymer's low thermal conductivity and the high feed rates of the laser. Thermal conductivity and heat capacity are implemented in dependence on temperature. In absorber-free laser transmission welding, deformations and melt blowout are usually avoided to prevent a blockage of the microfluidic system. Thus, only a thermal simulation neglecting deformations is sufficient. Heat transfer to the ambient air is negligible as well [6, 7]. After computing, the temperature distribution is compared with the materials melting temperature and all regions exceeding this temperature are defined as "seam". To obtain a good weld, the seam must cover both specimen with restriction to no molten surface.

In industrial applications, two irradiation regimes are mainly used [8]: Contour welding describes a process, where the laser is moved one time along the desired welding contour, locally heating and melting the material. This local heating prevents melt blowouts, making it advantageous for welding of microfluidic devices, since flow channels must not be affected. A major disadvantage is a low gap-bridging capability and high thermal gradients, leading to residual stresses. Quasi simultaneous welding provides a more homogenous temperature distribution with longer interaction times, leading to reduced residual stresses. A scanner with high feed rates is used to move the laser several times over the welding contour. Thus, the irradiation zone is heated nearly simultaneously. Molten material is squeezed out of the welding zone because of the applied clamping force. This leads to a relative movement between the two joining partners and therefore to improved gap closing capability.

To model the load, two different simulation approaches are studied, in order to find best compromise between simulation time and result precision.

B. Transient Simulation

A transient simulation model in ANSYS is built up. Thus, equation (2) is directly implemented as time dependent load function, representing the movement of the laser beam. The laser is guided one time over the specimen and the provided heat load is directly calculated in ANSYS. Heat conduction and dissipation is considered during heating phase (laser on considered section) and cooling phase (laser out of considered section).

The main drawback of this approach for simulation of quasi-simultaneous welding is, that the thermal load must be calculated again for every scan repetition, even if the applied load remains constant. Since the laser beam is moved several (approx. 50) times over the workpiece, the computation is time consuming.

C. Equivalent Heat Load

A thermal simulation model presented by Schmailzl [7] is used. Motivation of this approach is to reduce simulation time with a simplified model. As a result of a high feed rate of several hundred millimeter per second, the heat transfer during the passage of the beam can be neglected. Thus, the integral heat load of a laser passage (3) is calculated in advance in Matlab. After the transfer to ANSYS, the precalculated thermal load is directly applied to the model, simplifying the heating. The cooling phase is computed time dependent again.

$$Q_{(z,y)} = \int_{-\infty}^{+\infty} \frac{8 \cdot \alpha \cdot P}{\pi \cdot {d_{(z)}}^2} \cdot \exp\left(-\alpha \cdot z - 8 \frac{(v_x \cdot t)^2 + y^2}{{d_{(z)}}^2}\right) dt \quad (3)$$

D. Simulation and Comparision

Constant parameters for both simulation approaches are material parameters for PA 6 with absorption coefficient α = 0.85 1/mm, laser focus diameter d_{focus} = 0.18 mm and Rayleigh length z_R = 0.3 mm. Table 1 displays simulation's varying parameters energy per unit length, feed rate v and laser power P. Considering simulation time exposure, transient simulation needs a simulation time of approximately 5 min per beam passage. Equivalent heat load needs 4 minutes calculation time in Matlab, 30 seconds traffic time for loading data into Ansys and 3 minutes simulation time per beam passage.

At reduced feed rates (11.1 mm/s and 33.3 mm/s), the transient simulation approach leads to lower maximum temperatures, compared to equivalent heat load (see Fig. 2). This temperature difference decreases with rising feed rates and is negligible for feed rates larger than 100 mm/s.

Table 1: Correlation between the varying parameters for simulation. Feed rate (left, grey) and power P (white) define energy per unit input (light grey, top).

	Energy per unit length			
feed rate	0.1 J/mm	0.15 J/mm	0.2 J/mm	0.4 J/mm
11.1 mm/s	1.11 W	1.67 W	2.22 W	4.44 W
33.3 mm/s	3.33 W	5.00 W	6.67 W	13.3 W
100 mm/s	10 W	15 W	20 W	40 W
300 mm/s	30 W	45 W	60 W	120 W
900 mm/s	90 W	135 W	180 W	360 W

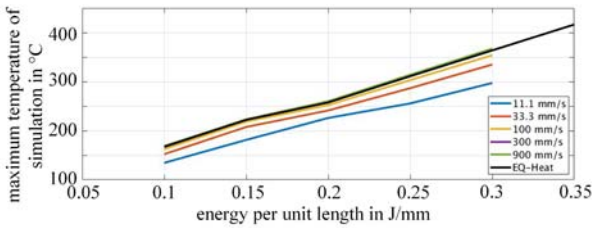


Figure 3: Influence of feed rate on maximum temperature reached in simulation.

Similarly, a reduced feed rate at the same energy per unit length leads to smaller and thinner weld seams. Fig. 3 displays the distance from upper and lower end of the weld seam to specimen's surface (left) and the maximum seam width (right). The weld seam border is defined as the isotherm of polyamide 6's melting temperature (T_{melt} = 235 °C). The welding result is good, when the upper end of the weld seam lies inside the upper joining partner with a distance to surface > 0 mm, and the lower end of the weld seam lies inside the lower joining partner (distance to surface > 1 mm). It shows, that both, upper and lower seam limits, are moving together with decreasing feed rates, causing smaller weld seams. In particular, a feed rate of 11.1 mm/s leads for energies per unit lengths smaller than 0.25 J/mm to seams not covering both joining partners. Therefore, the process window is smaller for slower feed rates. The effect of feed rate variation on weld seam width is similar.

Again, with feed rates exceeding 100 mm/s the difference between equivalent heat and transient simulation is negligible. Since slow feed rates offers more time for heat dissipation into ambient material, maximum temperature and seam extension are lower compared with higher feed rates.

Considering the definition of a good weld as one covering both specimen and leaving the upper surface unmolten, welding of two 1 mm thick PA 6 samples is possible at a Rayleigh length of 0.3 with energy per unit length between 0.2 J/mm and 0.3 J/mm. Furthermore, a faster feed rate is preferable, since the energy concentration is better because of less thermal losses, leading to a stabilized process. Additionally, welding with less energy is possible.

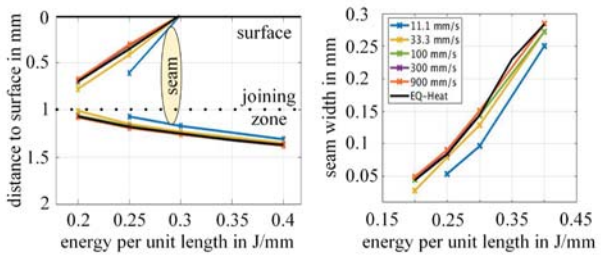


Figure 4: Distance of upper and lower end of weld seam to specimen surface (left) and weld seam width (right) calculated at different feed rates (colored lines), or with equivalent heat model (black line).

III. Experimental

A stand with high NA fixed focus optic is built up to verify the simulated process window of PA 6. In addition, tests with COC are performed.

A. Experimental Setup

The fixed focus test stand consists of a rail carrying optical elements. The beam of a thulium fiber laser (λ = 1940 nm) is guided and shaped by an optical fiber, a collimator, an adjustable beam expander and a Galilean telescope with high NA (> 0.6) focusing lens. Fig. 4 shows the experimental

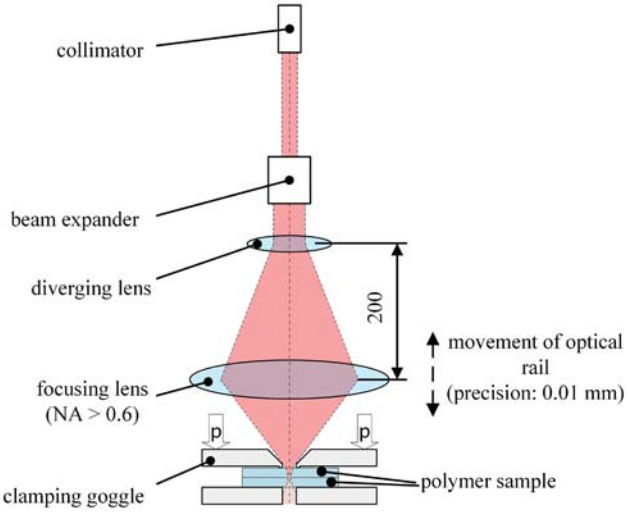


Figure 2: Experimental setup

setup. To enable distance variation between optics and specimen, a fine threaded spindle moves the rail. A measurement system with 0.01 mm resolution controls the position of the rail. Two specimens (50x20x1 mm^3 each) are fixed in an overlap by a clamping device and moved by a 2-axis-linear system with feed rates up to 300 mm/s.

B. Welding Tests – PA 6

Prior to the experiment small seams are processed with varying focus position. Thin cuts of these seams are prepared, using a rotational microtome. Using polarized light under a microscope, the seams are evaluated and the seam best meeting the joining zone is determined. The associated z-position of the optical rail will be used as reference for the joining zone.

Welds are processed at 200 mm/s feed rate and varying laser power. Fig. 5 shows the distance between upper and lower weld seam end to specimen's surface. With rising energy per unit length, the seams height increases from 0.21 mm (0.16 J/mm) to 0.65 mm (0.35 J/mm). Consequently, the experiment shows that welding of polyamide 6 is possible between 0.16 J/mm and 0.35 J/mm. This verifies the simulated results.

Anyhow, welding is even possible at slightly lower energy per unit lengths compared to simulated results. A possible explanation for this is the reduction of absorption coefficient α of PA 6 when it gets melted, which is ignored in the simulation. This absorption coefficient drop could lead to a deeper penetration of the laser and therefore to successful welds, even at low energy per unit inputs. Furthermore, simulation's melting range is simplified with one isothermal, while in reality the melting range includes a much broader temperature range.

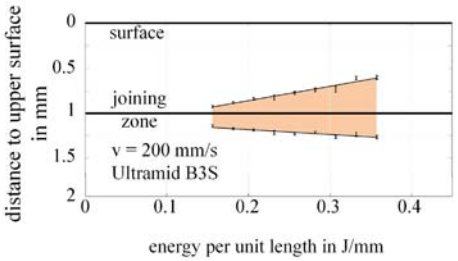


Figure 5: Distance between upper and lower weld seam end and specimen surface in dependency of energy per unit input. Material: PA 6 Ultramid B3s, Rayleigh length = 0.3 mm, λ = 1940 nm.

IV. Pyrometry for Process Monitoring

To enable process monitoring, a pyrometer is integrated in the fixed focus test stand. Welds are performed using COC (Topas 8007-04) and monitored online.

A. Experimental Setup - Pyrometer

Fig. 6 shows the existing experimental setup with integrated pyrometer. A customized pyrometer, based on a Micro-Epsilon CTM-3CF1-22 is used. Due to the necessity of a filter, eliminating laser's emission wavelength from the measurement spectrum, spectral sensitivity is decreased in the range of 2.0 μm to 2.5 μm. The analog output signal in the range of just some mV is processed using a cRIO-9035 from National Instruments with 100 kHz. The measurement spot diameter is 1.5 mm at a distance of 30 mm, measured from the front edge of the pyrometer's optic. Due to lack of space the measurement distance to specimen surface is 36 mm at a measurement angle of 25 degree towards specimen surface and thus slightly defocused. This leads to an enlarged and elliptical distorted measurement spot with 5.6 mm in feed direction and 3.1 mm rectangular to feed direction. Alignment of the pyrometric measurement spot is done with one horizontal and one vertical fine threaded spindle, each with a resolution of 0.01 mm.

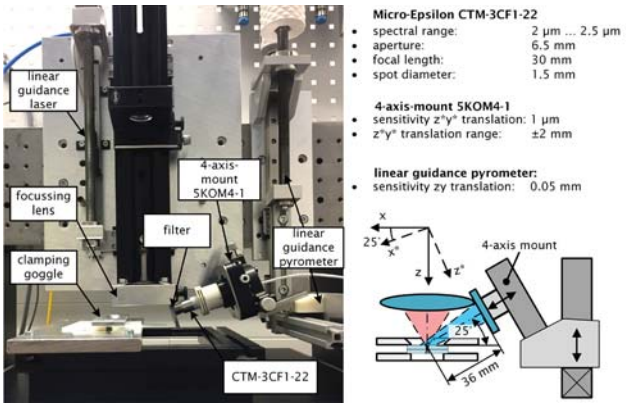


Figure 6: Integration of a pyrometer as a process monitoring unit

B. Definition of Critical Signals for Welding Success

Initially, it is important to define critical parameters for welding success. These parameters can be used to find out corresponding signals of a pyrometric measurement and therefore create the boundaries of a process monitoring unit.

As stated before, a crucial condition for welding success is the seam covering both specimens, forming a tight joint. While it is easy to find seams failed in entirety in a post processing routine, it is challenging to find local disruptions during the welding process. Furthermore, many industrial applications demand a clear surface without defects, caused by warpage or a molten surface. Thus, as a second critical factor for welding success a good surface is desired. Assessing these requirements, a locally resolved measurement is required to find signals corresponding to even the slightest defects inside seam or specimen surface.

C. Welding Tests – COC

Welds are processed using 60 W and 300 mm/s feed rate at five different focus positions. Starting at optimum focal position for welding, the optical rail is moved upwards 0.1 mm per test series, causing a focus shift inside the specimen towards the upper surface. After welding, thin cuts are produced and analyzed. The welds are classified as good, when there is a tight joint and an undamaged surface, or as damaged when there is evidence of either surface damage or loose joint. The focus shift is performed to deliberately create these defects and compare their appearance to pyrometer's signal.

Fig. 7 shows characteristic thin cuts (right) and the corresponding pyrometer signal with standard deviation calculated for 12 samples per test series (left). Signal I represent a good weld, with a tight joint and no surface defects. The signal is between 1 mV and 1.3 mV. Moving the optical rail 0.2 mm upwards, leads to beginning surface damages and a rising signal between 2.3 mV and 2.9 mV (II). These beginning surface damages may or may not lead to visible and palpable defects of the surface and can be detected evaluating thin cuts. With a rising signal over 3.5 mV a

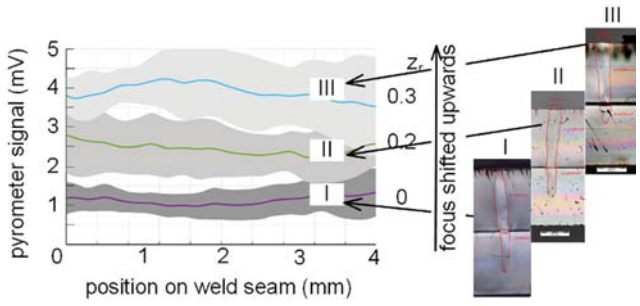


Figure 7: Pyrometer signal (left) of a tight weld with no surface damage (I), a tight weld with beginning surface damages (II) and a highly damaged surface (III). On the right side: characteristic thin cuts of each signal.

damaged surface is indicated (III). Burns of the surface layer lead to an unsteady signal and thus to a rising standard deviation. Signals under 1 mV are below signal threshold and lead to no analyzable data. These welds often have a loose joint and must be checked.

The experiment shows, that it is in principle possible to classify welding results in good and faulty parts using a pyrometer. Due to the limited pyrometer signal a loose joint cannot clearly divided from a tight joint. These welds have to be checked manually, to reduce the amount of false negative parts. However, operating between 1 mV and 2 mV leads to tight joints without surface impairments.

V. Conclusion and Outlook

Aim of this work was the identification of process parameters resulting in a good weld as well as development and application of pyrometer-based process monitoring. First, two different simulation approaches are set up to examine the influence of the interplay between feed rate and laser power for process result. Using a fixed focus test stand, welds are performed, confirming the simulated parameter. Subsequently, a customized pyrometer is added to the test stand and welds are performed under online temperature surveillance. It appears, that indirect weld seam localization with online pyrometry is possible. Further work on interaction and impact of measurement spectrum and material parameters for an optimized measurement signal will be done.

Acknowledgment

The authors gratefully thanks the Bavarian Ministry for Economic Affairs, Media, Energy and Technology for funding the project “3D-Laspyrint-Scanner” and the project partners Micro-Epsilon Messtechnik GmbH & Co. KG, Bayerisches Laserzentrum GmbH, Nexlase GmbH, LPKF WeldingQuipment GmbH and Gerresheimer Regensburg GmbH for kindly providing technical support and good teamwork. Thanks to Futonics Laser GmbH for providing a laser.

References

[1] N. Nam-Phong, M. Brosda, A. Olowinsky, and A. Gillner, “Absorber-Free Quasi-Simultaneous Laser Welding For Microfluidic Applications,” JLMN, vol. 14, no. 3, 255-261, 2019, doi: 10.2961/jlmn.2019.03.0009.

[2] Mamuschkin, V., Olowinsky, A., van der Straeten, K. u. Engelmann, C.: Laser transmission welding of absorber-free thermoplastics using dynamic beam superposition. High-Power Laser Materials Processing: Lasers, Beam Delivery, Diagnostics, and Applications IV. SPIE Proceedings. SPIE 2015, 93560Y

[3] A. Schmailzl, S. Steger, and S. Hierl, “Process Monitoring at Laser Welding of Thermoplastics,” *LTJ*, vol. 12, no. 4, pp. 34–37, 2015, doi: 10.1002/latj.201500029.

[4] A. Schmailzl, B. Quandt, M. Schmidt, and S. Hierl, “In-Situ process monitoring during laser transmission welding of PA6-GF30,” in *10th CIRP Conference on Photonic Technologies*, pp. 524–527.

[5] V. Mamuschkin, A. Haeusler, C. Engelmann, A. Olowinsky, and H. Aehling, “Enabling pyrometry in absorber-free laser transmission welding through pulsed irradiation,” *Journal of Laser Applications*, vol. 29, no. 2, p. 22409, 2017, doi: 10.2351/1.4983515.

[6] A. Schmailzl, S. Hüntelmann, T. Loose, J. Käsbauer, F. Maiwald, and S. Hierl, “Potentials of the ALE- Method for Modeling Plastics Welding Process-es, in Particular for the Quasi-Simultaneous Laser Transmission Welding // Potentials of the ALE- method for modeling plastics welding processes, in particular for the quasi-simultaneous laser transmission welding,” *Mathematical Modelling of Weld Phenomena*, no. 13, S. 965–975., 2019, doi: 10.3217/978-3-85125-615-4-51.

[7] A. Schmailzl, B. Geißler, F. Maiwald, T. Laumer, M. Schmidt, and S. Hierl, “Transformation of Weld Seam Geometry in Laser Transmission Welding by Using an Additional Integrated Thulium Fiber Laser,” in *Lasers in Manufacturing Conference 2017*.

[8] S. Polster, Laserdurchstrahlschweissen transparenter Polymerbauteile: Zugl. Diss: FAU Erlangen- Nuernberg, Univ. Bamberg: Meisenbach, 2009.

Modelling the Mechanical Behavior of the Intervertebral Disc

Anna Heinz
Laboratory of Biomechanics
OTH Regensburg
Regensburg, Germany
anna.heinz@st.oth-regensburg.de

***Abstract*— The intervertebral disc (IVD) and its mechanical properties have been subject of many studies. Different rheological models have been developed to describe the non-linear, time-dependent deformation behavior when subjected to loading. The material parameters used in those models are generally obtained by optimization methods, minimizing the error between experimental creep data and the model predictions.**

In this study, we aimed to develop a rheological model, which could then be calibrated using in vivo data obtained from measurements of spinal shrinkage using a precision stadiometer.

We built our model based on a model published in literature. It is a three-parameter standard linear solid (SLS) model, which uses time-varying material parameters ($E_1(t)$, $E_2(t)$, $\eta(t)$) instead of constant material parameters. The time dependent material parameters are obtained by optimization. Trying to replicate results from the study we used as reference, we utilized three different optimization methods. From the resulting material parameters, we then calculated a linear least-squares regression to obtain a linear relationship between the material parameters and time.

The different optimization methods resulted in different optimized material parameters over time. Only the values for the instantaneous elastic modulus were similar for all three methods. When comparing the results with the published data, variations in the curve shape were visible. The calculated linear regressions were also different from the reference data, yielding other values for slope and intercept. Squared correlations coefficients (R^2) calculated for the regressions indicated weak correlations, ranging from 0.163 to 0.420, with only one exception of 0.949.

For further use of the model, the optimization methods will have to be investigated to find the best fit and a different regression method will have to be applied.

Keywords—Intervertebral disc, viscoelasticity, creep, rheological model

I. INTRODUCTION

The IVD, located between two vertebral bodies, fulfill an important function within our spine. They provide flexibility to the spine, absorb and transmit loads. The IVD is made up of three main structures: The gelatinous nucleus pulposus in the center of the IVD, which is surrounded by the concentric lamellae of collagen fibers of the annulus fibrosus, which is again confined by the cartilaginous structures of the vertebral endplates.

The application of external loads results in a deformation of the IVD and especially changes in disc height. Subjected to a complex loading pattern throughout the day, the overall disc height has been found to decrease during the day but is known to return to its original height following recovery overnight [1]. Many studies have been conducted to measure the decrease of IVD height following specific activities, often using a device called 'precision stadiometer' [2]. With this device the decrease in disc height can be measured by examining the integral spinal shrinkage in vivo [3, 4].

However, when analyzing the mechanical properties of the spine, in vitro experiments are often conducted. The reaction of the IVD to an applied load is not linearly but exhibits a time-dependent strain response. This creep behavior under axial compression force has been reported by many researchers [5–9]. To characterize the mechanical properties of the disc, studies have tried to model the viscoelastic behavior IVD using rheological models. Rheological models are used to mathematically describe the relation between stress and strain, using the simple mechanical models of springs and dashpots. (With a combination of springs and dashpots, representing elastic and viscous material behavior, viscoelastic materials can be modeled.) There are different implementations of such lumped parameter rheological models. One which is commonly used is the three-parameter standard linear solid (SLS) model [5, 8]. It combines a Kelvin body (spring in parallel with a dashpot), representing the creep response, with a spring in series, accounting for the initial elastic response. Other variations used were 5-parameter models (2 Kelvin bodies and a spring in series), or the Maxwell fluid (spring in series with dashpot) in parallel with a spring [10, 11]. Stretched exponential functions such as the Kohlrausch-Watts-Williams (KWW) model have been used to fit experimental creep data, providing a good fit [12, 13]. The IVD experiences an increase in stiffness, when subjected to axial strain or following cyclic loading [12, 14, 15]. This nonlinear response cannot be accounted for with the SLS model, whose parameters are strain independent. To account for this nonlinearity, Groth and Granata [16] included a strain-dependency of the instantaneous elastic modulus, creating the standard nonlinear solid (SNS) model. Yang et al. [17] addressed this matter by using time-varying parameters in the SLS model.

Many such models have been developed to describe the creep behavior of the IVD, using experimental data from in vitro experiments to obtain the material parameters used within these models.

The primary goal of this study was to develop a method to model the mechanical behavior of the IVD. This model could then be used with in vivo measurements of spinal shrinkage obtained from a precision stadiometer, allowing the calibration of the model to a living subject.

Table 1
Bounds used during optimization, oriented by the range within which the optimized material parameters within the model by Yang et al. [17]

Bounds	E_1 [MPa]	E_2 [MPa]	η [MPa s x 10^4]
Lower	10	5	1
Upper	20	15	5

II. Material and Methods

The rheological model used in this study is based on the one implemented by Yang et al. [17]. It is a version of the classic SLS model, which consists of a Kelvin body (spring and dashpot parallel) in series with a spring. E_1 and E_2 are the elastic moduli of the Kelvin body and the spring respectively, η the viscosity coefficient of the Kelvin body. The model is described by the following differential equation:

$$\sigma + \frac{\eta}{E_1+E_2}\dot{\sigma} = \frac{E_1E_2}{E_1+E_2}\varepsilon + \frac{E_1\cdot\eta}{E_1+E_2}\dot{\varepsilon} \quad (1)$$

where σ is the total stress, ε the total strain, and $\dot{\sigma}$ and $\dot{\varepsilon}$ the rate of change of stress and strain with respect to time. For a constant stress σ, the following creep equation can be obtained by integration:

$$\varepsilon(t) = \frac{\sigma}{E_2} + \frac{\sigma}{E_1}\left(1 - e^{-t/\tau}\right) \quad (2)$$

and

$$\tau = \frac{\eta}{E_1} \quad (3)$$

To transform the constant parameter model to a time-varying parameter model, equation (2) can be written as follows for a constant loading frequency:

$$\varepsilon(t) = \frac{\sigma}{E_2(t)} + \frac{\sigma}{E_1(t)}\left(1 - e^{-t/\tau}\right) \quad (4)$$

and

$$\tau = \frac{\eta(t)}{E_1(t)} \quad (5)$$

The time-varying material parameters can then be optimally determined for each time step by minimizing the error between experimental data and the predicted deformation of the model. As no experimental data was available yet, an exemplary creep curve was created, based on the data published by Yang et al. [17]: The creep curve was created for a constant force amplitude and frequency (F=100N, f=8 Hz) with the material parameters for t=0 calculated by the model:

$$E_1 = 18.889\ [\text{MPa}] \quad (6)$$

$$E_2 = 12.421\ [\text{MPa}] \quad (7)$$

$$\eta = 1.638\ [\text{MPa s x } 10^4] \quad (8)$$

To find the best fit, different optimization methods available through the SciPy library available for use in Python were compared [18]. The least-squares method (*scipy.optimize.least_squares)* (LSQ), the minimize algorithm (*scipy.optimize.minimize)* (MIN), and the basin-hopping method *(scipy.optimize.basinhopping)* (BH) were utilized, trying to replicate the results presented by Yang et al. [17]. For the minimize and basin-hopping, which combines the minimize algorithm with a global stepping algorithm, the so-called method 'L-BFGS-B' was used. All methods were used with the bounds that can be found in Table 1 and as initial guess for the to be optimized material properties the values from equations (6-8) were used. These values were updated after each iteration to the results of the last iteration.

To obtain a linear relationship between the optimized time-varying material parameters and the time, according the work of Yang et al. [17], a linear least-squares regression was calculated. The squared correlation coefficient R^2 was calculated to evaluate the goodness of the regression.

Those resulting equations can then be substituted into equations (4) and (5), resulting in the final creep equation $\varepsilon_{F,f}(t)$.

Table 2
Fitted parameter equations obtained by the linear least-squares regression for each material parameter and methods and the according squared correlation coefficient R^2.

Material Parameter	Optimization Method	Fitted Parameter Equation	R^2
$E_1(t)$	**LSQ**	14.811 – 2.75 x 10^{-4} t (MPa)	0.360
	MIN	15.496 – 3.173 x 10^{-4} t (MPa)	0.404
	BH	14.282 – 3.472 x 10^{-4} t (MPa)	0.271
$E_2(t)$	**LSQ**	9.010 – 2.264 x 10^{-4} t (MPa)	0.352
	MIN	9.006 – 2.358 x 10^{-4} t (Mpa)	0.369
	BH	8.998 – 2.315 x 10^{-4} t (MPa)	0.359
η(t)	**LSQ**	1.331 – 1.278 x 10^{-1} t (MPa)	0.163
	MIN	1.637 – 5.240 x 10^{-4} t (MPa)	0.420
	BH	1.669 – 2.544 x 10^{-1} t (MPa)	0.949

III. Results

The optimization utilizing three different methods yielded differing results. The graphs resulting from the material parameters plotted over time are shown in Fig. 1. For E_2 the graphs of all three methods are similar, while for E_1 only the LQS and the MIN method show comparable results. The calculated regression equations and the corresponding squared correlation coefficients (R^2) can be found in Table 2. The resulting regressions resulted in a linearly declining function for all methods and material parameters. R^2 of the calculated regression ranged between 0.163 and 0.420 for 8 out of 9, with one exception of 0.949 (η, BH).

IV. Discussion

To describe the deformation of the disc over time, different approaches can be found.

Stretched exponential functions generally provided a good fit, also when used for cyclic loading, but were not able to correctly describe to initial and final parts of creep [12]. Hwang et al. [11] further showed that the model was unable to detect the influence of preload on the creep behavior, in contrast to a lumped parameter rheological model. One of the simplest lumped parameter models is the standard linear solid (SLS) 3-parameter model. Burns et al. [8] showed that

the model yielded good results for a constant axial stress. However, the application of this model was limited when utilized for cyclic loading [19]. Addressing the nonlinearities arising from an increase in stiffness in the response to axial strain and cyclic loading [12, 14, 15], models with nonlinear components were introduced. The model by Groth and Granata [16] included a strain-dependent instantaneous elastic modulus. This improved the prediction of the response to cyclic loading at low frequencies of 0.01 Hz but was not able to predict it correctly at higher frequencies, suggesting that viscosity coefficient might also be nonlinear. Yang et al. [17] also used the SLS model as a basis, but adapted it by adding time-varying material parameters. If vibration amplitude and frequency are held constant, the fact that the strain is time-dependent can be utilized to introduce time-dependent material parameters. The model description that was derived from those non-constant material parameters showed better prediction results than a model with constant material parameters. This model was used as the basis for the model developed in this study.

Using different optimization methods, we tried to replicate part of the results in the study from Yang et al. [17] Looking at the graphic illustration of the optimized material parameters over time, it can be seen that for E_2 all methods show similar results. E_2 is the instantaneous elastic modulus, which is directly proportional to the creep, therefore yielding the best and similar results during the optimization. The other parameters however show more differences between the different methods, which arise from the different algorithms that underly each of the methods.

As Yang et al. presented a linear relationship between material parameters and time in their study, we also tried to find a linear curve fit for our optimization results. The squared correlation coefficients greatly variied, with the lowest being 0.163 (η, LSQ) and the highest 0.949 (η, BH). No other R^2 is higher than 0.42, which indicates a very weak correlation in general. When examining the regression curves plotted together with the original material parameter optimization data (Fig. 1), it can be seen that a linear regression doesn't represent the curve shape very well. The obtained material parameters mostly resemble the shape of a negative exponential function when plotted over time, therefore a linear fit is not ideal to represent the relationship between material parameters and time. The only exception with a R^2 of 0.949 (η, BH) didn't show this characteristic curve shape.

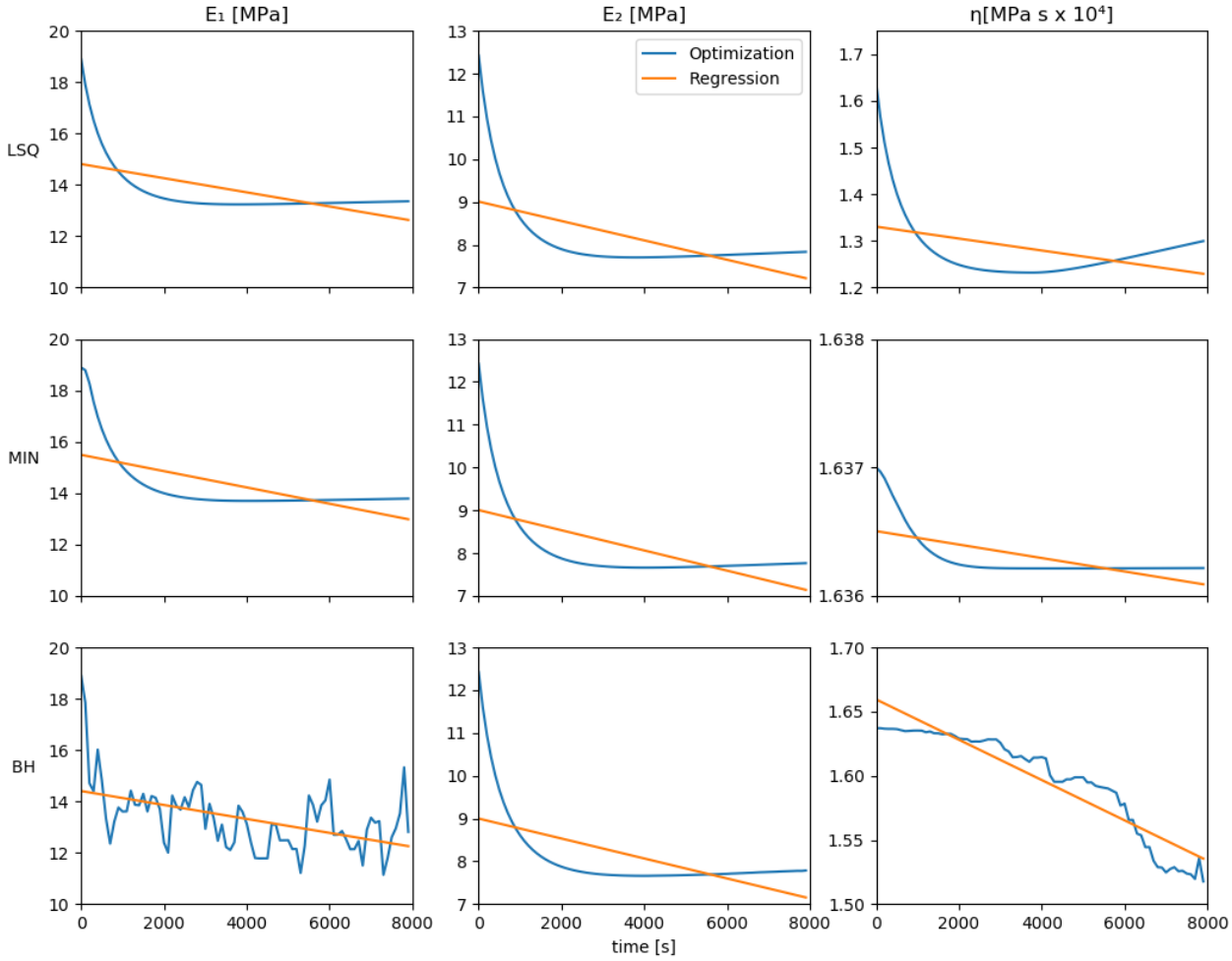


Fig. 1 Effect of optimization method on material parameters. Top row shows results obtained by using the least-squares method (LSQ), middle row the minimize method (MIN) and bottom row results obtained by the basin-hopping method (BH). The regression lines (orange) obtained by the linear least-squares regression have been added to the original data (blue).

When comparing the results to those of Yang et al., clear differences can be found. In their study, the elastic moduli (E_1 and E_2) both decline with time, whereas the viscous coefficient (η) increases, which is consistent with what Pollintine et al. reported in their study [20]. These trends can't be found in our data. The elastic moduli E_1 and E_2 obtained by LSQ and MIN (and also E_2 obtained by BH) indeed decline with time, although not steadily (as Yang et al. report) but rather sharp in the first quarter, followed by a small incline towards the end. The viscous coefficient however differs from the positive trend reported by Yang et al. but instead shows the same characteristics as the elastic moduli, when calculated with the LSQ and MIN method, even though within different ranges. When comparing the resulting regression equations with respect to slope and intercept, the slope they predicted was around 2.5 times larger than ours and the intercept differed by around 2-3 MPa. For E_2, the slope was similar, but the intercept differed by around 3 MPa again. While Yang et al. reported a positive slope of 3.8×10^{-4}, our slopes varied between -1.3 $\times 10^{-4}$ and -5.2 $\times 10^{-4}$, whereas the intercepts were similar. The curves they obtained by fitting the data show an R^2 of over 0.94, which indicates a very good fit. It is unclear however, whether they used the whole timespan from t=0 to t=8000 or if they fitted their data over a shorter timespan, which might result in different curves.

We were not able to reproduce the results presented by Yang et al. [17] within our study. It remains to be clarified, whether further refinements regarding the parameters of the optimization methods, such as bounds or tolerances, will lead to different results and to results, which show more resemblance with those we tried to replicate. The linear regression method used in this study was used because we wanted to recreate existing data, but the squared correlation coefficients indicate weak correlations for most cases. To be able to further use the material parameters as a function of time, a different regression method will have to be used. The resulting time-varying material parameters can then be used within the creep equation (Eqs. (4) and (5)). Using spinal shrinkage data from measurements with a precision stadiometer, the model could then be calibrated using in vivo data instead of in vitro creep data.

As the model is developed based on a constant force amplitude (F) and frequency (f), the resulting creep equation ($\varepsilon_{F,f}(t)$) would only be applicable within this scope. To use it for different loading scenarios, a range of creep equations would have to be obtained, requiring several experiments with varying loads and frequencies.

V. Conclusion

This study aimed to develop a method to model the mechanical parameters of the IVD. With the model chosen, a time-varying three parameter SLS model, replication of published results was not possible. This models' applicability for the use with in vivo data therefore remains to be verified. Different modeling approaches should be considered for the aim of this study. Models such as the standard SLS model, which have been used successfully by other researchers, might be more suitable for the application with in vivo data.

VI. References

[1] T. Reilly, A. Tyrrell, and J. D. Troup, "Circadian variation in human stature," *Chronobiology international*, vol. 1, no. 2, pp. 121–126, 1984, doi: 10.3109/07420528409059129.

[2] J. A. Eklund and E. N. Corlett, "Shrinkage as a measure of the effect of load on the spine," *Spine*, vol. 9, no. 2, pp. 189–194, 1984, doi: 10.1097/00007632-198403000-00009.

[3] I. Althoff, P. Brinckmann, W. Frobin, J. Sandover, and K. Burton, "An improved method of stature measurement for quantitative determination of spinal loading. Application to sitting postures and whole body vibration," *Spine*, vol. 17, no. 6, pp. 682–693, 1992, doi: 10.1097/00007632-199206000-00008.

[4] M. Magnusson, E. Hult, I. Lindstrom, V. Lindell, M. Pope, and T. Hansson, "Measurement of time-dependent height-loss during sitting," *Clinical biomechanics (Bristol, Avon)*, vol. 5, no. 3, pp. 137–142, 1990, doi: 10.1016/0268-0033(90)90016-Y.

[5] T. S. Keller, D. M. Spengler, and T. H. Hansson, "Mechanical behavior of the human lumbar spine. I. Creep analysis during static compressive loading," *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, vol. 5, no. 4, pp. 467–478, 1987, doi: 10.1002/jor.1100050402.

[6] M. A. Adams and W. C. Hutton, "The effect of posture on the fluid content of lumbar intervertebral discs," *Spine*, vol. 8, no. 6, pp. 665–671, 1983, doi: 10.1097/00007632-198309000-00013.

[7] L. E. Kazarian, "Creep characteristics of the human spinal column," *undefined*, 1975. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/1113976

[8] M. L. Burns, I. Kaleps, and Kazarian L.E., "Analysis of compressive creep behavior of the vertebral unit subjected to a uniform axial loading using exact parametric solution equations of Kelvin-solid models—Part I. Human intervertebral joints," *Journal of Biomechanics*, vol. 17, no. 2, pp. 113–130, 1984, doi: 10.1016/0021-9290(84)90129-5.

[9] W. Koeller, F. Funke, and F. Hartmann, "Biomechanical behavior of human intervertebral discs subjected to long lasting axial loading," *Biorheology*, vol. 21, no. 5, pp. 675–686, 1984.

[10] G. D. O'Connell, N. T. Jacobs, S. Sen, E. J. Vresilovic, and D. M. Elliott, "Axial creep loading and unloaded recovery of the human intervertebral disc and the effect of degeneration," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 4, no. 7, pp. 933–942, 2011, doi: 10.1016/j.jmbbm.2011.02.002.

[11] David Hwang, Adam S. Gabai, Miao Yu, Alvin G. Yew, and Adam H. Hsieh, "Role of load history in intervertebral disc mechanics and intradiscal pressure generation," (in En;en), *Biomech Model Mechanobiol*, vol. 11, no. 1, pp. 95–106, 2012, doi: 10.1007/s10237-011-0295-1.

[12] W. Johannessen, E. J. Vresilovic, A. C. Wright, and D. M. Elliott, "Intervertebral disc mechanics are restored following cyclic loading and unloaded recovery," *Annals of biomedical engineering*, vol. 32,

no. 1, pp. 70–76, 2004, doi: 10.1023/b:abme.0000007792.19071.8c.

[13] J. J. Sarver and D. M. Elliott, “Mechanical differences between lumbar and tail discs in the mouse,” *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, vol. 23, no. 1, pp. 150–155, 2005, doi: 10.1016/j.orthres.2004.04.010.

[14] W. T. Edwards, W. C. Hayes, I. Posner, A. A. White, and R. W. Mann, “Variation of lumbar spine stiffness with load,” *Journal of biomechanical engineering*, vol. 109, no. 1, pp. 35–42, 1987, doi: 10.1115/1.3138639.

[15] A. Kaigle, L. Ekström, S. Holm, M. Rostedt, and T. Hansson, “In vivo dynamic stiffness of the porcine lumbar spine exposed to cyclic loading: influence of load and degeneration,” *Journal of Spinal Disorders*, vol. 11, no. 1, pp. 65–70, 1998.

[16] K. M. Groth and K. P. Granata, “The viscoelastic standard nonlinear solid model: predicting the response of the lumbar intervertebral disk to low-frequency vibrations,” *Journal of biomechanical engineering*, vol. 130, no. 3, p. 31005, 2008, doi: 10.1115/1.2904464.

[17] X. Yang, X. Cheng, Y. Luan, Q. Liu, and C. Zhang, “Creep experimental study on the lumbar intervertebral disk under vibration compression load,” *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine*, vol. 233, no. 8, pp. 858–867, 2019, doi: 10.1177/0954411919856794.

[18] P. Virtanen *et al.*, “SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python,” *Nat Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[19] S. Li, A. G. Patwardhan, F. M.L. Amirouche, R. Havey, and K. P. Meade, “Limitations of the standard linear solid model of intervertebral discs subject to prolonged loading and low-frequency vibration in axial compression,” *Journal of Biomechanics*, vol. 28, no. 7, pp. 779–790, 1995, doi: 10.1016/0021-9290(94)00140-Y.

[20] P. Pollintine, M. S. L. M. van Tunen, J. Luo, M. D. Brown, P. Dolan, and M. A. Adams, “Time-dependent compressive deformation of the ageing spine: relevance to spinal stenosis,” *Spine*, vol. 35, no. 4, pp. 386–394, 2010, doi: 10.1097/BRS.0b013e3181b0ef26.

Efficient Implementation of Neural Networks on Field Programmable Gate Arrays

Kilian Märkl
University of Applied Sciences Regensburg
Regensburg, Germany
Email: kilian1.maerkl@st.oth-regensburg.de

***Abstract*—This paper investigates efficient implementation methods for Neural Networks on Field Programmable Gate Arrays, especially in the context of System on Chips. Programmable logic is becoming more and more available as a resource on such devices and offers, especially due to its true parallelism and flexibility, the ability of custom hardware accelerators. Introductory three basic concepts for the implementation of accelerators with regard to Neural Networks are discussed. Following, a CNN for traffic sign recognition is selected as application, which provides the basis for verifying the introduced methods. The first one is the implementation strategy and the overall design. Thereby a full realization of the Neural Network architecture with half-pipelining on the Field Programmable Gate Array is proposed, for a maximum relief of the processor and a straightforward development. The second method specifies a generic layer design, which can be used as a basis for all types. This allows a fast realization of the layers, especially for new ones. As final method for an efficient execution on Field Programmable Gate Arrays, the quantization of the parameters and the use of fixed-point arithmetic is introduced. Therefore, a statistical evaluation of the possible numerical intervals inside the Neural Network is used. In addition, further improvements and extensions of the introduced methods are given in the outlook.**

***Index Terms*—Neural Networks, CNN, FPGA, Embedded Systems, System-on-Chip, Hardware Accelerator, Implementation**

I. Introduction

Neural Networks (NNs) are solving more and more real-world problems, like classification [1], object detection [2], segmentation [3] and speech recognition [4], just to name a few. Therefore, the ambition to run NNs also on embedded systems is steadily growing, because they offer low costs, small shapes and low power consumption compared to mainly used workstations with Graphics Processing Units (GPUs). This opens up new fields of application and an enormous market. But up to now, there have been only few realizations, due to the limited computational resources of embedded processors. Thus, real time requirements, especially for state of the art NNs, could hardly be fulfilled yet.

However, there is also an increasing number of embedded systems which combine processor(s) and Field Programmable Gate Array (FPGA), so called System on Chips (SoCs). These devices allow to implement custom hardware or accelerators and thus represent a powerful system. Especially FPGAs are particularly well suited for the calculation of NNs, due to their true parallelism and their flexible configuration and design possibilities. Furthermore, they have hundreds or thousands of Digital Signal Processor (DSP) slices which can perform computational operations like multiply-accumulate in one clock cycle.

Therefore, this paper researches efficient implementations for NNs on SoCs to enable their economical use. There are different approaches and ideas of how to realize or accelerate NNs on FPGAs. Some concepts are explained in more detail in the following section.

II. Theory

There are three fundamental implementation strategies to speedup the execution of NNs using FPGAs, into which almost every approach can be divided. First, there is the use of simple accelerator(s) inside the FPGA. The second strategy is to implement one Processing Element (PE) for each type of layer of the NN and multiplex these resources to replicate the sequence of calculations. The third method is to rebuild the whole structure of the NN inside the FPGA with multiple PEs.

A. Accelerators

The first kind of accelerators are very simple and generic. They perform basic operations like dot products or matrix multiplications. Since the major part of the computations of NNs consists of such operations, the processor uses these accelerators to outsource and speedup a great part of the calculations. In addition, these types are not limited to NNs and can be used for various applications. But there are some drawbacks, such as a huge data exchange between the processor and the FPGA and consequently a high bus traffic. Furthermore, the processor has a significant software load, since it still has to coordinate the whole execution of the NN. Nevertheless, one example for an implementation of such an accelerator is [5], where a matrix processing unit is used to speedup big data analysis.

Another type of accelerators calculates the result from an entire layer of a NN, e.g. a convolution layer [6] of a Convolutional Neural Network (CNN). They are more complex and sometimes specialized for their application. The processor uses these accelerators in a similar way as the type mentioned before, but greater parts of the computations are outsourced as coherent entities. As a consequence, fewer interactions are required between the processor and the FPGA, leading to lower overhead and less bus traffic than the generic

accelerators. Despite this, there is one drawback, the processor still has a considerable software load to prepare, process and coordinate the data flows of the NN. An example of such a type is [7], where a convolution accelerator is implemented on FPGA cards to speedup the forward propagation and reduce the power consumption of CNNs calculated by Microsoft's data centers.

The general architecture of FPGA accelerators on SoCs is shown in Figure 1. The processor within the processing system is connected to the FPGA, referred to as programmable logic, via the on-chip bus system. The FPGA contains one or several hardware accelerators, exemplified in this figure by the two types mentioned above.

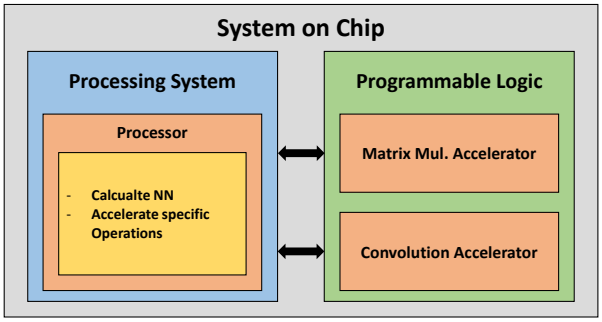


Figure 1. System architecture with accelerators

B. Multiplexed Layers

Another way to accelerate the execution of NNs on FPGAs is to use multiplexed layers. In this approach, each type of layer used in the network is implemented as a single PE or also called Processing Unit (PU). For a common CNN, e.g. one Convolution PE, one Max-Pooling PE and one Fully-connected PE are required. These modules are parallel to each other and each output of the PEs can be fed back to any input. This allows to multiplex and reuse the layers, represented by the corresponding PEs, for the execution of one pass through the NN. The configuration of how the PEs are multiplexed can either be fixed in hardware or adjustable via software. With such an implementation, the processor only has to input the data, read back the results, and initially set or coordinate the control of the PEs. This further reduces the load of the processor compared to the accelerators. Due to the multiplexing of the layers and the resulting reuse of the PEs, this approach is also resource-efficient and very adaptable, since new structures of NNs only change the configuration of how the layers are switched. However, there are also some disadvantages, such as an increased complexity and thus more control logic, which has to be implemented in the FPGA. Furthermore, the kernels of the convolution and the weights of the fully-connected layers inside the PEs have to be updated on each pass through. Examples for such implementations are [8] and [9]. Both designed and realized multilayer NNs, also known as Deep Neural Networks (DNNs), and specifically fully-connected feedforward nets with the multiplexed layers technique. The required resources are significantly less than for a full implementation, and nevertheless a recognition speed of 15,900 frames per second has been achieved for handwritten digits with the MNIST database in [9].

In Figure 2, the general structure of a NN with multiplexed layers on a SoC is shown in a simplified form. This exemplary illustration demonstrates this for a CNN and therefore consists of a Convolution PE, a Max-Pooling PE and a Dense (Fully-connected) PE. Each output can be switched to each input via the feedback path.

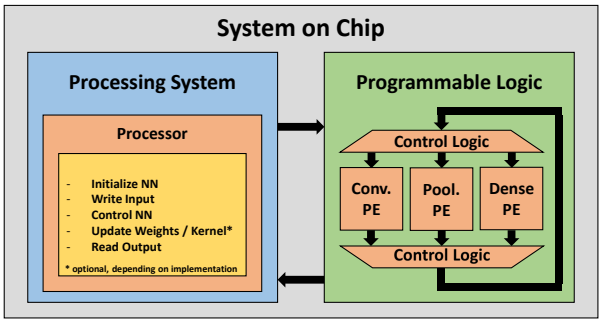


Figure 2. System architecture with multiplexed layers

C. Full Implementation

The third method is a complete implementation of the NN structure. Therefore, each layer is realized as an independent module inside the FPGA. These layers are connected according to the architecture of the NN, e.g. for a feedforward net they are cascaded in series. Thus, no complex control logic is needed to switch layers or update weights, unlike the multiplexed layers method. The implementation of the modules varies depending on the hardware. For layers of small fully-connected NNs (about less than few thousand neurons), each neuron can be realized entirely using flip-flops and DSP slices, allowing to calculate them in one or few clock cycles. Modules of greater NNs are typically using Random Access Memory (RAM) or on-chip memory to store the inputs and outputs and one or several computation units to sequentially process all the calculations of the corresponding layer. Furthermore, such an approach enables the possibility to pipeline each layer and thus speeds up the execution additionally, since all layers can be calculated simultaneously and in parallel. Note, the time for one pass through remains the same as for a non-pipelined implementation, but the throughput increases. This leads to the fastest execution time of all mentioned approaches. But even this approach has some downsides. The most significant one is the immense resource consumption on the FPGA by implementing the whole NN structure. Therefore, this method is not always feasible. An example for such a type is [10], where a small fully-connected NN with two hidden layers is implemented as a pipelined and a non-pipelined version.

In addition to the forward propagation, this NN has also the possibility to perform an online-backpropagation in parallel.

The general architecture of a fully implemented NN on the programmable logic of a SoC is shown in Figure 4, exemplary for a CNN. The layers are independent modules, cascaded in series whereby all of them can be computed in parallel.

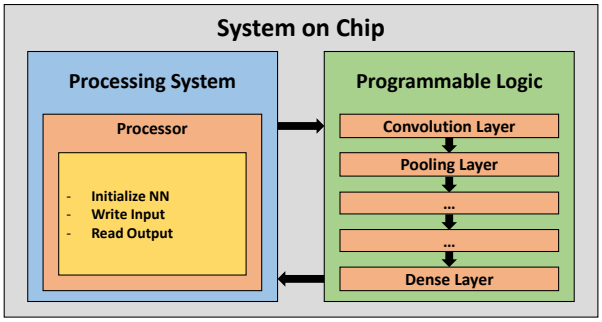


Figure 4. System architecture with full implementation

III. Neural Network

As concrete starting point and for further practical implementations and testings, a CNN for traffic sign classification is used as application. This network has already been developed, trained, and validated for investigations on a processor of a SoC in previous work [11]. Therefore, this CNN offers excellent comparative values and allows to determine a concrete acceleration factor for the FPGA.

The used network relies on the architecture of LeNet 4 [12], but with some modifications, like a different number and size of convolution kernels and neurons in the fully-connected layers and a changed activation function. The German Traffic Sign Recognition Benchmark (GTSRB) [13] was used as dataset to train the network and it reached an accuracy of 97.65% on the validation set. Sure, state of the art NNs like [14] achieve validation rates up to 99.71%, but they require 14,629,801 parameters. In comparison, the selected CNN has only 641,787 parameters and thus represents a moderate-size network, which is easier to implement for experiments and tests. The information about the individual layers are listed in Table I and the architecture of the whole CNN can be seen in Figure 3.

Table I. Description of the CNN layers [11]

Layer	Input-dimension	Kernel- / Weights-dimension	Activation-function	Output-dimension
Convolution	$64 \times 64 \times 3$	$(3 \times 3 \times 3) \times 32$	ReLU	$62 \times 62 \times 32$
Max Pooling	$62 \times 62 \times 32$	(2×2)	-	$31 \times 31 \times 32$
Convolution	$31 \times 31 \times 32$	$(3 \times 3 \times 32) \times 32$	ReLU	$29 \times 29 \times 32$
Max Pooling	$29 \times 29 \times 32$	(2×2)	-	$14 \times 14 \times 32$
Dense	6272	6272×100	ReLU	100
Dense	100	100×43	Softmax	43

IV. Methods

Following, the methods and the developed design for the implementation of the chosen CNN on the FPGA are explained. The focus is on three aspects, the overall design of the NN, the generic design of the layers, and the quantization of the parameters.

A. NN Design

The fundamental design of the NN is based on the previously mentioned full implementation approach. This reduces the load of the processor the most, and the network is completely described in hardware, thus no complex separation between soft- and hardware is required. Furthermore, this approach is straightforward to implement due to its moderate complexity.

The individual layers of the CNN are realized as independent modules and cascaded according to the architecture. This overall structure is shown in Figure 5. Each layer has an input and output RAM and, depending on the type, additional a weights and bias RAM. All modules, except the first and the last, share their input with the previous layer and their output with the following one. Therefore, these memories are referred to as buffer RAMs. The input and output memory of

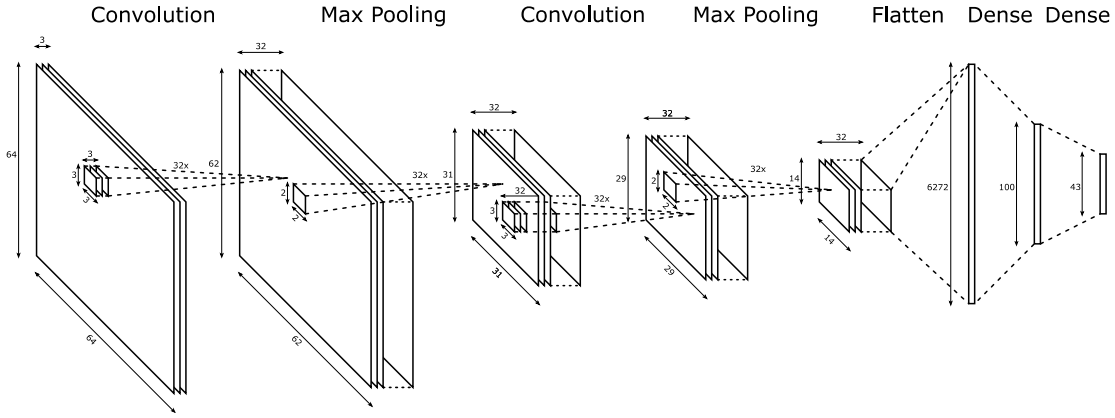


Figure 3. Architecture of the CNN

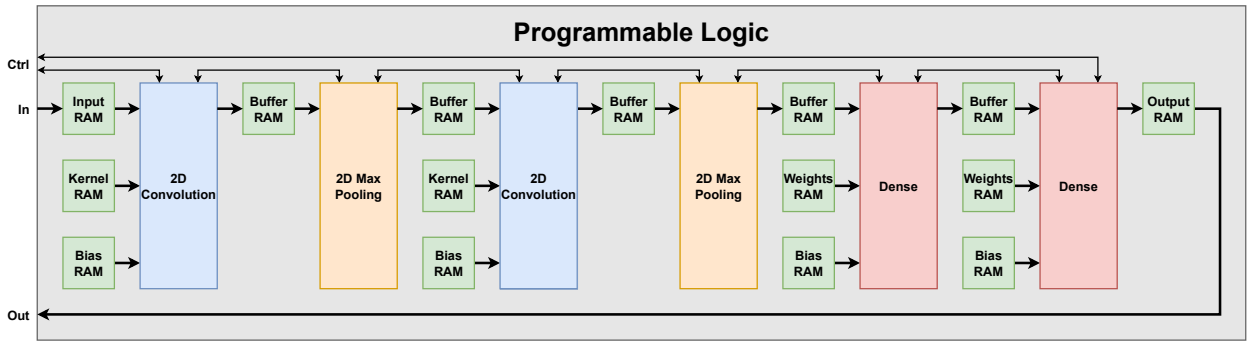


Figure 5. Overall design of the FPGA implementation

the entire network is shared with the processor of the SoC to load images and to read the results. The NN is controlled via the status and command signals of the first and the last layer, which are accessible as registers. Inside the FPGA, these signals are connected to the neighboring modules. Thus, there is no global coordination or management, instead the layers control the execution decentralized. As mentioned in theory, the full implementation allows a pipelining. However, the developed design does not support the execution of all modules in parallel. Since two layers share one buffer RAM, only one module is allowed to work with or on it, for consistency reasons. Therefore, only every second layer can run at the same time. That means, in the first execution step, every odd module runs, in the second step, every even. This reduces the execution frequency, but is more resource efficient in terms of required memory. Note, the time for a pass through the whole NN is the same.

B. Layer Design

The general design of all layers is based on one common structure. This allows a generic layout and thus fast and straightforward adaptations to potential extensions or adjustments. Figure 6 illustrates this design, which consists of the following main components:

- Finite State Machine (FSM): Indicates the status of the layer for the previous one, e.g. ready, busy or finish, and coordinates and controls its sub-modules during execution. Furthermore, the FSM checks the status of the following layer and triggers it after a complete calculation.
- Address Generator: Creates address sequences depending on the layer type for input, output, and optional for weights and bias memory to provide the required data for the calculations and to save the results. For dense layers, this is a simple up-counter, whereas for convolution and pooling layers [6], this is a complex series containing jumps and multiple calls of addresses.
- Arithmetic Logic Unit (ALU): Performs the specific calculations depending on the layer type. In case of dense and convolution layers, these are multiply-accumulate operations realized with DSP slices and compare operations for max-pooling layers implemented with logic cells. The processing is done sequentially and therefore one operation per cycle can be performed. This implies that several operations may be necessary for the calculation of one output value. E.g. for the computation of one output pixel of the fist convolution layer, 27 ($3 \times 3 \times 3$) clock cycles are required. The sequential execution requires only one ALU per layer and thus represents a very resource efficient method.
- Activation Function: Is applied on the result of the ALU and thus calculates the final value. The operation and its implementation depends on the selected function of the respective layer. E.g. for ReLU [6], it is a simple check of the sign bit, and in case of a negative number, zero is forwarded. This operation is realized with logic cells only and performed asynchronously.

Note that pooling layers neither have an activation function nor a weights and bias interface.

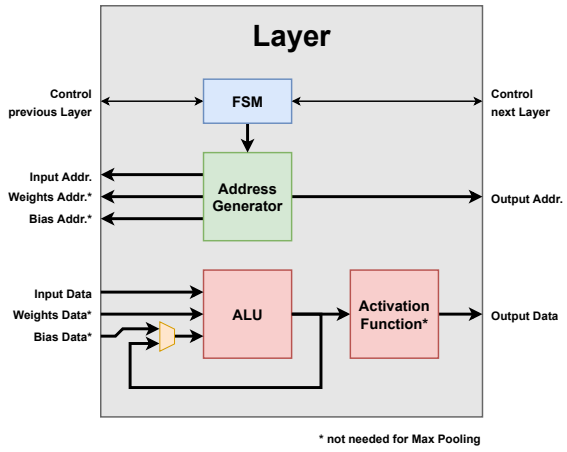


Figure 6. Generic design of a layer

Together, ALU and activation function are representing one neuron. Due to the sequential processing within the layer, this one neuron is used for all calculations and therefore constitutes a multiplexed neuron.

C. Quantization

In order to additionally accelerate the execution and to save hardware resources on the FPGA, the NN is implemented and calculated in fixed-point arithmetic. The reason for this is that most mathematical operations are implemented by using DSP slices. However, these usually support only integer or fixed-point data types, such as the DSP48E2 form Xilinx [15]. For an implementation of floating-point arithmetic, one ALU would require several DSP slices, additional logic and several clock cycles for a calculation.

To perform the computations in fixed-point arithmetic, a suitable format has to be determined first, into which the NN parameters are converted afterwards. However, it is not advisable to simply project the entire range and maximum precision of the floating-point type, since this would result in needlessly high memory consumption. The parameters and values between the layers of most NNs only vary in a certain and limited range. This range can easily be determined for the parameters with the largest and smallest value. More difficult to estimate is the range of the output values for each layer, because it cannot be predicted or limited absolutely. Therefore, in this paper, a statistical evaluation of the range is performed with a sufficient number and variation of test data. For this purpose, the training dataset is used. A similar method is used for the 8-bit quantization from TensorFlow Lite, introduced in [16]. The determined ranges define the integer part of the fixed-point type. For the fractional part, hence the accuracy, the number of bits of the floating-point mantissa is used. Another common approach, which can be applied additionally, is the determination of the required precision by gradual reduction and verification using the validation data set.

V. Conclusion

In this paper, three methods for the practical realization and implementation of a CNN on a SoC were introduced. The first one is the overall design of the NN on the FPGA. Due to the selection of a full implementation, the load on the processor is reduced to a minimum. This allows the processing system to spend more computation time for other tasks. Furthermore, the modular structure not only keeps complexity low, it also ensures a straightforward implementation. Thus, extensions in the architecture of the NN can easily be integrated. The second mentioned method is the layer design. This is characterized by the generic structure, whereby all layers have the same basis and therefore require less realization effort. Furthermore, new layers can be developed rapidly. The use of fixed-point as data type also saves hardware resources and execution time by allowing a DSP slice to perform one operation per clock cycle. In general, the developed design requires few resources like DSP slices and logic, e.g. flip-flops. Only the RAM consumption is moderate, but even this is low due to the sharing between the layers.

However, there are also some limitations in the developed design. One aspect is the execution time, which can significantly increase due to the sequential calculations, especially for large convolution layers. This leads to a further problem. Since the layers can have significantly different computational requirements, unbalanced execution times in the network arise and especially affect the efficiency of the pipelining. A further but less significant limitation is that the FPGA has to be synthesized and implemented again when the architecture of the NN is modified.

VI. Outlook

The methods developed and explained in this paper have to be implemented and tested in practice. Therefore they are realized on a Xilinx Multi-Processor SoC. In particular, the ZCU104 evaluation kit is used. Subsequently, crucial values such as execution latency, frequency and power consumption will be determined. These will be compared with the results of [11], where the selected CNN was implemented on the processor of the SoC. From these outcomes, values like the acceleration factor and the energy efficiency can be derived.

In addition to the introduced methods, there are also some approaches for further improvements. One possible extension is a full-pipelining through double buffering, which increases the throughput of the NN. Another approach is the use of multiple DSP slices per ALU, which allows several operations per clock cycle and accelerates the computations within the layers. However, the memory interface has to be adapted accordingly, to ensure that the required data is provided in parallel and in one clock cycle. One solution would be the use of 3D RAMs. Nevertheless, all these extensions represent a higher resource consumption and therefore are not necessarily feasible with regard to typical SoCs or FPGAs.

References

[1] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: A multi-class classification competition," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 1453–1460.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[4] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[5] C. Chung, C. Liu, and D. Lee, "Fpga-based accelerator platform for big data matrix processing," in *2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, 2015, pp. 221–224.

[6] A. Afshine and A. Shervine. (2019, January) Convolutional neural networks cheatsheet. CS 230 - Deep Learning. [Online]. Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks

[7] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. Chung, "Accelerating deep convolutional neural networks using specialized hardware," February 2015. [Online]. Available: https://www.microsoft.com/en-us/research/publication/accelerating-deep-convolutional-neural-networks-using-specialized-hardware/

[8] S. Himavathi, D. Anitha, and A. Muthuramalingam, "Feedforward neural network implementation in fpga using layer multiplexing for effective resource utilization," *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 880–888, 2007.

[9] T. V. Huynh, "Deep neural network accelerator based on fpga," in *2017 4th NAFOSTED Conference on Information and Computer Science*, 2017, pp. 254–257.

[10] R. Gadea, J. Cerda, F. Ballester, and A. Macholi, "Artificial neural network implementation on a single fpga of a pipelined on-line back-propagation," in *Proceedings 13th International Symposium on System Synthesis*, 2000, pp. 225–230.

[11] K. Märkl, "Effiziente Implementierung und Evaluation von Neuronalen Netzen auf Embedded-Systems," Project report, University of Applied Sciences Regensburg, 2019, unpublished.

[12] Y. Lecun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," in *International Conference on Artificial Neural Networks*, January 1995, pp. 53–60.

[13] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608012000457

[14] Álvaro Arcos-García, J. A. Álvarez García, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Networks*, vol. 99, pp. 158–165, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608018300054

[15] *UltraScale Architecture DSP Slice*, Xilinx, September 2019, UG579. [Online]. Available: https://www.xilinx.com/support/documentation/user_guides/ug579-ultrascale-dsp.pdf

[16] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," *Computing Research Repository*, 2017. [Online]. Available: http://arxiv.org/abs/1712.05877

Research of the optimal mesh for a centrifugal compressor's volute using the GCE method

Martin SAUTEREAU
Ostbayerische Technische Hochschule Regensburg
Turbomachinery Laboratory
Regensburg, Germany
martin.sautereau@sigma-clermont.fr

***Abstract*—The strive for efficiency is a common topic for many scientific investigations. Indeed, improving the yield allows either to reduce the consumed energy or to improve the effective one. A ventilator yield can reach 80% but hardly better yet. Centrifugal compressors are made of an impeller, which rotates in the housing referential. The impeller provides rotating kinetic energy to the fluid, and the volute, transforms this energy into a linear kinetic one.**

Thanks to CAD and CFD, more and more studies have been carried out to improve these machines. These are mostly concentrated on the impeller. Volutes studies are often ditched. One main problem for the CFD study of such a shape is the inner volume's meshing. It has to have an optimal size to offer the best ratio between results' quality and calculation time.

The different options and criteria to make a good mesh are tested and their efficiency are compared using the grid convergence index method. A python program has developed to perform this comparison. Once the important parameters are found, the ideal mesh will be describe.

***Index Terms*—centrifugal compressor, volute, CFD, ANSYS CFX, mesh, optimisation, python, GCI method**

I. Introduction

As turbocharger equip almost all the new internal combustion engine on the market, it is becoming more and more important to have them optimized. Despite the ecological changes in our society, these engines will always be used, whether with new fuels or in certain areas, where they are needed. That is why the flow machines' laboratory of the OTH Regensburg has developed a test stand for these components.

Before the tests, prototypes first have to be investigated in a computational fluid dynamic (CFD) software. Discretization is maybe the trickiest part of the simulation process. In order to reduce the computing time, some simplification can be made. Knowing which computing method should be used also helps defining a proper mesh. All these choices depend on the geometry of the volute and the kind of fluid we want to simulate. The boundary conditions and dimensions of our simulations would be the same than on the test stand. The heart of this work would be the meshing of the said volume considering the flow in it.

At first, we have to define the geometry we want to use. Considering the study settings, the solving technique is chosen. This will be helpful to choose the proper type of mesh. The Gride Convergence Index (GCI) method will allow to determine, which meshing options are useful and which one increases the computing time without improving the quality. All simulation are carried under Ansys CFX.

II. Scientific and technical background

A. Volute's vocabulary

The volute is the snail shape of the centrifugal compressor, its goal is to collect the flow coming from the impeller and to bring it to the outlet pipe. The fluid first flows trough the volute inlet, then goes to the volute itself and goes to the axial diffusor through the volute outlet.

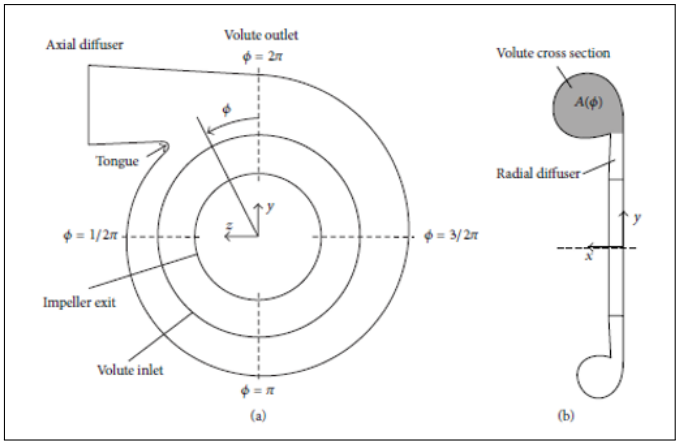


Figure 1. Definition of the volute geometry [8, p. 4]

Usually, the volute cross-section varies linearly with azimuthal angle (ϕ on Figure 1).

B. Different kinds of mesh

For further details cf. [11]

Most common CFD problems require the Finite Volume Method (FVM) or the Finite Difference Method (FDM). Finite Element Method (FEM) is only efficient for very complicated geometries. Depending on the simulation method that is used, the mesh can differ.

A mesh can be structured or unstructured. A structured mesh is generally made of hexaeder whereas an unstructured mesh can be made of almost any form but is usually made of tetraeders. Meshes can also be a mixture of structured and unstructured meshes.

Geometry fitted meshes must follow the contours of the geometry.

Cartesian meshes are the right alternative to geometry fitted meshes. CFD tools using this mesh type offer more precise and efficient algorithms. But the mesh lines cannot be fitted to bodies of complex geometries. In locations where body boundaries do not fit the cartesian mesh, interpolations must be used in order to take the effects of the misalignment between mesh and geometry into account.

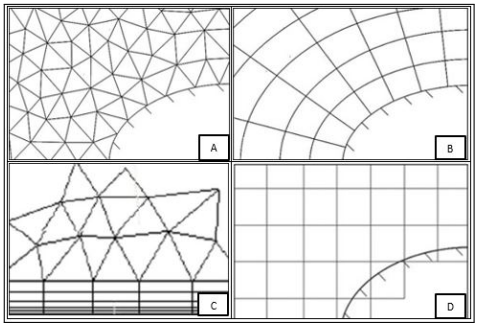


Figure 2. Different kinds of mesh: unstructured (A), structured (B), hybrid (C) and Cartesian (D) [11]

C. Different simulation methods

For further details cf. [7]

Depending on the flow, its geometry, the precision wanted and the calculation time available, the calculation method can change.

The most accurate calculations rely on the kinetic gas theory: it allows to simulate all kind of fluid with all the frictions and rotational considerations. The main problem of this method is its need for a Cartesian mesh which cannot fit to most geometries.

The Navier-Stokes equations simulate the same phenomenon but only for low Knudsen numbers. Indeed, if the Knudsen number is much lower than 1, the flow is considered as a continuum flow. In this case, the Navier-Stokes equations can be used. This is probably the most used method in aerodynamic and hydrodynamic: it can simulate the phenomenon with any kind of mesh.

In order to reduce the calculation time, the Euler equations or the potential theory can be used. None of them take the frictions into account. The potential theory does not even consider the rotations.

Table I sums up these paragraphs.

D. The outer layer

In order to model the friction caused by the walls, the mesh quality is usually improved in the near wall region. It usually implies to use a hybrid mesh as shown in Figure 2. Tohttps://de.overleaf.com/project/5f0ebf973a1e3600018e8c2d obtain such a mesh, the ANSYS "Inflation" option is used.

Because of its high Reynolds number(Re), the flow in a volute is turbulent, it's velocity u is spread into four zones, depending on it's distance to the wall y. This distance and

	friction	rotations
Kinetic gas theory (without restricition)	yes	yes
Navier Stokes equation (incl. boundary layer) (for continuum flows ($K_n << 1$) newtonian fluids)	yes	yes
Euler equations	no	yes
Potential theory	no	no

Table I
DIFFERENT CALCULATION METHODS [7, P. 11]

speed are described using y^+ [3] and u^+ [2] which are dimensionless speeds and distance. These are described as follow:

$$y^+ = \frac{u_* y}{\nu} \tag{1}$$

and

$$u^+ = \frac{u}{u_*} \tag{2}$$

where, according to [6] and [14]

$$u_* = \sqrt{\nu(\frac{\partial u}{\partial y})_{y=0}} \tag{3}$$

Calculating y^+ is made easier by some program such as the one presented by [1].

According to [15] and [7] the four velocity zones shown in Figure 3 have the following characteristics:

- at the wall contact, where $y^+ = 0$, the wall friction is so high that $u^+(0) = 0$,
- in the viscous sublayer, where $0 < y^+ < 5$, $u^+ = y^+$,
- in the buffer layer, where $5 < y^+ < 30$, the turbulence effects starts being stronger than the viscous one, y^+ is hard to estimate, the maximal deviation is by $y^+ \approx 11$ and
- in the log-law region, where $30 < y^+ <\approx 200$, $u^+ = \frac{1}{k} \ln y^+ + C$ where $k = 0,41$ and $C = 5,0$.

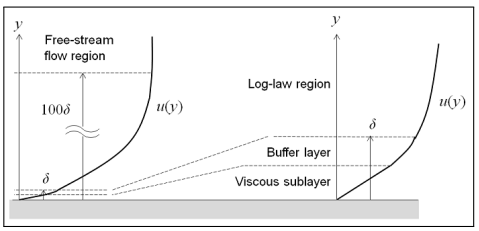


Figure 3. The four regime of turbulence [5]

The buffer layer and viscous sub layer are usually modeled rather than simulated.

E. Evaluating the quality

Different method can be used in order to evaluate the quality of a mesh. The aspect ratio investigates the shape of each cell whereas the grid convergence index (GCI) focuses on the deviation of the simulation related to the mesh quality.

1) Aspect ratio: The aspect ratio of a geometric shape is the ratio between its longest and its shortest edge. It is always be bigger than 1. The smaller it is, the better it is. An optimisation goal can be to minimise the average aspect ratio of a shape or to minimise the maximal one.

2) Grid convergence index (GCI): For further details cf. [10]
The GCI quantifies the evolution between different meshes. It takes into account the result of the simulation; it allows the evaluation of the uncertainties related to the mesh in the final results. It usually compares three meshes based on their results for one or several performance criteria.

The first step of the process is the definition of the grid size for all the different meshes. For a 3D simulation, the grid size is defined as:

$$h = \sqrt[3]{\frac{1}{N}\sum_{i=1}^{N}\Delta V_i} \tag{4}$$

where N is the number of cells and ΔV_i the volume of the i^{th} cell. It actually is the cubic root of the average volume and therefore represents an average side length of each cell.

The three simulations should be named 1, 2 and 3 so that $h_1 < h_2 < h_3$. The grid size evolution between two meshes are $r_{21} = \frac{h_2}{h_1}$ and $r_{32} = \frac{h_3}{h_2}$. ϕ_k is the value of the k^{th} simulation. $\epsilon_{ij} = \phi_i - \phi_j$ is the result's deviation between two meshes. The grid refinement factor is defined as $r = \frac{h_{coarse}}{h_{fine}}$ it should be bigger than $1,3$.

Knowing that $s = \text{sign}\left(\frac{\epsilon_{32}}{\epsilon_{21}}\right)$:

$$q(p) = \ln\frac{r_{21}^p - s}{r_{32}^p - s} \tag{5}$$

and

$$p = \frac{\left|\ln\left|\frac{\epsilon_{32}}{\epsilon_{21}}\right| + q(p)\right|}{\ln r_{21}} \tag{6}$$

can be solved using the fixed point algorithm method [4]. The function q is only here to simplify Equation 6 whereas p is a dimensionless variable usefull to calculate the GCI.

This allows to calculate the extrapolated values and the different errors from coarse and fine simulations:

- The extrapolated value: $\Phi_{ext}^{21} = \frac{r_{21}^p\Phi_1 - \Phi_2}{r_{21}^p - 1}$,
- the approximate relative error: $e_a^{21} = \left|\frac{\Phi_1 - \Phi_2}{\Phi_1}\right|$,
- the extrapolated relative error: $e_{ext}^{21} = \left|\frac{\Phi_{ext}^{12} - \Phi_1}{\Phi_{ext}^{12}}\right|$ and
- the GCI: $GCI_{fine}^{21} = \frac{1.25 * e_a^{21}}{{r_{21}}^p - 1}$.

Φ_{ext}^{32}, e_a^{32}, e_{ext}^{32} and GCI_{coarse}^{32} would be calculated the same way. All these data must be calculated again for all of the different measured values. In our case, HEINRICH [8] recommends to optimise the static pressure recovery coefficient C_p and the total pressure loss coefficient K_p:

$$C_p = \frac{p_{out} - p_{in}}{p_{t,in} - p_{in}} \text{ and} \tag{7}$$

$$K_p = \frac{p_{t,in} \quad p_{t,out}}{p_{t,in} - p_{in}}. \tag{8}$$

The GCI is a percentage describing the mesh quality's evolution between two meshes.

III. SIMULATING

A. Form to mesh

In order to perform a CFD simulation on a CAD part, the inner volume should first be extracted. Indeed, the CFD program only deals with fluid volumes. The physical characteristics of the piece holding the flow would be given through the boundary conditions, the step just before computing.

The geometry is simplified but its dimensions respect the orders of magnitude of the test stand's volute. Indeed, the goal of this study is to develop a meshing method in order to always have a comparable mesh for each new volute shape to test.

To prevent numerical mistakes, an outlet cone is added after the diffuser. This outlet cone does not really exist, it is just added for the simulations' stability.

The results of our simulations are taken at the cone's inlet. Apart from the outlet cone, all the surfaces are producing shear stress on the fluid, these are named "no slip wall". The outlet cone is a friction free surface ("free slip wall") because it is not a real surface and therefore should not restrict the fluid in any way.

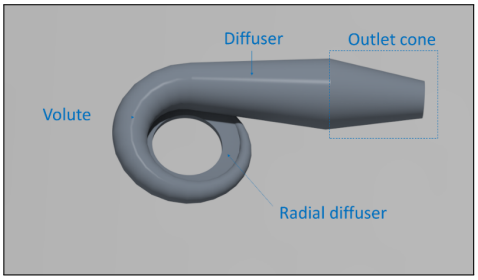


Figure 4. Modeled volute with the outlet cone highlighted

In order to measure C_p and K_p (as presented in Equation 7 and Equation 8) on the outlet of the volute, the geometry is separated into two geometries: the outlet cone and the volute itself. It is otherwise not possible to measure the outlet's pressure and static pressure.

B. Simulation method

Due to the complex shape of the volute, setting a cartesian mesh was not possible. That's why the kinetic gas theory equations are not usable.

According to GRÜN [7], the Navier Stokes equations allow to use almost any kind of mesh, depending on the computing method used.

The Navier Stokes equations allow to simulate rotation and friction. Therefore these are ideal for us. The only restrictive criteria is the Knudsen number of the flow. It should be much smaller than 1. This dimensionless number [7] compares the

average distance between two molecules λ and a characteristic length L:

$$Kn = \frac{\lambda}{L}. \tag{9}$$

In this case, we are working on a compressor, so the pressure obviously varies between the inlet and the outlet. The pressure's order of magnitude is around $1\,\mathrm{bar}$, the average free space is therefore around $68{,}10 \cdot 10^{-9}\,\mathrm{m}$. The order of magnitude of the characteristic length is one meter. Knowing all that, the Knudsen number can be calculated. Its order of magnitude is 10^{-7} which is much smaller than 1. Even if this calculation was carried out only with order of magnitudes, the results are so far from the limit that it does not require any further investigations. The calculation would all be carried using the Navier-Stokes equations

Not all turbulences can be simulated, the smallest ones must be modeled using turbulence models. Two of the most commonly used turbulence models are the k-omega $(k-\omega)$ and k-epsilon $(k-\epsilon)$ models. $k-\omega$ and $k-\epsilon$ are respectively efficient close and far from the wall, the $k-\omega-SST$ (SST means Shear-Stress-Transport) is a mix of the advantages of them both. That's why this model is used.

C. Evaluation of the meshes' quality

The chosen method for the evaluation of the mesh quality relies on the GCI. It would allow to evaluate the importance of each functionality of the meshing tool. It focuses on the deviation made by the mesh. The aspect ratio method mostly focuses on the aspect of the mesh. Even if this is an interesting data, it is not enough to compare two meshes.

In order to use the GCI method, a python program was developed to perform all the calculations previously mentioned. Depending on the programmer, the structure of the program may vary. In our case, it is structured into three parts: the importation of the python libraries, the definition of the functions, the operative part of the program.

As for all python programs, the first step is to import all the libraries which provide useful functions for the code. For this program only the numpy and math library were needed.

All data of each simulation are stored in a matrix where each line describes a mesh. Each column describes an element of this simulation, respectively: "name of the mesh", "number of cells" and "result" of the simulation for the watched criteria". Actually, a Python matrix is a list of lists. During the run of this program, a "grid size" column is added. This matrix is named "liste" in the python program. Using the function "tri4el_liste", a new matrix named "datas" is created, it would hold the same data than "liste" except that the lines would be ordered differently. They are sorted so that the first line shows the simulation with the smallest grid size and the last line the biggest grid size. In order to simplify the writing of the final program, each variable useful to obtain the GCIs has a dedicated function.

```
def tri4el_liste(hi):
    """sorts a matrix according to the value
    of its 4th column"""
    hf = []
    hf=sorted(hi,key=lambda column: column[3])
    return hf
```

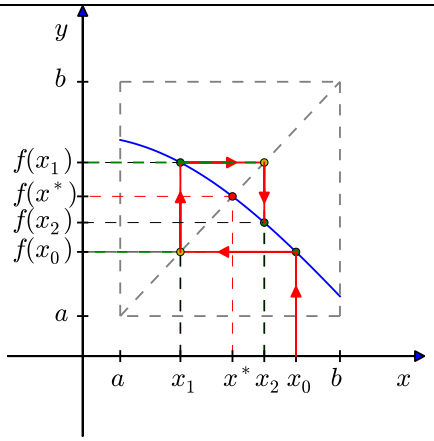


Figure 5. Illustrated iterative process [16]

This function makes sure that $h_1 < h_2 < h_3$ [10].

To calculate the grid size, the "h" function has been developed. In order to simplify the calculations, h has been defined as $h = \sqrt[3]{\frac{vol}{N}}$ where vol is the overall volume of the meshed volume instead of $h = \sqrt[3]{\frac{1}{N}\sum_{i=1}^{N}\Delta V_i}$.

```
def h(x, vol):
    """calculates h"""
    return (vol/x)**(1/3)
```

This approximation does not have a significant impact on the solution. The sum of all the volumes is very close from the overall volume. Such a sum is not possible to code in python. Indeed, iterative process are limited to 1000 iterations in python and the order of magnitude of the meshes' sizes is 100 000 elements. Even if this approximation may lead to some difference, it does not matter because the grid size's goal is only to say which of these three meshes is the coarsest and which one is the finest.

Having a grid refinement ratio bigger than $1,3$ is not the only constraint applying to the meshes. The three element number must be evenly distributed. If not, an over flow error might appear during the computing as shown below.

```
OverflowError: (34,'Numerical result out of range')
```

One important function of this algorithm is the fixed point iteration function. It is named fixedp in this program. This function has been taken from the glowing python which is an interactive python blog [13]. The fixed point iteration algorithm allows to solve equations looking like $x = f(x)$. The main idea is summed up in Figure 5. Starting from a x_0 value, the algorithm would then calculate x_1 using the iteration step $x_{i+1} = f(x_i)$. The iterations are repeating until the error $e = |x_i - x_{i-1}|$ is smaller than a defined value: 10^{-5} in our case.

The condition of convergence (for a function named g instead of f) according to the Indian institute of technology of Madras is:

"If $g(x)$ and $g'(x)$ are continuous on an interval J about their root s of the equation $x = g(x)$, and if $|g'(x)| < 1$ for all x in the interval J then the fixed point iterative process $x_{(i+1)} = g(x_i)$, $i = 0, 1, 2, \ldots$, will converge to the root $x = s$ for any initial approximation x_0 belongs to the interval J." [4]

The ASME [10] recommends this method, suggesting that these conditions are met.

The automatically sorting of the meshes with the function "tri4el_liste" is not the only user-friendly measure of this program. Indeed, everything is commented and explained, so that an unexperimented user can understand it and use this program. He must only read and understand what the GCI is.

IV. Experimenting

Different aspects of the mesh have been investigated: the default element size (DES), the default growth rate (GR) and the inflation parameters. By meshing the geometry, the outlet cone automatically tends to get a structured mesh with a low number of elements.

A. simulation parameters

To perform the simulation, the boundaries conditions are defined. As described in subsection III-A, the diffuser's walls are set to "no slip wall" and the outlet cone's walls as "free slip wall". The inlet's velocity is set to $90\,\mathrm{m\,s^{-1}}$. As said in subsection III-B, the used turbulence model is the $k-\omega-SST$ model. The outlet's is set to a defined mass flow, which corresponds to the inlet [12, p. 44].

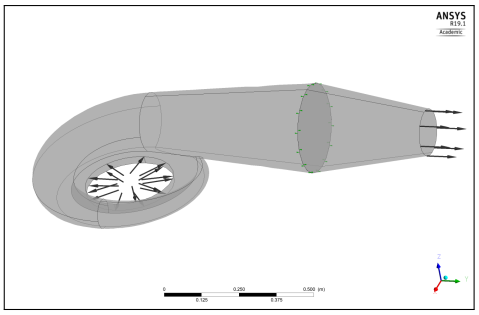


Figure 6. the geometry in the ANSYS-CFX, while setting the boundary conditions

B. Global mesh's parameters

1) Default element size: The DES is a value quantifying the mesh's refinement. For our geometry, the default element size can vary between $10\mathrm{mm}$ and $1000\mathrm{mm}$.

It was the first criteria to be inspected. The study has been carried out without using any inflation. Apart from the DES, all the mesh values have been left to the default one. The results in Table II were obtained simulating three time precisely the same conditions, just the DES changes between two simulations. C_p and K_p are calculated with the simulations' results using the Equation 7 and Equation 8.

Table II
DES STUDY

Mesh description	A DES=10mm	B DES=100 mm	C DES=1000 mm
number of element	267 039	49 788	29 147
P_{in}(Pa)	1 797,24	2 451,86	2 277,81
$P_{t,in}$(Pa)	6 601,25	7 096,42	7 010,6
P_{out}(Pa)	716,809	1 402,55	592,944
$P_{t,out}$(Pa)	1 705,91	1 416,97	1 589,88
C_p	-0,224902	-0,225922	-0,355998
K_p	1,01901	1,22282	1,14535

Table III
GCIs OF THE DES STUDY

	C_p	K_p
finest	$1,37335.10^{-9}\%$	$0,255011\%$
coarsest	$0,00564409\%$	$0,325490\%$

Table III shows how the pressure values can vary just by changing the meshes.

The only needed values are the GCIs between the coarsest and the finest meshes for the pressure recovery coefficient and the total pressure loss coefficient. Having different GCI values shows that, whether we want to investigate C_p or K_p, their are two different optimal meshes. Table III shows that total pressure loss coefficient, is way more impacted by the DES than the pressure recovery coefficient.

Table III shows that the DES have no influence on C_p. The DES' influence on K_p rate is very small too.

2) Growth rate: The influence of the GR have been investigated. The experimental results are presented in Table IV. The simulations have been carried out using the default mesh settings and no inflation. The default element size has been set to $100\mathrm{mm}$. The GCI study (see Table IV) shows us that even if C_p is well impacted by the GR, K_p feels it once again even more.

3) Global mesh's parameters' summary: This first part of the study has shown us that C_p is less impacted by these two

Table IV
GR STUDY

Mesh description	A GR=1,1	B GR=3,045	C GR=4,99
number of elements	111 741	18 090	15 107
P_{in} (Pa)	2 834,47	2 531,44	2 429,93
$P_{t,in}$ (Pa)	7 633,27	7 205,39	7 091,96
P_{out} (Pa)	1 015,02	277,586	128,873
$P_{t,out}$ (Pa)	2 141,11	1 270,38	1 118,98
C_p	-0,379146	-0,482216	-0,493574
K_p	1,14448	1,26980	1,28119

Table V
GCIs OF THE GR STUDY

	C_p	K_p
finest	$1,54691\%$	$2,27020\%$
coarsest	$1,48345\%$	$1,92980\%$

Table VI
FLH STUDY

Mesh $e_1 =$	A $0,014mm$	B $0,14mm$	C $1,4mm$	D $14mm$
Number of elements	81 691	73 548	58 976	34 469
P_{in} (Pa)	2661,56	3830,67	2837,47	1580,83
$P_{t,in}$ (Pa)	7020,18	8421,6	7525,77	6281,5
P_{out} (Pa)	831,465	1484,79	964,75	216,572
$P_{t,out}$ (Pa)	1754,66	2472,52	2001,27	1096,45
C_p	-0,419879	-0,510981	-0,399445	-0,290226
K_p	1,208070	1,295833	1,178359	1,103045

parameters than K_p.

Comparing Table III and Table V shows that in order to have a efficient mesh to study C_p, it is useless to have a fine DES. It is more important to have a fine GR. Even if the the DES has an impact on K_p, it stays neglectable in comparison to the GR.

Even if the most optimal mesh might not be the same for both C_p and K_p, having just one mesh for both of the criteria allows to run the simulation only once, which saves even more time.

C. Inflation parameters

Several criteria describe the inflation layers of a mesh. The tested ones are the first layer thickness (e_1) and the maximal number of layers (n, named "Maximum layers" in ANSYS). The inflation growth rate (IGR) cannot be tested because a fine enough grid refinement ratio cannot be reached. The total thickness (e_n) is described in Equation 10, it is basically the sum of the terms of a geometric progression (a geometric serie) of common progression IGR and start value e_1.

$$e_n = e_1 \frac{1 - IGR^n}{1 - IGR}. \tag{10}$$

Based on the y^+ calculation (see **the outer layer**), the first layer thickness should be set to $e_1 = 0,14\mathrm{mm}$ to include the whole buffer layer and viscous sub-layer ($y^+ = 30$).

1) First layer height: The "First layer height" (FLH) obviously describes the thickness of the first inflation layer. for this experiment, the DES has been set to $1000\mathrm{mm}$ and the GR to $1,2$. The "Maximum layers" has been set to 20.

The simulation results are presented in Table VI. In order to use a big enough grid refinement ratio, the A and D meshes must be used. The number of elements' difference between the A and the B meshes is too small, it generates an overflow error (as mentioned in **Evaluation of the meshes' quality**). Therefore, the used meshes are A, C and D.

The GCI study (see Table VII) shows that having a too thin first layer is not an effective meshing parameter. As $e_1 = 0,14\mathrm{mm}$ is not considered for this study, this result has to be mitigated.

2) Maximum layers: In order to study the influence of the "Maximum Layers" (ML) inflation option, the default "Element Size" has been set to $1000\mathrm{mm}$, the inflation GR and the default GR have been set to 1,2.

Table VII
GCIS OF THE FLH STUDY

	C_p	K_p
finest	0,045399%	0,081715%
coarsest	0,111666%	0,115298%

Table VIII
ML STUDY

Mesh Description	A n=25	B n=5	C n=1
Number of elements	76 647	53 779	34 320
P_{in} (Pa)	4144,57	-4283,34	2618,22
$P_{t,in}$ (Pa)	8698,49	-62,1356	7339,54
P_{out} (Pa)	1629,55	-6494,64	479,525
$P_{t,out}$ (Pa)	2653,27	-5531,09	1409,35
C_p	-0,552276	-0,523855	-0,452987
K_p	1,327476	1,295591	1,256045

To have a big enough grid refinement, the FLH has been set to $e_1 = 0,07\mathrm{mm}$. With a FLH of $0,14\mathrm{mm}$ the grid refinement would not have been enough.

The Table VIII presents the three meshes used for this experiment and their pressures.

The Table IX shows that the ML has an impact on C_p but not really on K_p.

3) Inflation parameters' summary: This second Part of the study has shown us that the inflation does not have a substantial impact on K_p. The "Maximum Layer" influences C_p.

The fact that none of these two inflation criteria can reach the GR's GCI, can show that most of the turbulence are situated in the main body of the volute, not in the boundary layers. But it is just an hypothesis which should be investigated deeper.

V. CONCLUSION AND PERSPECTIVE

What is clearly illustrated with Table X is that the DGR is the ultimate parameter for the C_p and K_p simulation. The maximal layer number is important for C_p but, for K_P the DES is the second most important parameter.

Table IX
GCIS OF THE ML STUDY

	C_p	K_p
finest	1,096944%	0,018781%
coarsest	1,263440%	0,021721%

Table X
PARAMETER RANKING BASED ON THEIR GCI REGARDING TO C_p AND K_p

GCIs	C_p	K_p	GCIs
DGR_{finest}	1,5469%	2,2702%	DGR_{finest}
$DGR_{coarsest}$	1,4835%	1,9298%	$GR_{coarsest}$
$ML_{coarsest}$	1,2634%	0,3255%	$DES_{coarsest}$
ML_{finest}	1,0969%	0,2550%	DES_{finest}
$FLH_{coarsest}$	0,1117%	0,1153%	$FLH_{coarsest}$
FLH_{finest}	0,0454%	0,0817%	FLH_{finest}
$DES_{coarsest}$	0,0056%	0,0217%	$ML_{coarsest}$
DES_{finest}	0,0000%	0,0188%	ML_{finest}

Table XI
THE BINARY RESEARCH PROCESS

stage	$e_1(mm)$	Nodes	Elements	$Sum = Nodes + Elements$
1	14	101166	346635	447801
2	7,7	103543	319021	422564
3	4,55	120327	309936	430263
4	2,975	140232	335734	475966
5	2,1875	155678	360182	515860
6	2,58125	146606	344193	490799
1	1,4	180353	402867	583220

Starting from the coarsest mesh possible, it will be improved step by step until the limit of 512 000 cells and knots allowed by the ANSYS student version is reached [9]. As shown on Figure 7, the first step is to set the DGR its finest level ($GR = 1,1$), the medium DGR is skipped (it will not make sense to go backward to generate a mesh with the medium DGR). Based on the same logic, the parameters will be improved following this sequence:

- DGR would be set to its finest value,
- ML would be set to its medium value,
- ML would be set to its finest value,
- DES would be set to its medium value,
- DES would be set to its finest value,
- FLH would be set to its medium value and at least
- FLH would be set to its finest value.

GCIs	C_p	K_p	GCIs
DGR_{finest}	1,5469%	2,2702%	DGR_{finest}
$DGR_{coarsest}$	1,4835%	1,9298%	$GR_{coarsest}$
$ML_{coarsest}$	1,2634%	0,3255%	$DES_{coarsest}$
ML_{finest}	1,0969%	0,2550%	DES_{finest}
$FLH_{coarsest}$	0,1117%	0,1153%	$FLH_{coarsest}$
FLH_{finest}	0,0454%	0,0817%	FLH_{finest}
$DES_{coarsest}$	0,0056%	0,0217%	$ML_{coarsest}$
DES_{finest}	0,0000%	0,0188%	ML_{finest}

Figure 7. Optimisation steps illustrated on Table X

Once one of the criteria can not be improved without overcoming the ANSYS student's limit, a binary researched is performed to get as close as possible to the limit.

The mesh can be improved until the first layer is set to $1,4\text{mm}$. Their is then 58 3220 (which is obviously bigger then 512 000) elements and nodes in the mesh. The goal is now to find a value for e_1 between 14mm and $1,4\text{mm}$ so that the elements and nodes number is just below the limit. Generating a mesh with a "First Layer Height" bigger than the DES means that no inflation will be generated. In our case, as long as $e_1 < 10\text{mm}$, no inflation will be generated. With a hand made binary research shown in Table XI, the optimal mesh is set with an "First Layer Thickness" of $2,6\text{mm}$. It then have 489 602 knots and elements which correspond to $95,6\%$ of the biggest mesh available with ANSYS Student.

According to this study, the most efficient mesh, will have the characteristics presented in Table XII

These mesh parameters should be used for the next step of this project: finding through CFD simulations the optimal shape for a centrifugal compressor's volute.

Table XII
BEST MESH'S SPECS

parameter	value
DGR	1,1
Maximal Layers	25
DES	10mm
First Layers Height	$2,6\text{mm}$

Table XIII
ABBREVIATIONS

Abbreviation	**meaning**
CFD	Computational fluid dynamic
CAD	computer assisted design
FEM	Finite element method
FVM	finite volume method
FDM	finite difference method
GCI	Grid convergence index
ASME	American society for mechanical engineers
SST	Shear-stress-transport
DES	default"element size"
GR	Growth rate
IGR	Inflation growth rate
FLH	First layer height
ML	Maximum layers

ACKNOWLEDGEMENT

This paper is written as part of the RARC Regensburger Applied Research Conference 2020. Also it has to be noted that this paper is a summary of the current state of the project in July 2020 and therefore cannot fully represent the final results of this project.

LEXICON

For nomenclature, abbreviations and indices see tables XIII to XV.

REFERENCES

[1] *CFD Online - Y-Plus Wall Distance Estimation*, [Online; accessed 2. Jul. 2020], Jul. 2020. [Online]. Available: https://www.cfd-online.com/Tools/yplus.php.

[2] *Dimensionless velocity – CFD-Wiki, the free CFD reference*, [Online; accessed 2. Jul. 2020], May 2006. [Online]. Available: https://www.cfd-online.com/Wiki/Dimensionless_velocity.

[3] *Dimensionless wall distance (y plus) – CFD-Wiki, the free CFD reference*, [Online; accessed 2. Jul. 2020], Mar. 2014. [Online]. Available: https://www.cfd-online.com/Wiki/Dimensionless_wall_distance_(y_plus).

[4] *Fixed Point Iteration Method*, 2013. [Online]. Available: https://mat.iitm.ac.in/home/sryedida/public_html/caimna/transcendental/iteration%20methods/fixed-point/iteration.html (visited on 06/08/2020).

Table XIV
FORMULA SYMBOLS

Symbol	Name	Unit
ϕ	Azimuthal angle	rad
Re	Reynolds number	∅
Kn	Knudsen number	∅
ρ	Density	$\mathrm{kg\,m^{-3}}$
η	Dynamic viscosity	$\mathrm{kg\,m^{-1}\,s^{-1}}$
ν	Cinematic viscosity	$\mathrm{m^3\,s^{-1}}$
v	Velocity	$\mathrm{m\,s^{-1}}$
δ	Boundary layer thickness	m
N_i	Number of cells of the i^{th} mesh	∅
Φ_i	Result of the simulation with the i_{th} mesh	may vary
h	Grid size	∅
ΔV_i	Volume of the i^{th} cell	$\mathrm{m^3}$
r_{ij}	Grid size evolution between mesh i and j	∅
ϵ_{ij}	Result's deviation between mesh i and j	may vary
Φ_{ext}^{ij}	Extrapolated value from meshs iand j	may vary
e_a^{ij}	Approximate relative error	∅
$e_e^{ij}xt$	Extrapolated relative error	∅
C_p	Pressure recovery coefficient	∅
K_p	Total pressure los coefficient	∅

Table XV
INDICES

index	meaning
in	Inlet
out	Outlet
t	Total
a	Approximate
ext	Extrapolated

[5] *Flow of a fluid over a flat plate*, [Online; accessed 1. Jul. 2020], Jul. 2020. [Online]. Available: https://cdn.comsol.com/wordpress/2013/09/Flow-of-a-fluid-over-a-flat-plate.png.

[6] *Friction velocity – CFD-Wiki, the free CFD reference*, [Online; accessed 2. Jul. 2020], Dec. 2008. [Online]. Available: https://www.cfd-online.com/Wiki/Friction_velocity.

[7] N. Grün, "Numerische Strömungsmechanik - Computational Fluid Dynamics (CFD) Vorlesung Wintersemester 2019/20", Skript, Ostbayerische Technische Hochschule Regensburg, Regensburg, 2019.

[8] M. Heinrich, "Genetic algorithm optimization of the volute shape of a centrifugal compressor", PhD thesis, 2015.

[9] A. Inc., *ANSYS Free Student Software Downloads*, [Online; accessed 8. Jul. 2020], Jun. 2020. [Online]. Available: https://www.ansys.com/academic/free-student-products.

[10] Ismail B. Celik, Urmila Ghia, Patrick J. Roache, Christopher J. Freitas, Hugh Coleman, and Peter E. Raad, "Procedure for estimation and reporting of uncertainty due to discretization.", *journal of fluid engineering vol 130*, 2008.

[11] Karim Segond, Günther Zwarg, and Dr. Gunether Siegl, "Maillage", [Online]. Available: http://www.e-cooling.com/fr/maillage.htm.

[12] Martin Heinrich and Rüdiger Schwarze, "Genetic optimization of turbomachinery components using the volute of a transonic centrifugal compressor as a case study.", 2016.

[13] glowing python, "Fixed iteration point", 2012. [Online]. Available: https://glowingpython.blogspot.com/2012/01/fixed-point-iteration.html.

[14] *Wall shear stress – CFD-Wiki, the free CFD reference*, [Online; accessed 2. Jul. 2020], Dec. 2012. [Online]. Available: https://www.cfd-online.com/Wiki/Wall_shear_stress.

[15] *Which Turbulence Model Should I Choose for My CFD Application?*, [Online; accessed 1. Jul. 2020], Jul. 2020. [Online]. Available: https://www.comsol.eu/blogs/which-turbulence-model-should-choose-cfd-application/?setlang=1.

[16] Wikipedia, Ed., *Iteração de ponto fixo*, 2019. [Online]. Available: %5Curl%7Bhttps://pt.wikipedia.org/w/index.php?title=Itera%C3%A7%C3%A3o_de_ponto_fixo&oldid=56505720%7D (visited on).

SESSION B2

Understanding and Designing an Automotive-Like Secure Bootloader

Markus Schrötter B.Sc.
Ostbayerische Technische Hochschule Regensburg
Regensburg, Germany
Email: markus1.schroetter@st.oth-regensburg.de

***Abstract*—This paper describes the approach of designing a customized secure bootloader. Purpose will be the distribution of CTF-styled automotive security challenges onto a target platform for educational purposes. This is done by analyzing current practices, requirements and conditions of bootloaders in the real automotive industry in order to provide a more accurate picture of the real world. The most important steps in designing a bootloader, likely even secure software in general, are setting clear security goals, as well as carefully considering the general conditions introduced through the environment. While in our application the internal structure of the bootloader does not have to accurately represent reality, there are other factors that show the difficulty in designing such a concept. Especially poor hardware support and the fact that the challenges can be seen as a firmware that is insecure by-design.**

***Keywords*—Secure Bootloader, Security, Automotive, Security Concept**

I. Introduction

Everything in our modern society is getting more connected. Close to everybody uses a smartphone. In recent years it became popular to modify a house with many small computer-like devices - forming a “Smart Home”. A similar development is seen in the automotive industry. Cars nowadays can be seen as a big network of up to 100 or more small computers and they are even being connected with wireless interfaces to the internet. All of these devices have something in common: they run software. In many cases, software is not final when the hardware is being manufactured in the factory or when it is sold.

We have become used to updates. Usually feature update or security updates. While a cars’ features may be final before selling, security is very unlikely to be final. In general, there are are two security requirements that have to be achieved. “Requirements for the presence of desired behavior and requirements for the absence of undesired behavior”[1]. While it is quite easy to prove feature is existent simply by testing and demonstrating it, it can be quite hard to prove that a system doesn’t have any vulnerability and therefore is secure. Therefore it is necessary to keep a device flexible in terms of the software running.

This is where bootloaders are used. Figure 1 illustrates the flash memory layout of such a device. A bootloader is the first code that runs on a device after it is powered or reset. It initializes the hardware, e. g. the microcontroller, into a usable state and then jumps to the firmware in order to boot into it.

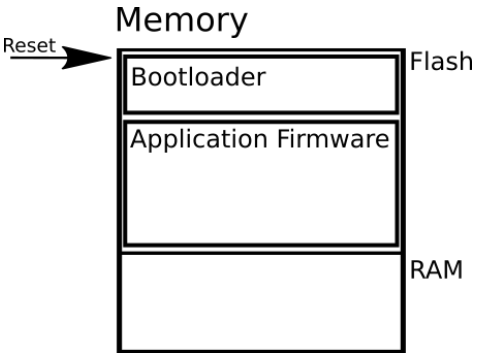


Fig. 1. Example bootloader memory layout.

It often comes with a firmware flash process, that allows for a new firmware to be loaded into the devices storage instead.

II. Bootloaders in Automotive

In the automotive industry, security is of special interest because of the requirements for functional safety. While many devices in peoples usual life use an update mechanism (e. g. phones, computers or IoT-Devices), most of the time a malfunction of these devices is not life-threatening. This is different with automotive software, which therefore has to fulfill many functional safety requirements and is tested very well. Nonetheless, even in the automotive industry, updates are sometimes necessary when the car is already in the hands of the consumer. Examples here would be the update for some diesel cars[2] or also sometimes non-critical feature-updates. Security vulnerabilities in cars may allow for threats found in the traditional computer world[3] and cases like the famous Jeep Cherokee-hack[4][5][6] show just how much functional safety depends on security. The difficulty of creating and implementing a flawless security concept, as argued in chapter I, makes an update mechanism necessary to mitigate dangers introduced through malicious intent or tampering with the device. There has to be long-term security, especially when considering the life-cycle of a car. While many day-to-day computer-devices are often just used for a limited amount of years, cars often have to be secured for several decades.

A. General requirements

The following section lists general goals of a bootloader in automotive when flashing a new firmware, based on a specialist book[7, Chapter 9.4]:

- Programming conditions
  Special conditions may have to apply while programming. This is specified by the memory chip manufacturer, examples can be: voltages, temperatures or timing constraints.
- Access controls
  The bootloader should only allow authorized entities to have access to the programming interfaces.
- Hardware-Software-compatibility
  The bootloader should check the compatibility of the new firmware to the hardware. Because of Original Equipment Manufacturer (OEM) internal structures there may be many versions of firmwares and Electrical Control Units (ECUs). Compatibility has to be ensured.
- Memory area checks
  A bootloader has to check different memory areas to decide whether a flash process can be executed. A fault condition could be that the target memory is not writeable.
- Robustness against programming errors
  The bootloader has to be secured from unintentionally being overwritten. Even in case of errors while programming the bootloader must stay operational to allow a recovery of the device.

B. Main Security Goals

A bootloader not only should provide an update mechanism for keeping the firmware running secure, but the process should also be secure itself. In the eyes of an attacker, a bootloader is an appealing target, since it contains a mechanism for changing the software running and hijacking that process may be a possible target for controlling the ECU.

In general, a secure bootloader provides two essential main properties that distinguish it from a normal bootloader. These are explained here related to the automotive industry:

- Confidentiality of the firmware
  The goal is to ensure confidentiality of secrets within the firmware. Different parts of a firmware can be seen as secret or not. In an open-source approach, often there are no secrets at all. Sometimes there are just small pieces to be kept secret (e. g. an authentication token). So it must be clear what has to be secured. In automotive, most of the time the OEM wants to protect his intellectual property and therefore has an interest for keeping the entire firmware secret. While updating the firmware of an ECU, it has to be transferred through several, possibly insecure channels. The tool to be used here is usually encryption.
- Integrity and authenticity of the firmware
  The goal is to ensure the integrity and authenticity of the software running. This is usually achieved through a process commonly known as “Secure Boot”. It describes a boot process, that only allows authorized software to run. It makes sure, that only firmware that was approved through the OEM is allowed. Usually signed fingerprints of the firmware that can only be produced by the OEM are used to achieve this. They can be generated by using public key cryptography. The software is verified every time the ECU boots. The background of this is to have a safe system. Running a modified firmware on a car’s ECU that is not tested and certified may impose danger to the driver and occupants, as explained before. It also hinders an attacker from easily hijacking the update process and injecting modified firmware, since he is not able to create a signed fingerprint.

While the automotive industry introduces many special attributes and conditions, it has still a lot of problems to be solved in common with other industries. By analyzing a sketch of a concept from the IoT-sector[8], the following (non-exhaustive) list of requirements should find consideration if applicable, when designing a secure bootloader:

- Key storage
  For cryptographic mechanisms within the bootloader to work, the secrets or keys have to be stored in a secure way. A hardware root of trust or secure key storage can be a good option. This is essential for fulfilling Kerckhoffs’ Principle[9].
- Revoke mechanism for signing keys
  In case private keys are leaked, a mechanism for revoking keys would mitigate possibilities to tamper with the device’s update mechanism.
- Rollback prevention
  It should not be possible to downgrade a firmware to a earlier version that might have security flaws.
- Readout protection
  Generally prevent possibilities to externally read data and extract data from the ECU. For example by disabling JTAG or encrypting external flash chips. This prevents common attack vectors.

These goals are supposed to provide a good starting point and orientation for designing a secure bootloader targeted at the automotive industry and are still open for discussion.

C. General conditions

Every software project depends on the different parameters that it is embedded in. Hardware capabilities, as well as requirements introduced through functional safety and general economic considerations may influence the requirements of a bootloader.

Besides the general circumstances concerning functional safety introduced at the start of chapter II, hardware capability is of importance. In the automotive industry in particular it is tried to keep the costs in production low because of the high quantity of cars. An ECU usually just barely fulfills the memory and performance requirements to fulfill its tasks since any unused performance reserve is likely to cost a lot more when

considering the quantity. In contrast, cryptography in general increases the resources required and microprocessors with special-purpose security hardware modules built-in are even more expensive. In order to have a secure system, hardware support must allow for it. While not only the microprocessor has to provide security features like a secure key storage or readout protection, it must also be secure. If the hardware-level security fails, software is likely to be vulnerable. An example of how it should not be is the STM32F1 microprocessor series. There it was proven that the readout protection can easily be broken[10].
It also has to be kept in mind that even hardware attacks are possible, since the attackers are most likely able to get their hands on the car and thus have physical access. Therefore it is essential that it is carefully evaluated whether the conditions allow for the security concept to be implemented and if it is secure despite the attackers possibilities.

D. Current State: Industry

Currently, in the automotive industry bootloaders are widely used. This is due to the advantages they provide[7]. Without one, every ECU would have to be programmed in the factory for many different cars and versions of cars. This creates a logistical complexity because of many slightly different ECUs. One of the most important advantages is that an ECU can be reprogrammed or updated more easily when it is necessary, e. g. at a repair shop.
Bootloaders in the automotive industry come in different kinds of flavors, but usually work in a similar way. As most of the software in automotive, they use a highly modular approach and fulfill different standards. This allows for reusability of the code. The ECU manufacturer can chose from several companies that provide proprietary bootloader solutions which can be adapted to different ECUs. Therefore it is often the case that the bootloader is provided by an external company.

E. Current State: Technical Aspects

In general, the bootloader is interfaced with over the car-internal network. Nowadays the Controller Area Network (CAN) bus system in combination with the ISO-TP[11] transport-protocol is very common to be used for communicating with the bootloader. The communication itself is using diagnostic protocols. One popular example is the Unified Diagnostic Services (UDS)[12] protocol. It is a client-server based protocol, where the vehicle repair shop tester (client) sends requests to the ECU (server), which then answers with responses. Functionality for authentication for using the interface and flashing the ECU is defined.
Due to the modularity and hardware properties, an automotive bootloader consists of several components. They are called Boot-Manager, Flash-Loader and Flash-Driver[7]. Figure 2 shows how they could be aligned in memory.

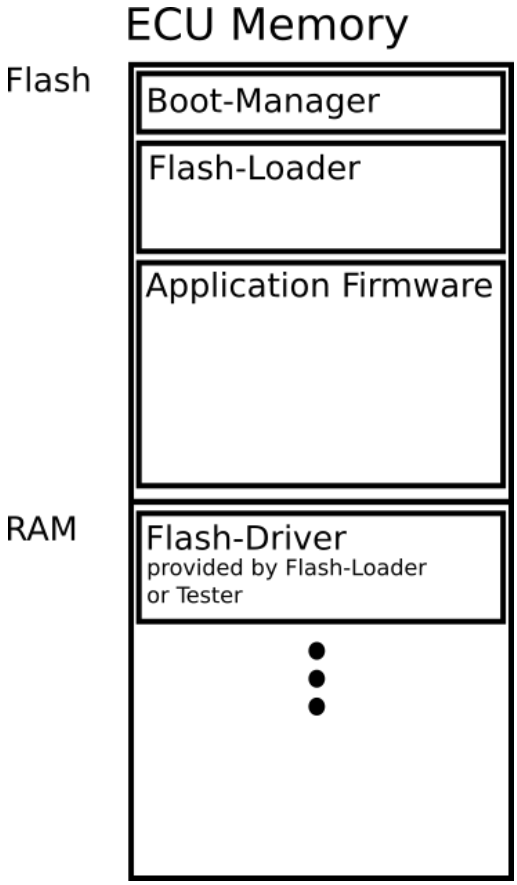


Fig. 2. Example automotive bootloader memory layout.

- Boot-Manager
  After powering on the ECU or a reset, the Boot-Manager is the first code to be executed. It decides whether the execution flow should continue towards the Flash-Loader or the Application Firmware. Criteria for this decision are whether there is a valid firmware or whether a switch into programming session has been requested.
- Flash-Loader
  The Flash-Loader contains the necessary code to control the execution flow while programming. It also must be able to communicate with the tester, therefore it contains an communication stack for the necessary protocols.
- Flash-Driver
  The Flash-Driver contains the actual programming algorithms in order to program the actual memory chips. These are usually routines for erasing memory and programming it. They can be already stored on the flash chip as part of the Flash-Loader or provided through the tester only when needed. Depending on the memory chip, different programming algorithms may be needed. As an example it could be a CPU-internal flash or an external Electrically Erasable Programmable Read-Only Memory (EEPROM). While reprogramming a memory chip, depending on the chip itself, it may be possible that other memory blocks of the same chip aren't accessible. Therefore the Flash-Driver is often copied to

RAM and run from there, since chips may not allow to be programmed and have code being run from at the same time.

This standardized approach also allows for addressing the in chapter II-B mentioned security goals.
In order to achieve confidentiality of the firmware, it is possible to specify an interchangeable encryption algorithm and transfer the firmware encrypted. A suitable and commonly used algorithm is called Advanced Encryption Standard (AES), as an example.
Integrity and authenticity of the firmware can be achieved on different security levels. A simple checksum can verify whether the programmed firmware is in its original state, but this can easily be manipulated. If security is needed, e. g. in case of a secure bootloader, cryptographic methodes like Hash-based Message Authentication Code (HMAC) and the Rivest–Shamir–Adleman (RSA) cryptosystem can be used.
In summary, the automotive industry uses a standardized approach that allows for mathematically secure concepts to be implemented within a bootloader. That said, it all depends on Kerckhoffs' Principle[9] and for the keys to be secret, as well as the bootloader running to not be manipulated.

III. Automotive-Like Bootloader for CTF-Challenges

In order to create a security concept, the conditions have to be carefully examined since different use-cases may require different approaches. In addition, every feature should be considered if it really is necessary, since the more complex a software becomes the more likely it is to contain security vulnerabilities.

A. Use-case

The final goal is a secure bootloader for education purposes, but it shouldn't be strictly limited to that. The idea is to have this bootloader be part of a course that provides students with realistic Capture-The-Flag (CTF) security challenges in the automotive context. In such a challenge, the student has to discover and exploit a purposefully implemented security vulnerability. If he succeeds, the student obtains a "flag", e. g. a random string of data, that proves he solved the challenge.
The secure bootloader itself will be used to distribute different challenges towards the hardware the challenges will run on. This will happen through automotive protocols in order to keep it realistic. The bootloader-interal structure does not have to be the same as in a real automotive bootloader, since the bootloader should just provide a proper environment and hacking the bootloader is not scope of the challenges. Nonetheless, the bootloader should be resistant to attacks.
Security purpose of the bootloader is to prevent any cheating while the challenges are being distributed. Since they will be flashed over CAN, students have access and could just sniff all challenges, extract the flag and therefore solve every challenge with the same exploit. Therefore, main goal is to protect the flag and to leak as little information about the challenge as possible. In addition only authorized code should be run to restrict any access further.

B. Conditions

The use-case and chip used introduce some special conditions that have to be considered when designing the bootloader. The chip used is the "ATSAME70N21B"[13]. It provides a few security-relevant features like a Memory Protection Unit (MPU), True Random Number Generator (TRNG), hardware AES and Secure Hash Algorithm (SHA) support.

- Insecure firmware by-design
  One special condition is that the firmware is insecure by-design due to it being a CTF challenge. This means the bootloader must not only provide security from outside attacks, but even from the firmware. Attackers are expected to even achieve arbitrary code-execution since it may be the challenges' goal. While it is intended to allow an extraction of the flag if the challenge is flashed, secrets within the bootloader, e. g. cryptographic keys, still have to be secured in order to not have the bootloader security be compromised.
- No hardware supported key management
  In order for the bootloader to provide security, the cryptograpic keys must not be leaked at any point in time. The microprocessor in use does not provide any hardware key management. Since the bootloader is the first code running, it has full control over the microprocessor and therefore it should be possible to configure the MPU to provide memory space that only the bootloader itself can access. In addition the MPU should be used to restrict access to hardware modules or memory spaces further that are not needed for the firmwares.
- Hardware access
  The challengees are provided the physical hardware and have access to the hardware not only when solving the challenge, but also when the challenge is distributed. This is important to consider because hardware attacks can be very hard to defend against. Such attacks can be side-channel and glitching attacks. While they are often of great effort they can often be quite successful. Nonetheless, the hardware is still managed and controlled by the providers of the challenges. If for example keys are leaked, they can be changed and possible countermeasures to attack may be implemented.

IV. Conclusion

Bootloaders in the automotive industry have some special properties due to the economy of the industry. Especially the interchangeability, standards and the fact that programming is very close to the embedded hardware is the reason to that. Due to the different requirements, it is reasonable to design the CTF bootloaders' internal architecture in a vastly different way to keep complexity and therefore the attack surface low. Nonetheless, from the outside it will seem like an automotive

bootloader since it implements the protocol properly. This paper shows that it is necessary to have clear security goals. Without the knowledge of what to protect, security measures may be inadequate. Also the general conditions are of importance. Possible attack vectors should be considered and the hardware support is important. If the latter is insufficient, it may not be possible to achieve the intended level of security at all. As for further work, an exhaustive attack surface analysis should be done in order to understand the threat fully.

REFERENCES

[1] H. Bidgoli, *Handbook of Information Security: Key Concepts, Infrastructure, Standards, and Protocols Volume 1*. Wiley, 2006, pp. 816–816.

[2] Bussgeldkatalog.org. (2020) Das vw-software-update im detail: Was bewirkt das update genau. Accessed: June 6, 2020. [Online]. Available: https://www.bussgeldkatalog.org/vw-software-update/

[3] N. Weiss, M. Schroetter, and R. Hackenberg, "On threat analysis and risk estimation of automotive ransomware," Oct. 2019.

[4] C. Miller and C. Valasek, "A survey of remote automotive attack surfaces," 2014.

[5] ——, "Remote exploitation of an unaltered passenger vehicle," 2015.

[6] ——, "Can message injection," 2016.

[7] W. Zimmermann and R. Schmidgall, *Bussysteme in der Fahrzeugtechnik: Protokolle, Standards und Softwarearchitektur*. Springer Vieweg, 2014.

[8] Cybergibbons. (2020) So what does an iot device need? Accessed: June 6, 2020. [Online]. Available: https://twitter.com/cybergibbons/status/1220846988020912130

[9] F. A. P. Petitcolas, *Kerckhoffs' Principle*. Boston, MA: Springer US, 2011, pp. 675–675. [Online]. Available: https://doi.org/10.1007/978-1-4419-5906-5_487

[10] M. Schink and J. Obermaier. (2020, March) Exception(al) failure - breaking the stm32f1 read-out protection. Accessed June 6, 2020. [Online]. Available: https://blog.zapb.de/stm32f1-exceptional-failure/

[11] *Road vehicles — Diagnostic communication over Controller Area Network (DoCAN) — Part 2: Transport protocol and network layer services*, Std. ISO 15 765-2, 2016.

[12] *Road vehicles - Unified diagnostic services (UDS) - Part 1: Specification and requirements*, Std. ISO 14 229-1, 2013.

[13] Microchip. Atsame70n21. Accessed June 11, 2020. [Online]. Available: https://www.microchip.com/wwwproducts/en/ATSAME70N21

Assembly and Investigation of a Compact Adsorption Heat Storage Module

Daniel Malzkorn*, Makram Mikhaeil and Belal Dawoud
Laboratory of Sorption Processes
Faculty of Mechanical Engineering
OTH Regensburg
Galgenberg Street 30, 93053 Regensburg, Germany
*Daniel.Malzkorn@oth-regensburg.de

***Abstract*—The EU-Horizon-2020 project "SWS-HEATING" aims at developing a compact solar-assisted heating system with advanced materials and components, to achieve a solar fraction over 60 % of heating demand (both space heating and DHW) of energy efficient single-family houses in central/north Europe. The core of the proposed energy system is an innovative, almost loss-free, multi-modular adsorption seasonal heat storage based on an innovative adsorption module employing a new sorbent material of the Selective Water Sorbent (SWS) family and heat exchangers (HEXs) to act as adsorber, evaporator and condenser. To this aim, asymmetric plate heat exchangers (ASPHEX) available form one of the SWS-Heating consortium partners shall be experimentally investigated as adsorber/desorber and evaporator/condenser heat exchangers in an adsorption storage module.**

The construction work, which has been carried out to adapt the available ASPHEXs to realize the world-wides most compact, lab-scale test unit of an adsorption storage module based on ASPHEXs in the Laboratory of Sorption Processes (LSP) of OTH Regensburg are presented. In addition, the adsorption-evaporation process has been experimentally investigated under typical operating conditions of adsorption heat conversion processes and the obtained results have been discussed.

A design review has been carried out to optimize the performance of the investigated ASPHEXs for application in the multi-modular seasonal adsorption storage. The design review results shall be realized by the consortiums partner and the optimized heat exchangers shall be tested in single as well as multi-modular adsorption storage units.

***Index Terms*—adsorption heat storage, plate heat exchangers, SWS, solar thermal energy**

I. Introduction

The depletion of the fossil fuel sources and the increasing concerns regarding the greenhouse gas (GHG) emissions and climate change are leading the world towards intensifying the utilization of renewable energy sources. Solar energy is one of the most promising alternative energy sources. However, due to the fluctuating and unstable nature, it does not match, most of the time, to the consumers energy demand. This divergence can be compensated by suitable energy storage technologies [1]–[3].

According to the EU-commission, the primary energy consumption in the EU household sector amounts to 25.7 % of the total energy consumption in the EU. In the households, the energy consumed for space and water heating amounts to 79.2 % of the total energy consumption. Until now, the EU-household sector depends by far on the traditional heating systems, which consume either fossil fuel or electrical power. In addition, an abundant amount of energy is lost every year in the different energy conservation processes worldwide [4]. About 63 % of this energy is classified as low-grade waste heat ($<100\,^{\circ}\mathrm{C}$), which cannot be utilized in electricity generation due to its low temperature.

Therefore, the interest in exploiting such low-grade waste heat as well as solar thermal energy for space heating has been recently increased considerably. Since the low-grade waste heat has mostly fluctuating nature as well, an effective heat storage technology, which allows long-term energy storage and compatible with the fluctuating and low-grade heat sources, could make a remarkable achievement in reducing the GHG-emissions in the household energy sector.

There are three main thermal energy storage techniques: sensible, latent and thermochemical [5]. In sensible storage, the stored heat is proportional to the temperature difference of the storage medium, i.e. the temperature difference between the charge and discharge phases. The storage density depends on both the temperature difference and specific heat of the used medium. The heat storage medium can be liquid or solid and the most common example is the water (in the liquid phase). The use of water as a sensible heat storage medium is favourable due to the cost-effectiveness, environment-friendly and good thermodynamics characteristics, in terms of the specific heat [6]. The main drawbacks of using water as sensible storage medium is the limitation of temperature working range ($0\,\mathrm{to}\,100\,^{\circ}\mathrm{C}$) and the high corrosion potential of the storage system components. Although the sensible heat storage system has simple designs, low cost, it is classified as a low heat storage density technology and suffers from high thermal losses.

In case of latent thermal storage, the large change in enthalpy during the phase change process at almost constant temperature of the storage material is exploited to store heat. The material are often mentioned as Phase Change Materials (PCMs) [7]. The phase change process can be a transition process between solid and liquid phase (i.e., melting and solidification process) or between two different crystal structures of the solid phase. Water beside salts are the most common inorganic materials used for the latent heat storage systems, where the storage is usually in the form of a transition between solid and liquid phase. Zalba et al. [8] have reviewed various aspects of latent heat storage systems, like PCMs heat transfer and applications. Latent heat storage systems have the advantage of operation in low temperature range and possesses a higher energy storage density in comparison with sensible heat storage.

The main advantage of latent heat storage is the ability to carry out charge and discharge processes without temperature fluctuations. However, the slow kinetics of the phase change, the instability of the PCM, and the material volume variation associated with the phase change of the material are the main drawbacks of this kind of heat storage. In addition, it is not very much suitable for long-term heat storage, due to the heat dissipation from the PCM to its surrounding.

Thermochemical energy storage is achieved by means of a reversible thermo-physical or thermos-chemical reaction between two components for storing thermal energy [9]. The heat storage systems associated with chemical reactions demonstrate high storage capacity compared with those associated with physical reactions. However, regarding the reaction kinetics, the physical reaction systems demonstrate superior performance. The sorption technologies involve physical reactions, which need generally lower charging temperatures compared to the chemical reactions. A sorption process involves a reaction between a refrigerant such as water or ammonia with a sorbent, which could be liquid (absorption system) or solid (adsorption system). Beside the high storage capacity, the almost zero-losses of the thermochemical energy storage system is the main advantage of using those systems for long-term, e.g. seasonal, heat storage.

Although absorption heat storage systems in e.g. Water-LiCl have received a lot of attention, they require relatively high charging temperatures, compared with adsorption systems. A greater risk of crystallization and pumping corrosive liquid are further drawbacks of such absorption systems [10]. In case of adsorption heat storage systems, these drawbacks do not exist. However, adsorption systems still have other challenges, which are presented in poor heat and mass transfer coefficients. More details about the advantage and disadvantage of the available heat storage technologies could be found elsewhere [5], [9].

It can be concluded that the compatibility of adsorption storage systems with the low-grade heat sources ($<100\,^{\circ}\mathrm{C}$) and the almost zero heat losses make it one of the most attractive heat storage technologies for exploiting solar thermal energy and the low-grade waste heat for space heating and cooling applications. The enhancement of the combined heat and mass transfer characteristics and the development of new adsorbent materials with higher adsorption capacities may lead to a remarkable improvement on the performance of such heat storage systems. Moreover, the system compactness, durability, reliability and fabrication cost should be taken into account when it comes to a commercial product.

The EU-project ”SWS-HEATING” aims, therefore, at developing an innovative seasonal thermal energy storage (STES) system based on the adsorption technology. A new system shall be developed with a novel storage material and a creative configuration, i.e., a sorbent material embedded in a compact multi-modular sorption STES unit. This will allow to store and shift the harvested solar energy available abundantly during summer to the less sunny and colder winter period, thus covering a large fraction of heating and domestic hot water demand in buildings. The targeted benefit of this next generation solar heating technology is to reach and overcome a solar fraction of 60 % in central/north Europe with a compact and high-performing STES system.

In this study, a compact single module, adsorption heat storage, with all components made of stainless steel and with a pair of a high-efficient asymmetric plate heat exchangers (ASPHEXs) developed from one of the SWS-Heating consortiums partners shall be assembled and pre-tested under relevant operating conditions of a real adsorption system. The setup module shall allow carrying out adsorption-evaporation and desorption-condensation processes representative to the processes carried out in the real adsorption heat conversion systems, like heat storage, heat pumps and chillers.

The applied ASPHEXs are identical and are not basically designed to act as adsorber/desorber. Therefore, a dedicated adaptation work should be carried out. The details of the adaptation and construction work, which has been carried out to adapt the applied ASPHEXs to produce the world-wides most compact, lab-scale test unit of a single-module, adsorption storage based on ASPHEXs in the Laboratory of Sorption Processes (LSP) of OTH Regensburg are presented. Based on the pre-test results, described and discussed in this work, the design review measures shall be derived, in order to design an optimized ASPHEX for application in the next developments of the multi-modular seasonal adsorption storage within the SWS-Heating project.

II. Assembly of the setup

A pair of identical closed-structure ASPHEXs produced by Alfa Laval, Sweden, was selected to assemble a compact single-modular adsorption heat storage system. Figure 1 depicts the configuration of the applied ASPHEXs. The ASPHEX is a stack of multi-nickel-brazed parallel plates made of stainless steel (SS). The brazing of the plates together forms two separated and non-identical domains. The ASPHEX are designed to exchange heat between two fluids (gas or liquid). Therefore, each domain shall be occupied by a flowing fluid, which enters the ASPHEX from an inlet port and leaves from the corresponding outlet port of that channel. Moreover, the design of the ASPHEX allows the draining of the condensed droplets in case of using the ASPHEX for cooling down a condensable gas.

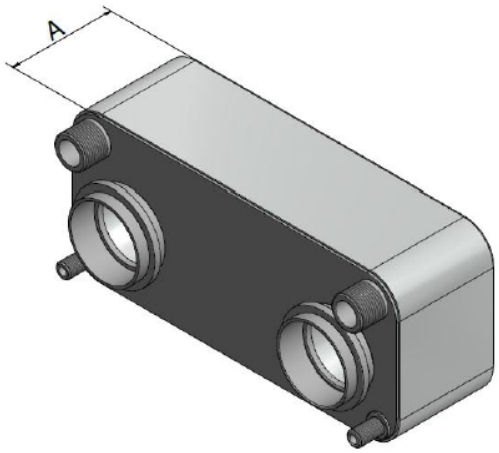
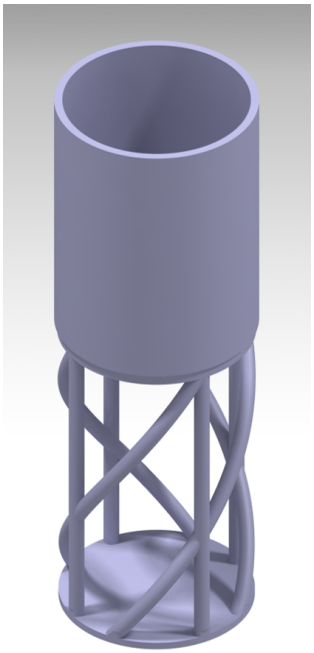


Fig. 1: ASPHEX Compact 26 with N-20 (A=83,0 [mm]) [11]

From the above description, it is clear that the applied ASPHEXs are not designed to act as adsorbers /desorbers. Therefore, some adaptation efforts are required to apply them as adsorber/desorber or evaporator/condenser in an adsorption storage module. The ASPHEX should enable the heat exchange between a HTF passing through one of the two separated domains of the ASPHEX and an adsorbent material occupying the other domain of the ASPHEX. As mentioned above, the two domains are not identical. The volume ratio of the domains is 1.66 and the bigger domain has the two big ports (see Fig. 1), which allows the application of such heat exchangers for heat recovery aspects in cooling down the exhaust gases of e.g. a combined heat and power system. From the Coefficient Of Performance (COP) point of view, it is favourable to select the smaller domain for the HTF and the larger one for the adsorbent material, because that leads to decrease the heat capacity ratio of the Adsorber (K_AdsHX) [12] and, accordingly, increase the COP. Another advantage of using the larger domain for the adsorbent is the



a) 4 co-rotating spiral with a diameter of 4 mm and a pitch of 210, 4 supports, base plate closed and hollow cylinder

b) 3-D printed cylindrical frame with the stainless steel sieve

Fig. 2: Cylindrical frames with 4 mm wall diameter and a pitch of 210 mm per 100 mm

larger ports (see Fig. 1), which can be utilized to connect the adsorber/desorber to the evaporator/condenser with lower mass transfer resistance.

The form of the adsorbent material plays also a significant role on the mass transfer resistance facing the vapour flow during the adsorption and desorption processes. Using the adsorbent in form of loose grains provides lower mass transfer resistance compared with using the adsorbent in a coated form. Indeed, the coating form demonstrated higher heat transfer rates between the adsorbent and the surfaces of the heat exchanger, however it suffers from poor mass transfer and high production costs. Several experimental studies [13]–[15] have demonstrated that adsorbent materials in form of loose grains can provide the same specific power of a coated adsorbent layer, if the right grain size is selected. In addition, with respect to stability, production cost and ease of production, the use of loose grain has many advantages [12].

In order to use adsorbent in form of loose grains without allowing the grains to fall out of the adsorbent domain into, e.g. the vapour manifolds (the two openings with the biggest diameter in Fig. 1) of the ASPHEX, a special construction allowing the vapour to pass through, while preventing the adsorbent grains from falling out has been designed.

Figure 2 depicts the design of this special construction. It is a cylindrical frame with a piece of fine stainless steel (SS)

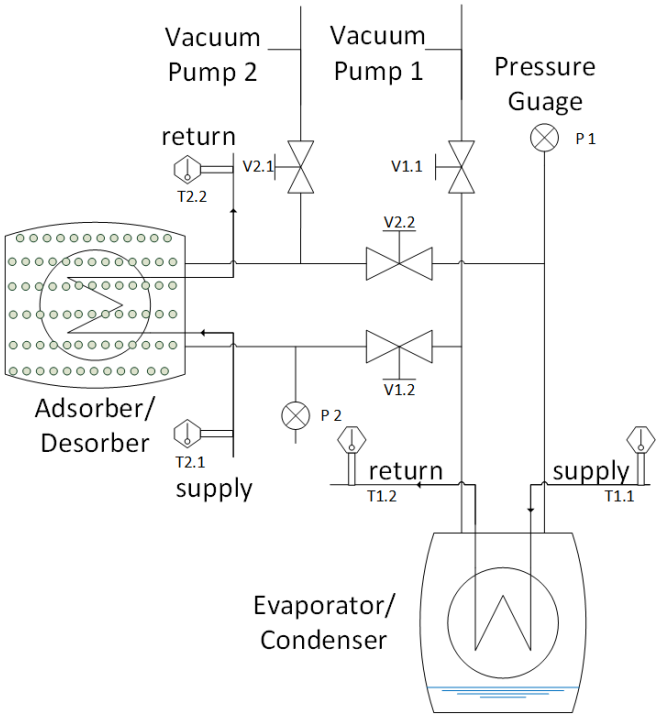


Fig. 3: Schematic layout of the experimental adsorption storage unit with the required sensors and actuators

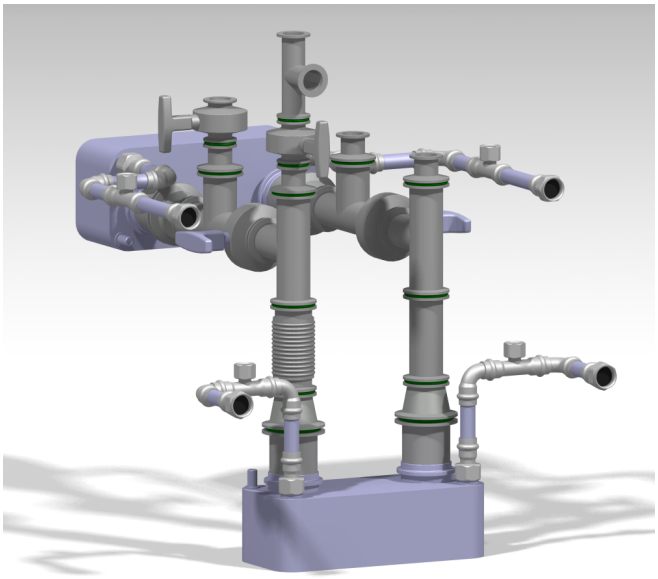


Fig. 4: Setup of the test rig with two ASPHEXs and the necessary vacuum components

sieve mounted annularly on the lower part, thus facing the different adsorbent domain channels. Two pieces of this special construction have been realized with the aid of the 3-D printing technology existing in the laboratory of Sorption Processes of OTH-Regensburg and mounted in the vapour manifolds of the ASPHEX. With appropriate stoppers, both frames have been mounted in the adsorber/desorber heat exchanger and prevented from any movement after the assembly of the whole storage module. To fill the adapted ASPHEX with the adsorbent grains, the two small ports for draining the condensed gases have been utilized. The Adapted ASPHEX has been mounted on a vibratory sieve shaker and filled in with the adsorbent grains. The adapted ASPHEX has been filled with 765 gram of dry silica-gel grains of type ”Siogel” produced by OKER CHEMIE, Germany, in the size range of 0.71-1.0 mm.

The layout depicted in Figure 3 illustrates a schematic for the assembled, single modular adsorption heat storage unit. The system consists, as depicted in Figure 3, mainly of two compartments.

The first is the top-left ASPHEX adapted to work as adsorber/desorber and the second is the second identical ASPHEX (without any adaptation) to work as an evaporator/condenser. The two ASPHEXs are connected together through two separated pipelines. The connection pipes allow the refrigerant vapour to transfer between the two ASPHEXs. Two vacuum gate-valves (V1.2 and V2.2) are mounted on the pipelines to allow separating the two ASPHEXs from each other. Two pressure transducers (P1 and P2) are mounted on the pipelines connecting the two ASPHEXs together (See Fig. 3) to measure the pressure of the refrigerant vapour in the two ASPHEXs separately, i.e., the vapour pressure inside the adsorber/desorber and the evaporator/condenser. Finally, two extra vacuum valves (V1.1 and V2.1) are mounted on the pipelines to allow evacuating the two ASPHEXs separately. In addition, the valve used to evacuate the evaporator/condenser (V1.1) is used also to fill it with the degassed refrigerant.

The pipelines and valves are all made of stainless steel. To prevent the undesired local vapour condensation on the inner surface of the connection piping and valves, especially during the condensation-desorption processes, a controlled heating cable is wrapped around them to keep their wall temperature higher than the condensation temperature. The set-up is well insulated to minimize the heat loss/gain to/from the surrounding. A leak test has been carried out with the helium leakage test unit to ensure the vacuum tightness of the whole assembly before starting the measurements. Figure 4 shows the 3D-drawing of the assembled components of the adsorption storage module depicted schematically in Figure 3.

A. Evaporator/condenser preparation

After evacuating the whole system, the two vacuum gate valves (V1.2 and V2.2) shall be closed to separate the evaporator/condenser from the adsorber/desorber. An external tank filled with degassed water, which shall be used as a refrigerant, is to be placed on a high-sensitive balance and be connected to the evaporator/condenser through valve (V1.1) mounted for evacuation and filling the evaporator /condenser with the refrigerant. A cold heat transfer fluid (HTF) at 5 °C, is allowed to pass through the evaporator/condenser to cool it down and enables condensation of the water vapour coming

out from the external tank. After accumulating 225 gram of the degassed water in the evaporator/condenser valve (V1.1) is to be closed. The amount of the water (refrigerant) has been determined based on the amount of the dry adsorbent in the adsorber/desorber and the planned testing conditions allowing a maximum water uptake of 29.4 g water /100g of dry adsorbent. The amount of water should be a little higher than the maximum amount of water, which could be adsorbed during any of the adsorption-evaporation processes (24 g/100g). A much higher amount of the refrigerant results in higher film thickness in the evaporator/condenser during the adsorption-evaporation and desorption-condensation processes, thus reducing the system dynamics by increasing the heat transfer resistance upon evaporation/condensation. On the other hand, a smaller refrigerant amount could lead to very much reducing the pressure at the end of the adsorption-evaporation process, thus reducing the adsorption dynamics and the overall performance of the adsorption unit.

B. Hydraulic Setup

In order for the experimental investigations of adsorption/evaporation and desorption/condensation operation phases of different adsorption heat transformation processes to be accrued out at controlled operating conditions, a dedicated hydraulic setup has been developed and mounted in the laboratory of Sorption Processes of OTH Regensburg. The hydraulic setup, which is depicted in Figure 5 , comprises two separated hydraulic circuits, a primary and a secondary circuit. The primary circuit (lower loop designated as HK) feeds the adsorber/desorber heat exchanger, whereas the secondary one (upper loop designated as NK) feeds the evaporator/condenser heat exchanger. A high precision control system has been established to allow realizing the desired temperature, pressure and flow rate of the HTF on each hydraulic circuit upon entering the respective heat exchanger.

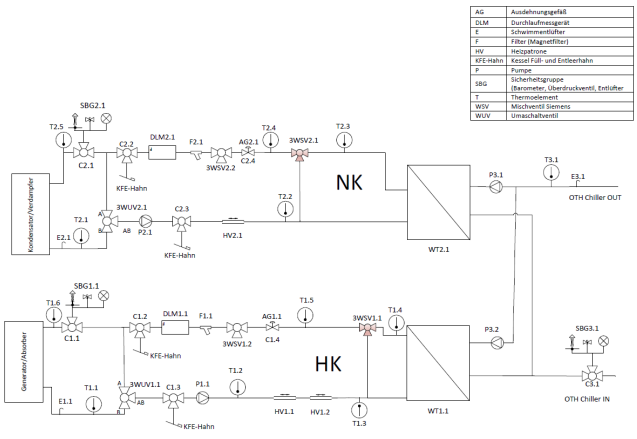


Fig. 5: Hydraulic setup [16]

In addition, the control system allows sudden falling and rising of the HTFs temperature on the primary circuit, enabling carrying out adsorption and desorption processes similar to the processes taking place in real adsorption heat pumps and chillers. Moreover, thanks to the two gate valves connecting the adsorber/desorber to the evaporator/condenser, adsorption and desorption processes similar to the processes taking place in adsorption storage systems can be experimentally investigated.

The above described setup allows conducting evaporation-adsorption and condensation-desorption processes according to the large temperature jump (LTJ) methodology described in more detail in [17], which corresponds to the processes taking place in adsorption heat pumps and chillers. Moreover, it allows the experimental investigation of adsorption units according to the large pressure jump (LPJ) method developed in [18], which replicates the processes taking place in adsorption storage and heat transformation systems.

C. Experimental procedure

In this study, adsorption-evaporation processes at different operating conditions corresponding to the LTJ methodology [17] have been conducted. Therefore, the gate valves connecting the adsorber/desorber to the evaporator/condenser are kept open during all conducted processes. Every adsorption-evaporation process comprised three successive phases. The first is the dehydration phase, during which the temperature of the HTF feeding the Adsorber/desorber heat exchanger is adjusted to $90\,^{\circ}\mathrm{C}$ for one hour.

The second is the preparation phase, which aims at realizing the adsorber/desorber equilibrium state condition for starting the quasi-isobaric adsorption process. This is done by adjusting the adsorber/desorber heat exchanger to the adsorption start temperature, calculated from the equilibrium vapour pressure diagram of the applied adsorbent-adsorbate pair (here Siogel-water) [19] and, in the same time adjusting the temperature of the evaporator/condenser heat exchanger to the required evaporator temperature of the tested adsorption process.

Details about the working principle of the thermally driven adsorption transformers could be found in Dawoud [20] as well as in Aristov et al. [21]. Using the equilibrium data of Siogel-water pair in [19], the starting adsorption temperatures and initial uptakes corresponding to evaporation temperature of $10\,^{\circ}\mathrm{C}$ and $15\,^{\circ}\mathrm{C}$, condensation temperature of $30\,^{\circ}\mathrm{C}$ and $35\,^{\circ}\mathrm{C}$ and driving heat source temperature of $90\,^{\circ}\mathrm{C}$ have been determined, (see Table I).

The end adsorption temperature is set equal to the condenser temperature or the respective process. This preparation phase takes 2 hours to ensure reaching the equilibrium state. The third test phase is the quasi isobaric adsorption phase, in which the temperature of the HTF feeding the adsorber PHE decreases rapidly to the end-temperature of the adsorption process. The LabVIEW code written to control the whole set-up allows to enter the desired end-temperature and realizing it at the inlet of the adsorber/desorber in about 2 minutes. The adsorption phase is measured over 2 hour as well to ensure reaching equilibrium conditions at the end of the process.

TABLE I: Applied operating conditions in the different experimental runs conducted in this work

$T_{Heat\,source}$, °C	T_{Evap}, °C	T_{Cond}, °C	$T_{ads-starting}$, °C	W_0, g/100g
90	10	30	65	4.22
		35	59	5.24
	15	30	72	4.15
		35	65	5.15

D. Power outputs and water uptake evaluations

The continuously stored readings from the volume flow rate sensors in each hydraulic circuit and temperature sensors installed at the inlet and the outlet of both evaporator/condenser and adsorber/desorber ($T_{1.1}$, $T_{1.2}$, $T_{2.1}$ and $T_{2.2}$ in Figure 3) are used to carry out the temporal energy balance of each unit. The uncertainties of the installed sensors are depicted in Table II. The outcomes of conducted energy balances are the instantaneous output power of the evaporator/condenser unit and the adsorber/desorber as described by the following equations (1) and (2).

$$\dot{Q}_{Evap} = \dot{m}_{HTF,Evap.} \cdot C_{pHFT} \cdot (T_{Evap,in} - T_{Evap,out}) \quad (1)$$

$$\dot{Q}_{Ads} = \dot{m}_{HTF,Ads} \cdot C_{pHFT} \cdot (T_{Ads,in} - T_{Ads,out}) \quad (2)$$

Where, $\dot{m}_{HTF,Evap.}$ and $\dot{m}_{HTF,Ads}$ are the mass flow rate of the HTF passing through the evaporator and adsorber, respectively. The HTF is water and its mass flow rate is set at 6 $\frac{\text{kg}}{\text{min}}$ in every heat exchanger. C_{pHFT} is the specific heat capacity of the HTF (4.2 $\frac{\text{kJ}}{\text{kg·K}}$). $T_{Evap,in}$,$T_{Evap,out}$, $T_{Ads,in}$ and $T_{Ads,out}$ are the inlet and outlet temperatures of the HTF passing through the evaporator and the adsorber, respectively.

As the vapour volume inside the system is small and the mass of the vapour existing inside it can be neglected, compared with the mass of the water liquid in the evaporator and the water adsorbed in the silica grains, the assumption of equality between the rate of water evaporation in the evaporator ($\dot{m}_{Evap.}$) and the rate of water adsorption in the adsorber ($\dot{m}_{Ads}$) is very reasonable. In addition, the maximum vapour pressure change inside the adsorption unit is below 2 mbar. The rate of water evaporation is calculated, accordingly, by equation (3).

$$\dot{m}_{Evap.} = \frac{\dot{Q}_{Evap.}}{\Delta h_{Evap.@T_{Evap.}}} \quad (3)$$

Where, $\Delta h_{Evap.@T_{Evap.}}$ is the latent heat of evaporation at a certain evaporator temperature ($T_{Evap.}$). The water uptake (w) is defined as the ratio of the weight of the absorbed water to the weight of the dry adsorbent and its instantaneous value could be calculated from:

$$w(t) = w_o + \frac{\int_0^t \dot{m}_{Ads} \cdot dt}{m_{Ads,dry}} \quad (4)$$

Where, $m_{Ads,dry}$ is the dry mass of the adsorbent and w_o is the initial water uptake (given in Table I).

TABLE II: Sensor used for measurements [22]–[25]

Sensor	Accuracy	Measured quantity
Balance KERN type EMB 6000-1	± 0.1g	Dry weight of adsorbent filling the adsorber/desorber
Pressure transducers PFEIFFER VACUUM type CMR 362	± 0.2 % of reading	Vapour pressure inside the adsorber/desorber and evaporator/condenser
RTD temperature sensors TMH type Pt100	1/10 DIN class B	HTFs temperatures at the inlet and outlet of both ASPHEXs
Flow meters SIEMENS type Sitrans F M MAG 100	$\leq 0.4\% \pm 1\frac{mm}{s}$	

The power due to the released heat of adsorption ($\dot{Q}_{Ads}(t)$) can be calculated from the following equation:

$$\dot{Q}_{Ads}(t) = m_{Ads,dry} \cdot \frac{dw}{dt} \cdot \Delta H_{is}(w) \quad (5)$$

Where, $H_{is}(w)$ is the isosteric heat of adsorption, which is almost constant at $w = 0.05 - 0.24\frac{kg}{kg}$ and equal to $(50.5 \pm 1.8)\frac{kJ}{mol}$.

III. Results and Discussion

A. Evaporator Power

Figure 6 presents the experimentally obtained evaporator power at the given operating conditions in Table I. Before starting each adsorption-evaporation process, the adsorbent grains, with their content of water (adsorbate), should be in a thermal equilibrium with the water vapour surrounding the grains. This is realized by the preparation phase over 2 h. To conduct an adsorption process as it takes place in real adsorption heat pumps, sudden cooling of the adsorbent martial is required. The hydraulic unit enables this temperature reduction over $\sim$ 100 to 140 seconds for the HTF temperature from the set adsorption start temperature ($T_{ad-start}$) to the set adsorption end temperature (T_{ad-end}). Once the adsorbent grains start to be cold down, the adsorption process is triggered and refrigerant vapour molecules starts to diffuse inside the pores of the grains. This interaction between the adsorbent and the refrigerant molecules is due to the well-known Van der Waals forces [26]. Consequently, a pressure difference between the vapour inside and outside the grains is developed , which keeps the adsorption process running until this driving force vanishes upon reaching the new equilibrium condition corresponding to the adsorption-end temperature (T_{ad-end}) and evaporator pressure.

The binding of the refrigerant molecules to the adsorbent surfaces results in releasing the so-called heat of adsorption ($H_{is}(w)$). The temperature of the adsorbent grains tends to increase and the ability of the grains to adsorb more refrigerant decreases. Therefore, effective cooling is required to ensure the continuity of the adsorption process. The higher the pressure of the adsorption-evaporation processes, the higher the evaporator power obtained. The blue and red curves in Figure 6 represent the evaporator power obtained at evaporator temperature 15 °C, which are higher than the evaporator powers obtained in case of applying evaporator temperature of 10 °C (curves represented in green and magenta colours). The evaporator power obtained in case of applying condenser temperature of 30 °C (T_{ad-end} = 30 °C) is, however, higher than the power obtained in case of condenser temperature of 35 °C (T_{ad-end} = 35 °C). This is attributed to the higher differential water uptake expected (and measured in Figure 7) at lower condenser temperatures.

In all cases, the evaporator power increases rapidly and reaches a maximum value (0.52 kW in case of T_{Ev} = 15 °C and T_{cond} = 30 °C) over the first few seconds. This is attributed to the existing high adsorption driving force, explained above, at the beginning of each adsorption process leading to higher adsorption rate and, accordingly, high evaporation rate. At the time of peak power, the rate of adsorption is at its maximum value. Afterwards, the adsorption rate decreases due to the decrease of the driving force (vapour pressure difference between inside and outside the adsorbent grains) and, consequently, the evaporator power decreases with a slow rate until reaching almost zero power after 3000 seconds.

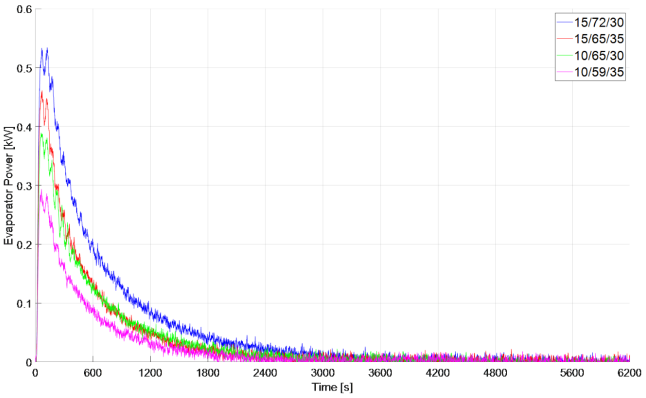


Fig. 6: Evaporator power obtained at the different applied operating conditions

B. Water uptake

Figure 7 illustrates a comparison between the water adsorption dynamics obtained at the different operating conditions. The maximum water uptake is obtained at evaporator temperature of 15 °C and condenser temperature of 30 °C, whereas the minimum value is at 15 °C and 35 °C evaporator and condenser temperature, respectively. This can be attributed to the increasing temperature left between evaporator and condenser from 15 to 25 K and the expected reduction of both evaporator power and adsorber performance. The initial water uptake at every operating conditions has been calculated from the equilibrium model of Siogel/water pair in [19] and the values are reported in Table I. The final water uptake obtained at the tested operating conditions has been checked against the values obtained from the equilibrium model, as a function in $T_{ads-end}$ and evaporator pressure (P_{Evap}). The maximum deviation obtained is 1.48 g/100g. The maximum uncertainty in estimating the water uptake out of the experimental investigations is estimated to be less than $\pm$0.42 g/100g.

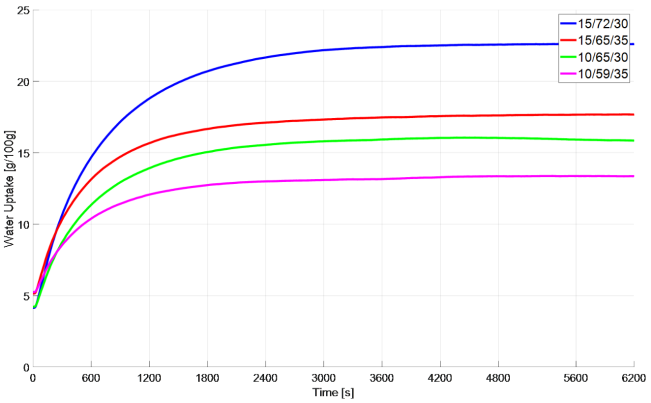


Fig. 7: Water uptake variation with the time obtained at the different operating conditions

C. Adsorber power

Figure 8 depicts the adsorber power variation with the time obtained experimentally at the different applied operating conditions. Since the large temperature jump method has been adopted for the all conducted experiments, the adsorber power increases very rapidly and reaches up to a maximum value (8.35 kW in case of T_{ev} = 15 °C and T_{cond} = 30 °C) during the first few seconds. This is attributed to the step change in the HTFs inlet temperature. At the time of peak power, the HTFs outlet temperature is still very close to the ASPHEXs initial temperature (72 °C in case of T_{ev} = 15 °C and T_{cond} = 30 °C), whereas the inlet temperature attains the adsorprion-end temperature (T_{ad-end} = 30 °C). Afterwards, the power output decreases rapidly and gets lower than 2 kW after 60 seconds. After that time, the adsorber power continues to decrease, but with a clearly slower and slightly fluctuated rate until it reaches almost zero power after 1800 seconds. The slight fluctuation in the adsorber power is attributed to the fluctuation in the HTF inlet temperature, as depicted in Figure 9, which depicts the HTFs inlet and outlet temperature variations with the time for the experimental run with T_{ev} = 15 °C and T_{cond} = 30 °C.

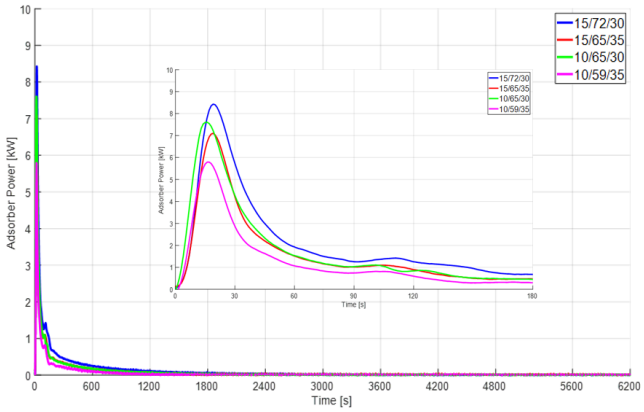


Fig. 8: Adsorber power variation with the time obtained at the different operating conditions

In order to evaluate the relative contributions of both sensible and adsorption heat on the adsorber power, the power due to only the heat of adsorption has been calculated from equation (5) and subtracted from the adsorber power, in order to estimate the contribution of sensible heat stored in the adsorber components on the total adsorber power. The results obtained for all applied operating conditions are illustrated in Figure 10, in which the adsorber power including both contributions (due to the release of the heat of adsorption and that due to the sensible heat stored) is represented by the solid blue lines, and the power due to only the release of the adsorption heat is represented by the solid red lines. The dashed red lines represent the resulted power due to the sensible heat stored in the ASPHEXs metal, HTF and the dry adsorbent. The power due to the sensible heat vanishes almost

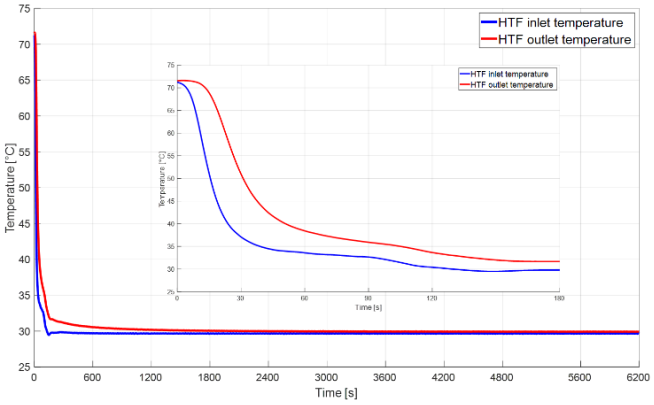
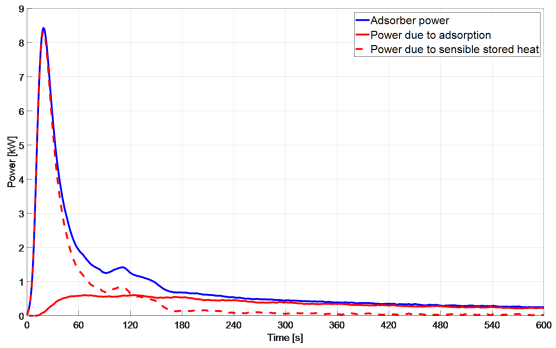


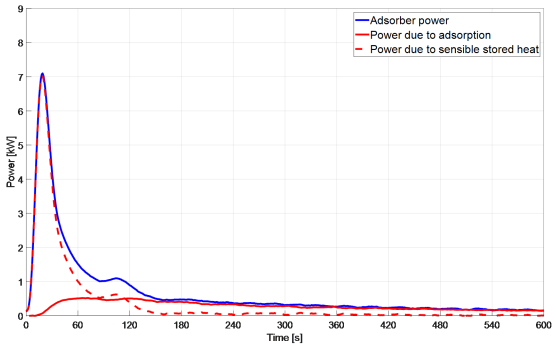
Fig. 9: HTFs inlet and outlet temperature variation with the time (in case of T_{ev}=15 °C, T_{cond}=30 °C and T_{source}=90 °C)

completely during the first 2.5 to 5 minutes of the process start depending on the boundary condition. The cross over between both contributions (heat of adsorption and sensible heat stored) takes place after 20 seconds from the beginning of the adsorption process and, from there on, the power due to the heat of adsorption dominates the process dynamics. The lower the sensible heat stored in the metal and the HTF of the adsorber HX compared to the heat of adsorption, the higher the COP of the adsorption system. From the system specific power (SP) point of view, this sensible heat has to be rapidly transferred to the HTF to allow a rapid and effective cooling of the adsorbent material and, consequently, a fast adsorption process leading to a high systems SP.

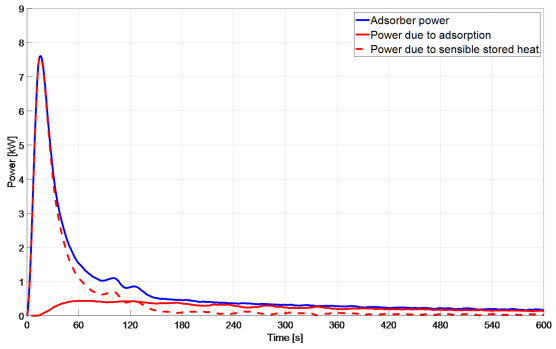
From the above analysis, it is clear that reducing the metal mass of ASPHEX and minimizing the domain of the HTF inside it shall result in decreasing the sensible heat stored in the adsorber heat exchanger and, consequently, improve the system COP and SP. This can be done by very much reducing the thickness of both endplates of the PHE (in the current design 4 mm, leading to a 30 % mass contribution to the total mass of the HEX. In addition, the degree of asymmetry between the two domains (volume of the adsorbent domain to the volume of the HTF domain) must be increased, from the current value of 1.66 for the investigated PHEX to at least 3 in the design reviewed design. Indeed, the tests have been carried out with Siogel-water as a working pair, which is not the most effective adsorbent in the market but could be acquired faster as it offers a limited differential water uptake leading to a small contribution by the stored heat of adsorption. The main target of this work was to offer the proof of concept of applying plate heat exchangers in the adsorption field for the first time worldwide and most importantly to derive the design review measures for the next generation ASPHEX, which is now under development for the SWS-heating project. In parallel, the target for the differential water uptake is three times that of Siogel, which has been proven in a parallel experimental campaign at OTH for the materials developed by the SWS-Heating project partners.



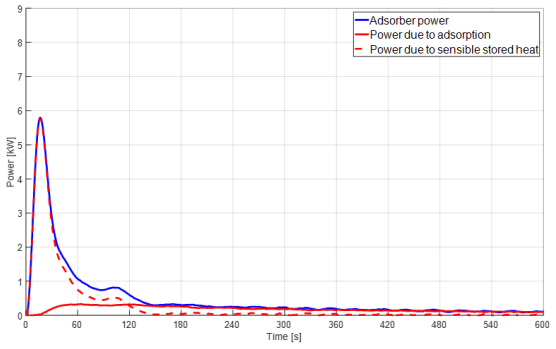
a) T_{ev}=15 °C, T_{cond}=30 °C *and* T_{source}=90 °C



b) T_{ev}=15 °C, T_{cond}=35 °C *and* T_{source}=90 °C



c) T_{ev}=10 °C, T_{cond}=30 °C *and* T_{source}=90 °C



d) T_{ev}=10 °C, T_{cond}=35 °C *and* T_{source}=90 °C

Fig. 10: Heat of adsorption and sensible heat stored in the adsorber contributions on the adsorber overall power

IV. Conclusion

A compact single adsorption heat storage module, with all components are made of stainless steel and with a pair of a high-efficient asymmetric plate heat exchangers (ASPHEXs) developed from one of the SWS-HEATING consortium partners is assembled. The setup module allows carrying out adsorption-evaporation and desorption-condensation processes representative to the processes carried out in the real adsorption storage systems, heat pumps and chillers. Experimental investigations under relevant operating conditions of a real adsorption heat pump are carried out. The study comes out with a recommendation to reduce the volume of the heat transfer fluid (HTF) domain of the adsorber/desorber heat exchanger with taking care of the influence on the heat transfer between the heat exchangers metal surfaces and the HTF. In addition, the thickness of the end plates of the plate heat exchanger shall be very much reduced.

References

[1] L. A. Chidambaram, A. S. Ramana, G. Kamaraj, and R. Velraj, "Review of solar cooling methods and thermal storage options," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 6, pp. 3220–3228, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.rser.2011.04.018

[2] I. Dincer and M. A. Rosen, "Thermal Energy Storage (TES)," *Thermal Energy Storage: Systems and Applications*, p. 93, 2002.

[3] Harald Mehling; Luisa F. Cabeza, *Heat and cold storage with PCM*, 2000, vol. 11, no. 3.

[4] "EU-commission, total energy consumption." [Online]. Available: https://www.iea.org/regions/europe

[5] L. F. Cabeza, I. Martorell, L. Miró, A. I. Fernández, and C. Barreneche, "1 - Introduction to thermal energy storage (TES) systems," in *Advances in Thermal Energy Storage Systems*, ser. Woodhead Publishing Series in Energy, L. F. Cabeza, Ed. Woodhead Publishing, 2015, pp. 1–28. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781782420880500018

[6] S. Vasta, V. Brancato, D. La Rosa, V. Palomba, G. Restuccia, A. Sapienza, and A. Frazzica, "Adsorption heat storage: State-of-the-art and future perspectives," *Nanomaterials*, vol. 8, no. 7, 2018.

[7] P. Pinel, C. A. Cruickshank, I. Beausoleil-Morrison, and A. Wills, "A review of available methods for seasonal storage of solar thermal energy in residential applications," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 7, pp. 3341–3359, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.rser.2011.04.013

[8] B. Zalba, J. M. Marn, L. F. Cabeza, and H. Mehling, "Review on thermal energy storage with phase change: materials, heat transfer analysis and applications," *Applied Thermal Engineering*, vol. 23, no. 3, pp. 251–283, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359431102001928

[9] T. M. Letcher, R. Law, and D. Reay, *Storing energy: with special reference to renewable energy sources*. Elsevier Oxford, 2016, vol. 86.

[10] S. D. Waszkiewicz, M. J. Tierney, and H. S. Scott, "Development of coated, annular fins for adsorption chillers," *Applied Thermal Engineering*, vol. 29, no. 11-12, pp. 2222–2227, 2009.

[11] Airec, "Compact-26_2016-12.pdf," p. 1, 2016.

[12] A. Freni, B. Dawoud, L. Bonaccorsi, S. Chmielewski, A. Frazzica, L. Calabrese, and G. Restuccia, *Characterization of Zeolite-Based Coatings for Adsorption Heat Pumps*, 2015. [Online]. Available: http://link.springer.com/10.1007/978-3-319-09327-7

[13] A. Freni, F. Russo, S. Vasta, M. Tokarev, Y. I. Aristov, and G. Restuccia, "An advanced solid sorption chiller using SWS-1L," *Applied Thermal Engineering*, vol. 27, no. 13, pp. 2200–2204, 2007.

[14] S. Santamaria, A. Sapienza, A. Frazzica, A. Freni, I. S. Girnik, and Y. I. Aristov, "Water adsorption dynamics on representative pieces of real adsorbers for adsorptive chillers," *Applied Energy*, vol. 134, pp. 11–19, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.apenergy.2014.07.053

[15] L. X. Gong, R. Z. Wang, Z. Z. Xia, and C. J. Chen, "Design and performance prediction of a new generation adsorption chiller using composite adsorbent," *Energy Conversion and Management*, vol. 52, no. 6, pp. 2345–2350, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.enconman.2010.12.036

[16] G. Dupont, "MAPR 2nd Semester Report Gabriel Dupont.pdf," p. 42, 2019.

[17] Y. I. Aristov, B. Dawoud, I. S. Glaznev, and A. Elyas, "A new methodology of studying the dynamics of water sorption/desorption under real operating conditions of adsorption heat pumps: Experiment," *International Journal of Heat and Mass Transfer*, vol. 51, no. 19-20, pp. 4966–4972, 2008.

[18] B. Dawoud and Y. Aristov, "Experimental study on the kinetics of water vapor sorption on selective water sorbents, silica gel and alumina under typical operating conditions of sorption heat pumps," *International Journal of Heat and Mass Transfer*, vol. 46, no. 2, pp. 273–281, 2003.

[19] A. Sapienza, A. Velte, I. Girnik, A. Frazzica, G. Füldner, L. Schnabel, and Y. Aristov, " Water - Silica Siogel working pair for adsorption chillers : Adsorption equilibrium and dynamics," *Renewable Energy*, vol. 110, pp. 40–46, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.renene.2016.09.065

[20] B. Dawoud, "Water vapor adsorption kinetics on small and full scale zeolite coated adsorbers; A comparison," in *Applied Thermal Engineering*, vol. 50, no. 2. Elsevier Ltd, 2013, pp. 1645–1651. [Online]. Available: http://dx.doi.org/10.1016/j.applthermaleng.2011.07.013

[21] Y. I. Aristov, I. S. Glaznev, and I. S. Girnik, "Optimization of adsorption dynamics in adsorptive chillers: Loose grains configuration," *Energy*, vol. 46, no. 1, pp. 484–492, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.energy.2012.08.001

[22] Siemens, "SITRANS F M MAG 1100," 2019.

[23] Pfeiffer Vacuum, "PTR24611.en.pdf," 2019. [Online]. Available: https://www.pfeiffer-vacuum.com/productPdfs/PTR24611.en.pdf

[24] THM Temperatur Messtechnik, "Mantelwiderstandsthermometer-000.pdf," 2019. [Online]. Available: https://www.temperaturmesstechnik.de/fileadmin/user_upload/pdf/tmh-mantelwiderstandsthermometer-000.pdf

[25] K. &. S. Gmbh, "Betriebsanleitung Präzisionswaagen," pp. 1–18, 2016. [Online]. Available: https://dok.kern-sohn.com/manuals/files/German/EMB-BA-d-1636.pdf

[26] D. M. Ruthven, *Principles of adsorption and adsorption processes*. John Wiley & Sons, 1984.

Development of a High Pressure Constant Volume Combustion Chamber for Investigation on Flammability Limits

Jan Triebkorn
Faculty of Electrical Engineering and Information Technology
Laboratory for Combustion Engines and Emission Control
OTH Regensburg
Regensburg, 93053 Germany
Email: jan.triebkorn@st.oth-regensburg.de

***Abstract*—The focus of this research work is to develop a constant volume combustion chamber (CVCC) which can be used to investigate the laminar burning velocity and thereby the flammability limits of gaseous mixtures of combustible gas, oxidizer and diluent. In this work specifically hydrogen and oxygen are uses as a combustible gas and an oxidizer, respectively. Argon and water (steam) are used as diluents. The chamber is designed to simulate the condition of the gas mixture inside of a combustion engine. Initial conditions prior to an ignition can be up to $70\,\mathrm{bar}$ and $450\,^{\circ}\mathrm{C}$. This work describes the approach designing the CVCC with regard to the boundary conditions, the material choice and the geometry. For this purpose a structural and thermal analysis have been made using the finite element method. To be able to precisely reach a certain gas composition inside the chamber three methods are used (Dalton's law of partial pressures, mass flow controllers and mass spectrometry). To calculate the laminar burning velocity the increase in pressure is measured and evaluated.**

***Keywords*—Constant volume combustion chamber; Laminar burning velocity, Flammability limits.**

I. Introduction

As part of the research project QUAREE100 a one cylinder combustion engine is in development. This engine uses a mixture of hydrogen, oxygen, argon and water as a working gas in a closed cycle process and thereby has zero CO_2 emissions. Prior to the initial operation of this engine fundamental information about the characteristics of the combustion process of this hydrogen-enriched gas mixture is required. The laminar burning velocity is an important characteristic of the reactivity of combustible mixtures and allows to extract even more basic flame properties. [1], [2]

While the combustion of mixtures containing hydrogen has been repeatedly investigated, the research was mainly limited to initial temperatures and pressures close to atmospheric conditions (as reported by [3]). In the last decade more investigations at elevated temperatures and pressures were made, but compositions with steam (water) and argon as diluents were not tested. As the described engine uses a mixture consisting of H_2, O_2, Ar and H_2O, there is no entirely comparable information about the combustion characteristics of such a mixture available.

As it is not recommended to obtain such experimental data using and possibly damaging an engine a constant volume combustion chamber (CVCC), which can easily be modified and is comparatively inexpensive is an optimal solution. It is designed to simulate the gas conditions at top dead center of the engine (at the end of the compression stroke).

In different literature aside from calculating the laminar burning velocity from the pressure rise at the initial stage of flame propagation often optical methods are used as well. However this study focuses only on the first method as good results are achievable. [2]

II. Measurement Technology

A. Test Bench Setup and Measurement Instruments

The general setup of the test bench with the CVCC, the gas supply and measurement instruments is illustrated in Figure 1. The chamber uses the following measurement instruments:

- piezo-electric pressure transducer with charge amplifier, $0\,\mathrm{bar}$ to $250\,\mathrm{bar}$, (Kistler 6052C and 5018A),
- piezo-resistive pressure transducer with amplifier, $0\,\mathrm{bar}$ to $100\,\mathrm{bar}$ (Kistler 4011A100 and 4624AK),
- $1.5\,\mathrm{mm}$ NiCr−Ni thermocouple (type K),
- BOSCH GDI injector for water injection,
- mass flow controller (Bronkhorst type FG-201AV),
- mass spectrometer (V&F EISense) to analyse gas composition and
- real-time target machine (Speedgoat) with programmable FPGA (8 analog I/O's, 200 ksps) for data acquisition and pressure indicating.

To ensure that a certain gas composition can precisly be achieved three different measurement methods will be used. The combustion chamber will be filled step by step based on the proposed procedure in DIN 6146 [4] and using Dalton's law of partial pressures and a piezo-resistive pressure transducer. Because at high temperature and high pressure especially in presence of steam (H_2O) the gas mixture can not be considered as an ideal gas, therefore the compressibility factor Z (calculated with REFPROP using the NIST

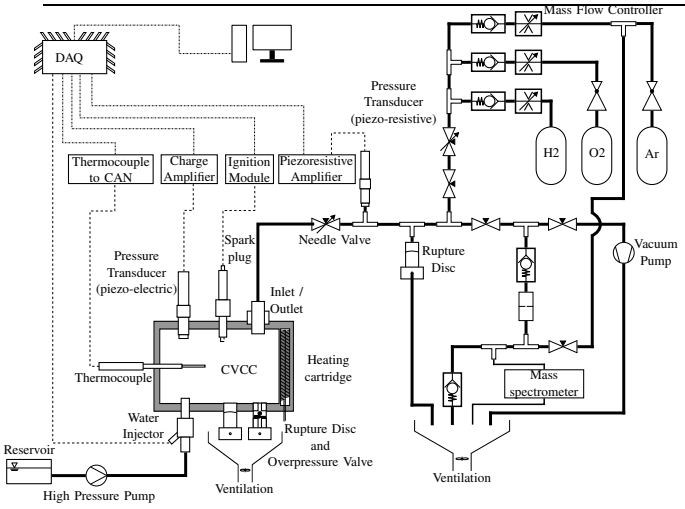


Figure 1. Schematic design of the test bench consisting of the CVCC, the gas supply and measurement equipment.

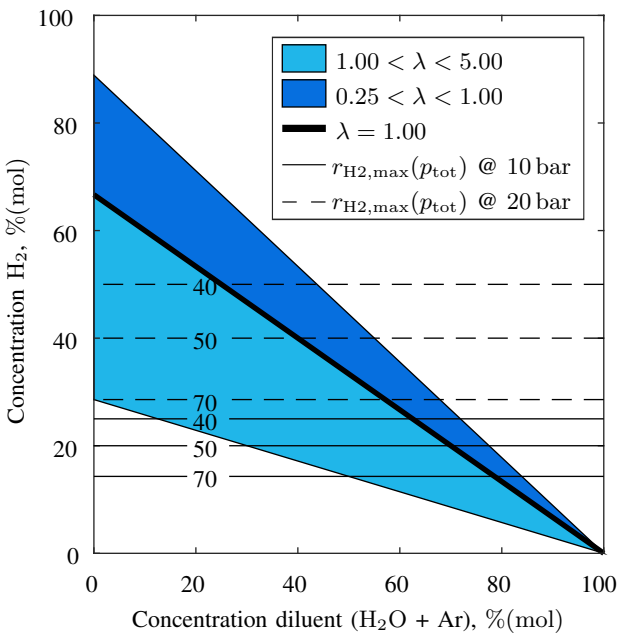


Figure 2. Graph illustrating lean and rich air-fuel ratios (λ) and limitations for the gas composition due to the available hydrogen gas supply.

database [5]) needs to be taken into account. In addition mass flow controllers are used to measure the amount of gas filled into the chamber for every single gas. After the filling process a small amount of the mixture can be analysed by a mass spectrometer to verify the gas composition.

To control the filling and evacuation process and for the data acquisition a real time machine (speedgoat) is used. After the ignition the pressure is measured by a piezo-electric pressure transducer. The increase in pressure at the initial stage of combustion needs to be recorded at very high frequency (up to 100 kHz) and resolution (16 Bit), therefore a FPGA is used similar to typical engine indication systems.

Heating cartridges are used to heat the CVCC and thereby the gas mixture. Multiple thermocouples are used to monitor the temperature of the case and the gas mixture.

B. Gas Composition and Limitations

While oxygen and argon can both be provided by 200 bar gas cylinders, the hydrogen supply is the current facility is limited to 10 bar. Figure 2 illustrates lean and rich gas compositions targeted in this work. In addition the maximum possible amount of H_2 is shown for a given total pressure (40, 50 and 70 bar respectively) depending on the maximum partial pressure of H_2 (10 and 20 bar).

While one study [6] reported that stoichiometric mixtures with less than 50 % diluent (steam) and at high pressures are susceptible to detonate even leaner mixtures ($\lambda > 2$) could only be tested down to a diluent level of about 70 % (at a total chamber pressure of 70 bar). Table I shows some limitations for the gas composition for an exemplary chamber pressure of 70 bar and $\lambda = 1$ due to the available hydrogen gas supply (10 bar) and the potential benefits of a higher available pressure (e. g. 20 bar).

C. Calculation of Laminar Burning Velocity

Typically two methods are used to calculate the burning velocity. This is on the one hand by optical measurements

Table I
LIMITATIONS FOR THE GAS COMPOSITION OF THE MIXTURE DUE TO LIMITATIONS ON THE HYDROGEN GAS SUPPLY EVALUATED FOR A TOTAL CHAMBER PRESSURE OF 70 bar AND A STOICHIOMETRIC RATIO $\lambda = 1$.

H_2-supply	minimal diluent concentration	maximal hydrogen concentration
$p_{H2,max}$	$r_{diluent,min}$	$r_{H2,max}$
10 bar	78.57 %	14.29 %
20 bar	57.14 %	28.57 %

using high-speed cameras and on the other hand by measuring the increase in pressure inside the chamber. In this study only the latter is used as multiple studies (e. g. [2] and [7]) reported good results for both methods.

The pressure method was proposed by Lewis and Elbe [8] and developed by Andrews and Bradley [9]. While in the past it was difficult to attain sufficient sensitivity and accuracy from a pressure transducer due to advancements in technology today the pressure can be measured very precisely and thus this method presents a relatively simple technique to obtain information about the flame propagation. [1]

This method takes the adiabatic change of pressure and temperature of the unburned gas into account. The laminar burning velocity S_L can be evaluated by: [2]

$$S_L = \frac{S_S}{\sigma}\left(1 + \frac{1}{\gamma_b}\frac{B_2 r_b^3}{(S_S^3 p_0 + B_2 r_b^3)}\right), \qquad (1)$$

where $\sigma = \rho_u/\rho_b$ is the expansion rate of burned mixture, r_b is the burned gas radius, γ_u and γ_b are specific heat ratios for the unburned and burned gas and p_0 is the initial pressure. B_2 is the polynomial coefficient of the correlation of the

experimental pressure-time history: $p(t) = p_0 + B_2 \cdot t^3$. S_S is the visual flame velocity given by:

$$S_\mathrm{S} = \left(\frac{B_2}{p_0} \left(\frac{\gamma_\mathrm{b} R^3 + \frac{\gamma_\mathrm{u}}{\sigma} r_\mathrm{b}^3 - \gamma_\mathrm{b} r_\mathrm{b}^3}{(1 - \frac{1}{\sigma})\gamma_\mathrm{b}\gamma_\mathrm{u}} - r_\mathrm{b}^3 \right) \right)^{1/3}, \qquad (2)$$

where R is the internal radius of the chamber. Initially this method is designed for spherical chambers and an ignition at the centre. However this method is still expected to be sufficient to obtain qualitatively information about the laminar burning velocity of different gas compositions.

III. Dimensioning and Design of the CVCC

The chamber material and the general design of the CVCC has been chosen and designed based on the following boundary conditions:

- initial pressure: up to 70 bar,
- initial temperature: up to 450 °C ,
- small volume and
- high corrosion resistance.

A. Material Selection

As high corrosion resistance is required different stainless steels are considered. At high temperatures austenitic steels are susceptible to hydrogen atoms diffusing into the steel lattice, accumulating and forming CH_4 (methane) which leads to a pressure buildup and results in blister and hydrogen embrittlement. However AISI 316Ti has high amounts of chromium (16.5 % to 18.5 %) and nickel (10.5 % to 13.5 %) which imparts good resistance to hydrogen attack. [10]

Figure 3 shows the yield strength (0.2 % offset) over a broad temperature range. It can be seen that the yield strength decreases significantly for a temperature around 450 °C. While Inconel 601 has a higher yield strength of 330 MPa at 550 °C it is much more expensive due to the higher amount if nickel ($\sim$ 58 % compared to $\sim$ 12 %). [11] Therefore 316Ti is selected for the chamber.

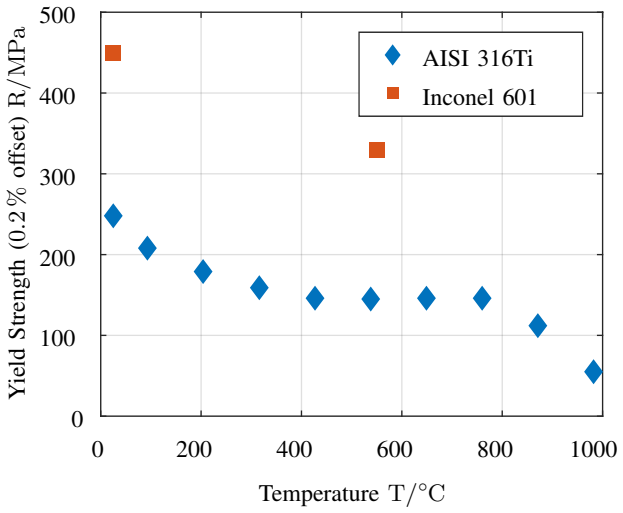


Figure 3. Yield strength (0.2 % offset) of two different alloys at elevated temperatures. [12], [11]

B. Geometry of the CVCC

The next step in the design procedure is the geometry. Regarding the volume of the chamber multiple factors are considered. Because CVCCs are often used to analyse the spray pattern of fuel injectors at varying pressures, the maximum spray penetration and maximum width usually need to be taken into account. However this is not an intended usecase for this CVCC. Injected water hitting the wall is not a concern in this situation as it evaporates quickly. The targeted experimental data in this study is quite fundamental, thus there is no need to take the actual geometry of the one cylinder engine into account. While typically CVCCs with an internal radius of 70 mm to 250 mm are used because of before mentioned reasons, in this case a smaller volume is more desirably.

The small volume comes with many advantages, less gas is needed to reach a certain pressure, which not only leads to a shorter time to fill, heat up and evacuate, but also means the energy from the combustion is considerably lower.

Typical shapes are cylinders and spheres, for easier manufacturing a cylinder is used. The chamber needs multiple ports for the gas inlet and outlet, spark plug, injector, thermocouple, pressure transducer, overpressure valve and rupture disk. To accommodate all those parts, which will mostly be placed lateral, a height of 20 mm and a diameter of 40 mm is suitable. This results in a chamber volume of 25 cm^3.

The gas inlet will be placed tangential to ensure a good mixing of the gases as demonstrated in [13]. This eliminates the need of a mixing fan which makes the design even more complex (as shown in [14]).

C. Structural Analysis

1) Calculation of Load: To approximate possible peak pressure while combustion Cantera is used with the GRI-Mech 3.0 reaction mechanism [15]. This detailed mechanism is widely used to model natural gas combustion.

As the partial pressure of hydrogen is currently limited to 10 bar the combustion of a stoichiometric test mixture of 10 bar H_2, 5 bar O_2 and 35 bar Ar was simulated. The total pressure prior to an ignition without taking non-ideality into account is 50 bar and initial temperature is 573 K. This results in a peak pressure of 249 bar. The same amount of hydrogen and oxygen diluted with steam instead of argon results in a peak pressure of 160 bar.

As described in [2] the burning velocity can be determined by measuring the pressure rise at the initial stage of the flame propagation. Therefore the maximum chamber pressure can be limited by an overpressure valve without losing valuable information. By limiting the chamber pressure independent of the combustion pressure the load on the chamber and the measurement instruments can be reduced. The maximum chamber pressure before the overpressure valve opens is set to 100 bar.

2) Calculation of Minimum Wall Thickness: For an initial estimate the combustion chamber can be approximated as a thick-walled cylinder. For cylindrical shapes hoop stress is the most critical stress, hence it is the only stress calculated. Hoop

stress for a thick-walled cylinder can be calculated by Lamé's equation:

$$\sigma_\mathrm{h} = \frac{p_\mathrm{i} \cdot r_\mathrm{i}^2 - p_\mathrm{o} \cdot r_\mathrm{o}^2}{r_\mathrm{o}^2 - r_\mathrm{i}^2} - \frac{r_\mathrm{i}^2 \cdot r_\mathrm{o}^2 (p_\mathrm{o} - p_\mathrm{i})}{r^2 (r_\mathrm{o}^2 - r_\mathrm{i}^2)}, \tag{3}$$

where r_i is the inner radius, r_o is the outer radius, r is the radius to a point in the wall, p_i is the inside pressure and p_o is the outside pressure.

For this calculation a maximum chamber pressure of $p_\mathrm{i} = 100\,\mathrm{bar}$, an outside pressure of $p_\mathrm{o} = 1\,\mathrm{bar}$ and an inner radius of $20\,\mathrm{mm}$ is assumed. A safety factor of 10 is desired. Taking a remaining yield strength of $150\,\mathrm{MPa}$ into account this results in a minimum wall thickness of $24\,\mathrm{mm}$. Because the chamber is not an actual cylinder and holes are not taken into account the minimum wall thickness is set to $30\,\mathrm{mm}$.

3) Strength Analysis by Finite Element Method: Followed by the calculation of the minimum wall thickness and the geometry of the chamber a CAD model of the CVCC was created. A section view can be seen in Figure 4.

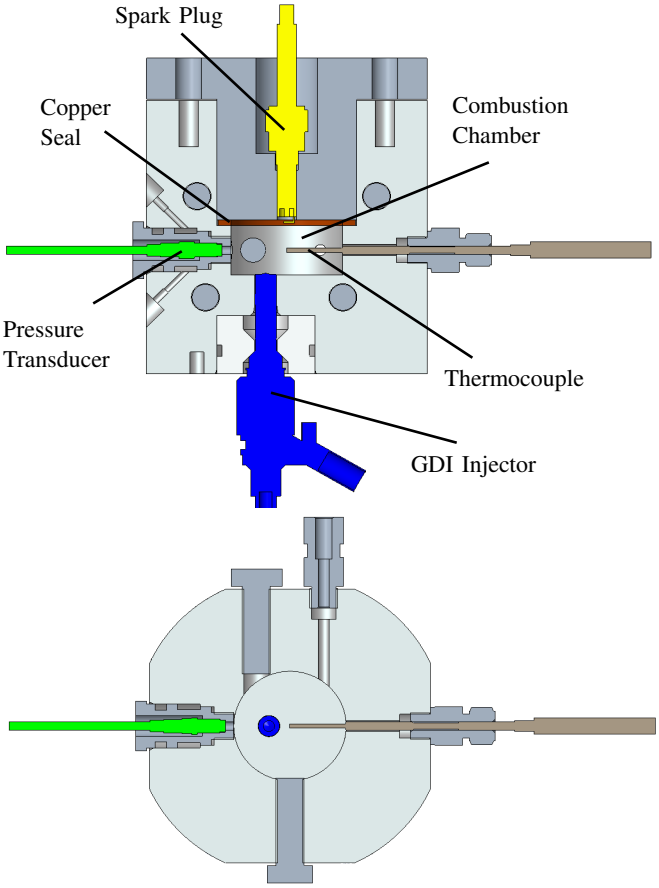


Figure 4. Section view of CAD model of the CVCC. Top: side view. Bottom: Top view.

In addition to the calculation of hoop stress the strength of the CVCC is analysed by the finite element method (see figure 5). The inside pressure is set to $100\,\mathrm{bar}$ and the top

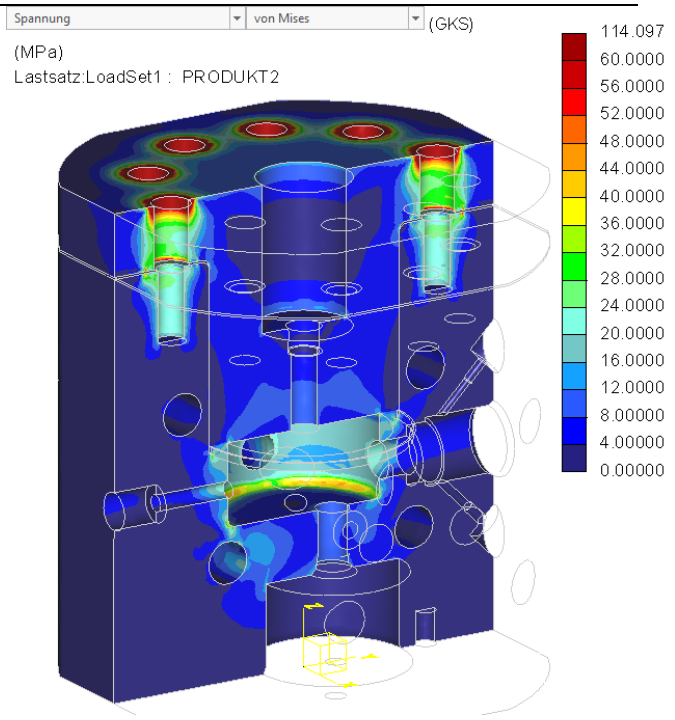


Figure 5. Analysis of the CVCC by finite element method. Stress in MPa by von Mises yield criterion.

part of the chamber which is screwed to the lower part has a preload of $50\,\mathrm{kN}$. The chamber is sealed with a copper ring. The model consists of 63.018 elements.

Stress in the side wall around the holes reaches $20\,\mathrm{MPa}$, in the corners of the chamber stress can rise as high as $40\,\mathrm{MPa}$ which is expected. While a safety factor of 10 was not achieved it is still considered as sufficient for the expected load.

As the injector and the piezo-electric pressure transducer both need to be cooled they represent heat sinks, thereby an even heat input into the chamber is difficult to achieve. To optimize the placement and the required power the same CAD model was used to simulate the thermal load. Four heating cartridges with a length of $60\,\mathrm{mm}$, a diameter of $8\,\mathrm{mm}$ and a maximum power of $500\,\mathrm{W}$ are used. The thermal simulation demonstrated that even if the heating cartridges are placed strategically and the power of each one is set independently an even temperature on the wall inside the chamber can not be achieved.

Because of this there is the risk of steam condensing on the relatively cold wall of the CVCC, as the boiling point at those high (partial) pressures is quite high.

IV. Conclusion and Outlook

Initial calculations point out that the currently available hydrogen gas supply ($10\,\mathrm{bar}$) restricts the possible gas compositions, thus solutions to overcome this limitation will be looked into.

According to the boundary conditions of an initial chamber pressure up to $70\,\mathrm{bar}$ and an initial temperature up to $450\,^\circ\mathrm{C}$

a CVCC was created. The CVCC is made from stainless steel AISI 316Ti and is designed to withstand a static pressure of 100 bar at 450 °C. The thermal simulation of the heating system brought the potential problem of an uneven temperature distribution to the attention. To solve this problem an insert like a bushing could be used to function as a heat shield in the area of the heat sink.

After the initial operation of this test bench further modifications can be made to acquire even more experimental information. For example the system can be modified to use a glow plug instead of spark plug, thereby surface ignition can be investigated as well.

To obtain as much data as possible in relatively short time the test bench needs to automated and optimised to precisely evaluate flammability limits with as little tests as possible.

In addition to the intended investigations, the CVCC can be used versatile to investigate different mixtures with different gases or even synthetic fuels.

In general this test bench makes it possible to obtain experimental information and data about hazards arising from the different applications with hydrogen like fuel cells.

ACKNOWLEDGMENT

The author would like to thank Prof. Dr.-Ing. H.-P. Rabl, head of the laboratory for combustion engines and emission control, for the supervision of this work and also his coworkers for their great assistance. The author also gratefully acknowledges the financial support of the Federal Ministry of Education and Research who make the construction and implementation of the developed test bench possible through the research project FEsMo-Tec [16].

REFERENCES

[1] Gavrikov, A. I. and Bezmelnitsyn, A. V. and Leliakin, A. L. and S. B. Dorofeev, "Extraction of Basic Flame Properties from Laminar Flame Speed Calculations," in *Proceedings of the 18th International Colloquium on the Dynamics of Explosions and Reactive Systems*, Seattle, Wash., July 2001, pp. 114.1–114.5.

[2] M. Kuznetsov, R. Redlinger, W. Breitung *et al.*, "Laminar burning velocities of hydrogen-oxygen-steam mixtures at elevated temperatures and pressures," *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 895–903, 2011.

[3] Y. Shebeko, S. G. Tsarichenko, A. Korolchenko *et al.*, "Burning velocities and flammability limits of gaseous mixtures at elevated temperatures and pressures," *Combustion and Flame*, vol. 102, no. 4, pp. 427–437, 1995.

[4] Deutsches Insititut für Normung e. V., "Gasanalyse - Herstellung von Kalibriergasen - Manometrisches Verfahren," Berlin, Oktober 2018.

[5] E. W. Lemmon, M. L. Huber, and M. O. McLinden, "NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 9.1, National Institute of Standards and Technology," 2013.

[6] M. Kuznetsov, J. Grune, V. Alekseev *et al.*, "Explosion Limits of Hydrogen-Oxygen-Steam Mixtures at Elevated Pressures and Temperatures," in *Proceedings of the 21th International Colloquium on the Dynamics of Explosions and Reactive Systems*, July 2007, pp. 281.1–218.4.

[7] H. Zhang, X. Bai, D. Jeong *et al.*, "Fuel combustion test in constant volume combustion chamber with built-in adaptor," *Science China Technological Sciences*, vol. 53, no. 4, pp. 1000–1007, 2010.

[8] B. Lewis and G. von Elbe, "Determination of the Speed of Flames and the Temperature Distribution in a Spherical Bomb from Time–Pressure Explosion Records," *The Journal of Chemical Physics*, vol. 2, no. 5, pp. 283–290, 1934.

[9] G. E. Andrews and D. Bradley, "Determination of burning velocities: A critical review," *Combustion and Flame*, vol. 18, no. 1, pp. 133–153, 1972.

[10] American Iron and Steel Institute, "High-Temperature Characteristics of Stainless Seel: A Designers' Handbook Series No 9004," 2020. [Online]. Available: https://www.nickelinstitute.org/media/4657/ni_aisi_9004_hightemperaturecharacteristics.pdf

[11] MatWeb, "Special Metals INCONEL Alloy 601." [Online]. Available: http://www.matweb.com/search/DataSheet.aspx?MatGUID=f3fb3ae6ebe54d98ad8fa01c74b6a3e8&ckck=1

[12] Allegheny Technologies Incorporated, "ATI 316 Austenitic Stainless Steel." [Online]. Available: https://www.atimetals.com/Products/Documents/datasheets/stainless-specialty-steel/austenitic/ati_316ti_tds_en_v1.pdf

[13] R. Munsin, C. Bodin, S. Lim *et al.*, "Design of Constant Volume Combustion Chamber (CVCC) with Pre-Combustion Technique for Simulation of CI Engine Conditions," in *Proceedings of the 4th TSME International Conference on Mechanical Engineering (TSME-ICOME)*, 10 2013.

[14] A. Phan, "Development of a Rate of Injection Bench and Constant Volume Combustion Chamber for Diesel Spray Diagnostics," Graduate Theses, Iowa State University, Iowa, 2009.

[15] G. P. Smith, D. M. Golden, M. Frenklach *et al.*, "GRI-Mech version 3.0." [Online]. Available: http://www.me.berkeley.edu/gri_mech/

[16] L. Langwieder, "Neue Anlage zur Entwicklung zukünftiger grüner Energiesysteme," 2020. [Online]. Available: https://www.oth-regensburg.de/new-startpage/hochschule/aktuelles/einzelansicht/news/neue-anlage-zur-entwicklung-zukuenftiger-gruener-energiesysteme.html

Cheap Car Hacking for Everyone - A Prototype for Learning about Car Security

Sebastian Baar (B.Sc.)
Faculty of Electrical Engineering
and Information Technology
OTH Regensburg
Email: sebastian1.baar@st.oth-regensburg.de

***Abstract*—Automotive security will become more and more important in the comings years. Cars will increase their interfaces to the outer world and because of that their attack vectors will increase. To confront this problem, ways to teach and learn about this topic are needed, but cars and electronic control units are too expensive to be used by the majority of students. Therefor we present a way of using consumer electronics and open source software to create a prototype for automotive network security education. This platform is set to be portable inside a coffer and, due to its nature of using free software, is able to simulate real life attacks on cars for demonstration purposes as well as learning about the security flaws modern cars contain.**

***Index Terms*—Car Security, CTF, Education**

I. Introduction

The demand for competent security researchers and developers in the automotive and embedded market is rising for years. Car manufacturers as well as universities are seeking appropriate ways and tools to qualify their students and developers for this thematic. The high cost of cars makes a hard entrance barrier for hands on introduction into automotive penetration testing. Yet this niche is becoming more and more important for security analyses. Already carried out attacks on cars [1][2] show the vulnerabilities and attack vectors [3], as well as the possible implications of such attacks.

This project aims to confront these entrance barriers and problems by trying to develop a cheap automotive network security platform, while trying to leave it as realistic as possible. For this goal we will show that it is possible to make an affordable hacking platform in which real life attack scenarios can be displayed and repeated. On top of that the widespread use cases for open source software for penetration testing cars and building up learning platforms will be shown. Some research and prototypes have already been developed, but with oftentimes different goals in mind. These are readable in section II. Due to it's nature of being primarily developed for university students, certain objectives for the learning process have to be set out. This takes place in section III.

The overall architecture of the platform is outlined in section IV. The architecture is divided in the hardware and the software part. The hardware part shows the measures which had to be taken to keep the costs of this project low. The software part shows the main programs and its use cases, where the importance of open source software is highlighted. After that, two implemented challenges are described, which are oriented on real life scenarios.

At the end, section V gives an overview of the accomplished work and shows possible future enhancements.

II. Related Work

Automotive networks in coffers for security research purposes are a niche concept, which is not researched much yet. But car manufacturers are starting to develop their own testbeds for teaching their own developers. Toyota for example started with the Portable Automotive Security Testbed with Adaptability (PASTA) project, their first adventure of this kind[4]. Because of the ongoing development in the autonomous driving sector, Toyota saw the need for an open and programmable research tool, which allows them to make security analyses of electronic control units. They fit a small coffer with four microcontrollers, all programmable and with an open design, and connected them together with the use of the CAN bus. They implemented four microcontrollers as ECU replacement inside this setup, which can be programmed depending on the use case. These are then on connected with each other over a CAN bus, which is also implemented inside the coffer. The system allows for future connections like for example Ethernet or Bluetooth.

Another similar model of a partial car communication build up was presented by the security researchers Charlie Miller and Chris Valasek. They woth work on research regarding the penetration testing of cars and ECUs. They are known for their Cheep Cherokee Hack[1][2][5] in which they successfully altered the internal CAN bus communication to control the car from distance. The model they presented had the goal to lower the entry barriers for security researchers in the automotive sector. Because of the high costs of cars, this limits the research possibilities for entry level analysts. They are in favor of buying used single car ECUs and the buildup of a testbench, in which the ECUs are connected depending on the attack scenarios.[6] It should be noted that this approach only applies to the execution of penetration tests, but leaves out the whole educational aspect of teaching car security and also capture the flag challenges (CTF).

The general usage of CTF challenges has already been researched by different institutions. A russian university for example used this kind of approach to teach students about

information security. The result of their study have shown that the participation and the motivation of the students reaches higher levels[7] when using this gamified way of teaching.
Another similar idea was tried by the makers of picoCTF [8]. In contrast to the already mentioned PASTA coffer of Toyota, they used a completely virtualized approach. A complete Open Source framework for build CTF style challenges was made with the intention of other people contributing with new challenges. During a competition, which was aimed at american high school students, over 2000 people participated. The projects main goal was to try and use CTF style competition on high schoolers. Due to the nature of these competitions, most of them are aimed more towards university students[9].

III. Objectives

Due to the aforementioned lack of fitting methods to learn about car security, the idea of a platform for learning about these topics was born. Certain objectives for this have been defined:

- **Low Costs** Existing platforms for learning about car security are not cheap. The PASTA platform of Toyota for example is still as costly as a middle class car. Buying ECUs and setting up a automotive network testbed, as recommended by Miller et al. [6], is also breaking the 1000 Euro barrier. Therefor the platform must use cheaper alternatives. To measure this feature the best case scenario is to be cheaper than 400 Euros, whereas the worst case is to be as expensive as a standard new car, which would be around 25.000 Euros.
- **Portable** Using a car to study automotive security, the portability is not existing. The space required for the car and the costs for a parking spot are well over the budget. Therefor the goal is to get the whole automotive network inside a small portable coffer, which can be used anywhere. Here the best case scenario is if the whole platform can be used anywhere in the world by being able to put it in a standard size coffer or bag. The worst case would be if it is stationary and only move able with a distinct temporal forerun.
- **Open Source** Due to the first objective, keeping the costs down, proprietary software can't be used. Another reason to choose open source is the ability to let people over the world participate. The Scapy framework already has many automotive protocols implemented and there are existing kernel modules for Linux operating systems to communicate over CAN. With open source researchers can use already implemented methods and contribute new ones. The success metric hereby ranges from "only proprietary software" to "only Open Source software".
- **Gamification Approach** For making users more interested in car security the learning should follow modern ways of teaching. Using CTF style challenges are used to create a competitive environment in which users are eager to learn more and accomplish more tasks, while keeping it entertaining and fun. The success metric for this ranges from "not suited for educational purposes" to "in line with education purposes for universities as well as the industry".

With the shown objectives an evaluation of our approach against the existing alternatives in section II is made by using the success metrics in table I:

TABLE I
This papers approach listed up with the objectives and their measurements from section III plus the "Automotive Security Topic" feature. The scale ranges from +3 (very strong) to -3 (very weak), where 0 is indifferent

	Our Approach	**PASTA**	**Chris Valasek & Charlie Millers Approach**	**picoCTF**
Low Costs	3	-3	1	3
Portable	3	3	-2	3
Open Source	3	3	1	3
Gamification Approach	3	1	-3	3
Automotive Security Topic	3	3	3	-3

The overall alignment of the platform is to make learning about the security of automotive networks cheaper and more accessible. The cost factor is an important of this project, which will be covered in the next chapter. Another important topic is that of the education part. Learning about car security should be interesting and fun for students, which means that the project is taking the gamification approach. To make this approach consistent with the alignment, certain goals, methods and content of the education part have to be defined:

- **Learning Goals** One of the most important goals is to enhance the security education on the topic of connected cars. The awareness regarding attack and manipulation vectors of modern cars should be raised. On top of that basic knowledge about the composition of modern car networks, as well as the approach of malicious software should be conveyed.
- **Learning Methods** Students should in principle solve exercises, so called challenges, which should be set up as realistic as possible. A notebook must be plugged into the interfaces of the coffers network and predefined tasks should be understood and solved by the students. For that a point based system will be used. Each challenge has a certain value, depending on the difficulty of the task.
  After solving the challenge, the students get a unique string, so called flag, which they have to enter in the accompanying website. Through this point based system, a competition should be created between the students in which they get motivated to use more of the learned knowledge and therefor solve more of the assignments.
  A russian university already applied a similar approach in using CTF games to teach students about network security. In the publication *A CTF-Based Approach*

in Information Security Education Alexander Mansurov showed that the gamification of learning material and the competitive part of CTFs enhance the motivation and the learning speed of the participants [7].
But due to high knowledge barriers in the security fields, authors like Cheung et al. plead for a workshop class in which theoretical and practical parts are combined [10]. This is implemented in our challenges in which a lecturer will start with explaining theoretical concepts and after that the penetration testing starts.

- **Learning Content** Due to the topic of automotive security the content of the learning experience covers a wide area. The network architecture of cars, as in the differences of older architectures to newer architectures is one of the first starting points. From there on the different protocols which are used will be discussed and learned. These range from CAN, ISO-TP, UDS to newer ones like Automotive Ethernet protocols. This covers the theoretical part of the learning experience.
  After that, the students will gain experience in using penetration testing tools like Scapy. This will be accomplished by using predefined challenges, which all contain a real life scenario of a attack on a cars ECU or network.

IV. Architecture

Derived from the objectives in section III the architecture of the platform can be build. It is divided in the hardware and software aspects of the platform.

A. Hardware

The objectives ”Low Costs” and ”Portable” are the ones influencing the hardware part the most. ”Portable” requires the platform to be able to be fitted inside a coffer. ”Low Costs” limits us in our selection of fitting ECU replacements. Figure

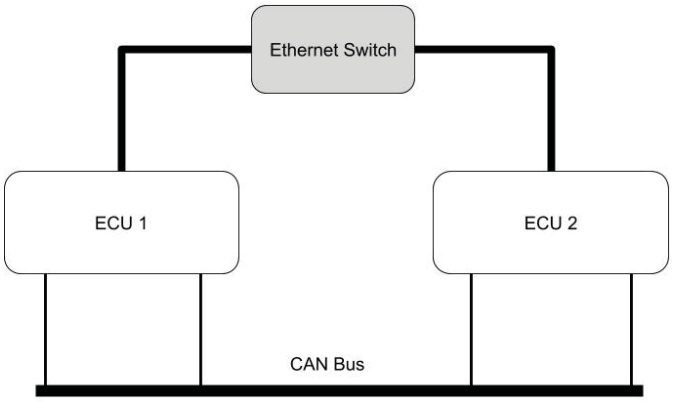


Fig. 1. Abstract graphical overview of the hardware architecture of the platform, with two ECUs replacements and two bus systems

1 shows the basic idea of how the networks inside the platform look like.

1) Electronic Control Units: Modern cars can contain over 80 different ECUs. Their performance ranges from small, bare metal microcontrollers to powerful machines with operating systems and web browsers. Therefor the hardware should represent that. Raspberry Pi Computers, which are used to simulate the ECUs, are able to offer both a low performance and high performance device.

2) Bus Systems: Following the idea of keeping the project as close to real car networks, the bus systems inside the platform consist of CAN and Ethernet. CAN is still the most used bus inside modern cars and will be for the foreseeable future. The easiest way of setting up this bus was to use a 9 pin ribbon cable, with the two communication lines CAN HIGH and CAN LOW being terminated with a 120 Ohm resistor. Interfaces for connecting the ECUs to the bus are spread all over the bus.

Automotive Ethernet is the most recent development in the car network sector. Due to the need for faster communication, more bandwidth and the problems of developing a new system, car makers choose the existing and well tested Ethernet. The OSI model contains 7 layers for communicating inside a network. Protocols from layer 2 and upwards are usable for the specific demands inside a car. The existing layer 1 technologies for Ethernet are also usable for diagnostic purposes like Diagnostic over Ethernet (DoIP) [11, p. 141]. But for the internal communication another physical layer was invented: BroadR-Reach. This layer allows for better electromagnetic compatibility and weight reduction. The compatibility with standard Ethernet connections is not possible with BroadR-Reach. Due to the availability and cost problems with this technology and the otherwise trouble-free usage of the higher layers, the car hacking platform uses standard Ethernet. Protocols like the aforementioned DoIP and SOME/IP are set in the layer 7 of the OSI model and are therefor unaffected by these lower layer changes.

3) Cost: With ”low cost” as a main objective for the project and with the guideline of being cheaper than the alternatives (e.g. automotive testbed for around 1000 Euros [6]), certain limitations had to be applied. The used electronics is mostly consumer electronics and therefor no automotive grade. Table II gives an overview of the total cost.

TABLE II
Cost statement of the project

Amount	Description	Cost (in Euro)
1	Coffer	50
2	Raspberry Pi 4 (2Gb)	100
1	Ethernet switch	30
3	Ethernet cable	25
1	9 pin ribbon cable	8
1	Raspberry Pi Display	65
1	Power supply	25
2	PiCAN 2 Duo	65
1	Miscellaneous	30
	Total Cost	**398**

The ECUs from figure 1 are Raspberry Pi single chip comput-

ers. They are equipped with an operating system which allows the development of more complex challenges. Their Ethernet ports allow the simulated Automotive Ethernet communication inside the platform. As for the CAN communication the PiCAN 2 Duo shields are used. They are equipped with two CAN interfaces, which enables us to simulate four ECUs in total on the bus.

B. Software

1) Raspbian OS: The most used operating system for Raspberry Pis is the Raspbian OS. It is a Debian-based Linux operating system. Due to its open source nature, certain modifications for the platform were able to be made. Restricted user accounts, who are not able to read and therefor bypass the task to solve the challenges. On top of that preconfigurations regarding the WiFi connection, the ability to configure the ECUs remotely and the ability to flash the challenges have been made.

2) can-utils: Debian based operating systems allow for the installation of the can-utils package from the official repositories. The package has implemented user space applications for communications over CAN in combination withe the SocketCAN Linux subsystem.

3) Scapy: Scapy is a python framework for packet manipulation. It allows the creation, modification, receiving, and sending of messages. Due to its open source nature, the contribution of new protocols is possible. Many automotive protocols like CAN, UDS and SOME/IP have already been implemented for usage. The challenges inside the coffer are all implemented with this framework.

4) Challenges: **CAN Man-In-The-Middle**

Man-In-The-Middle (MITM) attacks are a well known and used threat on every network. They are a "[...] clever way to circumvent encryption [12, p. 406]". The basic idea is that instead of two communication partners A and B communicating with each other, a third participant, the attacker, is sitting right between them. When A thinks he is talking to B, he is instead communicating with the attacker and vice versa. Due to the CAN bus' missing encryption, the question for a MITM attack use case on the CAN bus arises. One real life example of such an attack is the manipulation of the odometer value. The YouTube Channel bigclivedotcom shows such an attack with a common microcontroller [13]. In this example the microcontroller with two CAN interfaces is used to intercept the correct message containing the odometer value, create a new message with the original odometer value minus 40.000 and then send it to the dashboard. The main objectives of such an attack are letting the car appear less used and increasing the selling price.

Due to the simple nature of this attack and it's high potential of manipulating the cars worth, it was chosen as a introduction challenge for this project. For the two ECUs communicating with each other, the Raspberry Pis where used. The master Raspberry Pi containing the Display is able to link the messages' values to the real life items, like for example the tachometer or the rev counter. With these visible informations and the use of the *candump* command from the *can-utils* package, or the Scapy functions, the user is able to link the CAN identifiers to the corresponding values. Here starts the physical part of the challenge, where he has to remove both ECU mockups from the common bus and connect each of them to one of the CAN interfaces. From there on he needs to write a program which intercepts only the CAN message with the odometer value, decrease its number by a predefined value and send it to the correct mockup. After completing these steps the flag for solving the challenge will appear on the display.

UDS Scanning

Unified Diagnostic Services (UDS) is a specification for diagnostic purposes. It allows the communication and maintenance of ECUs. Services like "Read Data By Identifier (RDBI)" or "Read Data By Address (RDBA)" are implemented for this purpose. Because of this wide reaching access on an ECUs internal program code and the security implications of this, UDS is often used during attacks on cars [14].

For the challenge users need to gain access to certain security levels. This can be done by using the RDBI and RDBA services to gain knowledge about the system. With enough information gathered the security level change can be started. Even though the implementation of the security parts of UDS are not standardized, an often used mechanism is the seed key procedure. The tester requests a seed from the car, which generates it and sends it to him. Then a predefined algorithm computes the seed and sends the key to the car. If the key generated inside the car is the same as the one sent from the tester then security access is granted. The length of the seed key pairs is not defined, but the most common length is 16 bits. These 16 bits are an easy target for modern computers and can be broken in approximately 110 hours using brute force attacks [14]. This overstretches the time constraints for challenges of the platform. Therefor other ways for authentication are used. In the UDS Scanning challenge certain hints are given to the user when he uses the UDS method RDBI. With these informations the user is able to guess the authentication method as a byte wise XOR operation of the seed with the number 1337. By sending the correct key to the ECU the challenge is solved.

V. Conclusion and Future Work

In this work we showed different existing approaches for learning about car security. We described our own approach and its goals, objectives and functionality. For each objective we used an individual scale to rate each one's success in reaching these objectives. We put our own prototype against the available solutions and compared its features.

Learning about car security is, as explained in the beginning, oftentimes expensive and not accessible for most people. Only one out of the three other approaches we showed, the Toyota PASTA coffer, was specifically for a similar use case as our prototype. But this coffer still is not financially accessible enough for most people. As far as the portability goes, only two other approaches could be rated the same as ours. One of them being the PASTA coffer, the other the picoCTF,

which is missing the automotive security focus. In the specific use case of a low cost, automotive network security learning environment, which is also easily portable and not as static as car or a test bench, we are able to offer our prototype as a accessible solution.

As for future work on this prototype, more and more challenges will be implemented. Another major change could also be to follow a similar approach as the picoCTF and to implement challenges inside a virtualized environment, therefor cutting the cost factor even more and nearly completely cutting the hardware part. Upcoming trends in the automotive world like automotive ethernet give a better chance of using existing transportation protocols from the web development field, as for example the CAN and UDS protocols. The accessability and portability could also be improved with this approach, therefor widening the group of people who can learn about car security.

REFERENCES

[1] C. Miller and C. Valasek, " Remote Exploitation of an Unaltered Passenger Vehicle ," http://illmatics.com/Remote%20Car%20Hacking.pdf, [Online; last access 10.06.2020].

[2] ——, " CAN Message Injection ," http://illmatics.com/can%20message%20injection.pdf, [Online; last access 10.06.2020].

[3] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, K. Koscher, A. Czeskis, F. Roesner, and T. Kohno, "Comprehensive experimental analyses of automotive attack surfaces," in *Proceedings of the 20th USENIX Conference on Security*, ser. SEC'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 6–6. [Online]. Available: http://dl.acm.org/citation.cfm?id=2028067.2028073

[4] T. Toyama, T. Yoshida, H. Oguma, and T. Matsumoto, "PASTA: Portable Automotive Security Testbed with Adaptability," https://i.blackhat.com/eu-18/Wed-Dec-5/eu-18-Toyama-PASTA-Portable-Automotive-Security-Testbed-with-Adaptability-wp.pdf, [Online; last access 10.06.2020].

[5] "Hacker steuern Jeep Cherokee fern," https://www.heise.de/security/meldung/Hacker-steuern-Jeep-Cherokee-fern-2756331.html, [Online; letzter Zugriff 07.01.2019].

[6] C. Miller and C. Valasek, "Car Hacking:For Poories," http://illmatics.com/car_hacking_poories.pdf, [Online; last access 10.06.2020].

[7] A. Mansurov, "A ctf-based approach in information security education: An extracurricular activity in teaching students at altai state university, russia," *Modern Applied Science*, vol. 10, no. 11, pp. 159–166, 2016. [Online]. Available: http://www.ccsenet.org/journal/index.php/mas/article/download/60685/33468

[8] P. Chapman, J. Burket, and D. Brumley, "Picoctf: A game-based computer security competition for high school students," in *3GSE*, 2014.

[9] J. Werther, M. Zhivich, T. Leek, and N. Zeldovich, "Experiences in cyber security education: The mit lincoln laboratory capture-the-flag exercise," 08 2011, pp. 12–12.

[10] R. S. Cheung, J. P. Cohen, H. Z. Lo, F. Elia, and V. Carrillo-Marquez, "Effectiveness of cybersecurity competitions."

[11] W. Zimmermann and R. Schmidgall, *Bussysteme in der Fahrzeugtechnik*, 5th ed., ser. ATZ/MTZ-Fachbuch. Wiesbaden: Vieweg, 2014.

[12] J. Erickson, *Hacking: The Art of Exploitation, 2nd Edition*, 2nd ed. USA: No Starch Press, 2008.

[13] C. Mitchell, " Naughty CANbus odometer "interface". (Fakes mileage.) ," https://www.youtube.com/watch?v=f4af1OBU5nQ, [Online; last access 10.06.2020].

[14] M. Ring, T. I. A. S. rensen, and R. Kriesten, "Evaluation of vehicle diagnostics security – implementation of a reproducible security access," in *SECURWARE 2014*, 2014.

SESSION C1

Lukas Escher
Design and characterization of an in air nitrogen dioxide trace gas detection sensor

Mario Aicher
Evaluation of different hardware platforms for real-time signal processing

Lukas Reinker
Measurement of kinematics and muscle activity of athletes under stress

Ludwig Brey
Development of Automatized Procedures for the Generation of Complete Measurement Time Series

Design and characterization of an in air nitrogen dioxide trace gas detection sensor

Lukas Escher, Thomas Rück
Sensorik - Applikationszentrum (SappZ) Regensburg
OTH Regensburg
lukas.escher@st.oth-regensburg.de

***Abstract*—The concentration of nitrogen oxides in ambient air is an important contributor to air pollution. As a consequence of high soil in the air, considerable health outcomes as well as an increase in climate change can be the result. The measurement of nitrogen dioxide (NO_2) as part of the nitrogen oxides is therefore a crucial point for assessing air quality. In this work a small, transportable sensor for trace gas detection of NO_2 in air is developed in order to provide a system capable of monitoring local concentration changes. Photoacoustic spectroscopy (PAS) is used as measuring principle in the sensor setup. A periodically amplitude-modulated laser excites the analyte molecules through absorption of photons. The subsequent non-radiative relaxation via energy transfer results in a local increase in temperature, which leads to a pressure change. As the laser is modulated periodically, the pressure oscillation forms a sound wave detectable with a microphone. The sound volume is then an indicator for the NO_2 concentration.**

***Index Terms*—Photoacoustic, Photoacoustic spectroscopy, Nitrogen dioxide detection**

I. Introduction

Despite well known health effects, negative environmental impacts and premature deaths, humans all over the world are still exposed to high air pollution concentrations which exceed reference concentrations set by EU and WHO [1]. Especially in densely populated areas with heavy traffic, a higher NO_2 exposure can be measured. Citizens living there are at a higher risk of developing asthma or irritations of the airways which can lead to premature death. Further a increased incidence of cancer is suspected but not yet scientifically proven [2].
The WHO proposed a hourly mean of $200\mu g/m^3$ (106.4ppb) and an annual average limit of $40\mu g/m^3$ (21.3ppb) for NO_2 [3], which has been adopted by the Federal Environment Agency in Germany [4].
Due to cost and size of highly accurate measurement setups, long term monitoring of NO_2 so far has been carried out only at a few fixed locations, resulting in a poor spatial resolution of NO_2 distribution. To remedy this situation, a sensor network, consisting of small low-cost sensors is required.
Therefore, this research work aims to develop a small, transportable and inexpensive measurement system for a continuous in air trace gas detection of NO_2 using photoacoustic spectroscopy (PAS) as measurement principle.

II. Photoacoustic Spectroscopy

In 1880 the photoacoustic (PA) effect was discovered by Alexander G. Bell. It describes the production of a sound wave induced by absorption of light in solids [5]. After understanding the process of light absorption by the theory of quantum mechanics at the beginning of the 20th century, the lack of suitable light sources delayed the application of Bells discovery to the second half of the century [6].
The standard absorption spectroscopy computes the concetration by determining the lost light energy through a sample. The lower the concentration in the specimen, the more similar the optical power in front of and behind the sample, which restricts the maximum limit of detection. Contrary to this, PAS determines the concentration of trace gases with a direct method, as the absorbed light energy is not compared with a value, but converted directly into the sound signal that indicates the concentration. As the sensitivity is proportional to excitation optical power, the performance of PAS based sensors can benefit from the high output power levels achieved as a result of technology developments by the semiconductor industry [7] .This makes PAS a powerful measurement principle to detect low concenrations of NO_2 in ambient air.

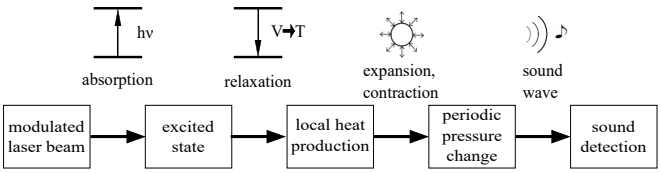


Fig. 1. Signal generaion using PAS.

Fig. 1 shows how the Signal in a PAS sensor is generated. First the analyte sample in a measurement cell is irradiated by light with wavelength coinciding with an absorption band of the substance. The molecules are energetically excited to electronic or vibronic states by absorbing this light and can relax via collisions with another molecule. This non-radiative relaxation converts the energy of the absorbed photon into translational energy which leads to a local heat production in the gas, also regarded as pressure change. Due to the periodical modulation of the light source, the temperature and therefore the pressure oscillates with the modulation frequency applied. PAS measurement cells are designed as acoustic resonators to increase the signal-to-noise ratio (SNR). Therefore, the excitation at the resonance frequency, which depends on the geometry of the resonator, forms a standing acoustic wave that can be detected by different sensors e.g. microphones or

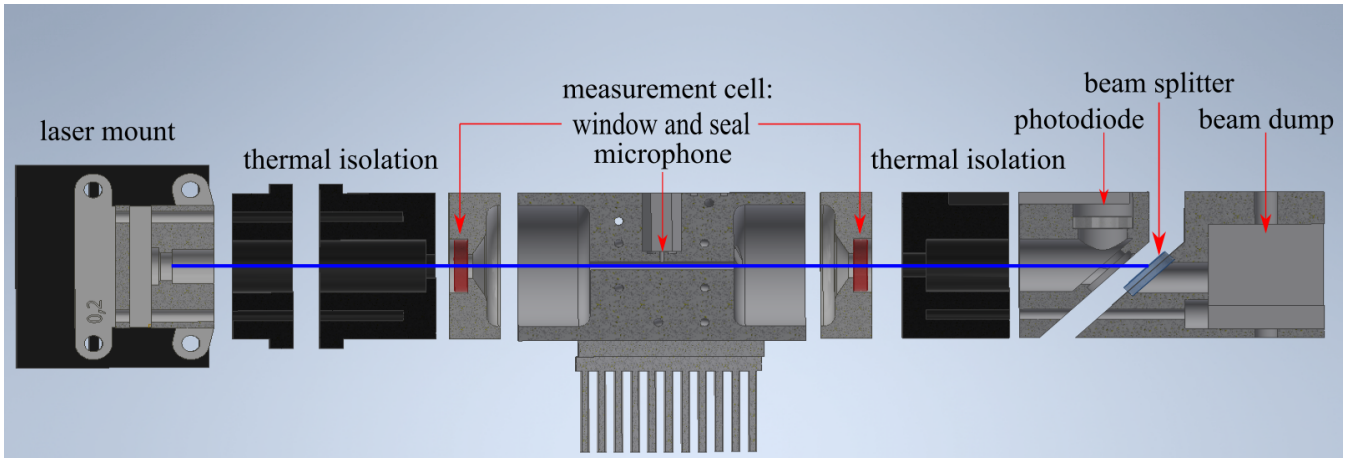


Fig. 2. Cut through an expanded 3D model of the designed measurement setup. The blue line indicates the optical pathway.

quartz tuning forks. The amplitude or volume of this signal is proportional to the gas concentration [8].

III. Measurement Setup

The individual parts of the setup shown in Fig. 2 are designed with CAD and manufactured by 3D-printing. The laser mount, measurement cell and its seals consist of Metal (AlSi10Mg) whereas the other components as the thermal isolation and the parts of the beam splitter are printed in plastic for thermal isolation issues, cost and production speed. As a whole, the setup has a length of 20cm and width and height of 4cm. Not shown in Fig. 2 is the peripheral electronics, consisting of a board laser and thermoelectric cooler (TEC) driver (meerstetter LDD-1121 & TEC-1122), frequency generator (Keysight 33522B) for triggering, a lock-in amplifier (LIA; Signal Recovery 7270) to evaluate the microphone signal and an external power supply. All devices are controlled by self-written code via LabView on a PC and the signals are read and processed.

A. Laser diode

For NO_2 detection a blue 447nm laser diode (Osram PLPT 450D_E A01) in a TO90 package is used. The diode is supplyed with a 50% duty cycle square wave amplitude modulated current from the laser diode driver at a operating point of 2,1A with a frequency fitting to the first harmonic of the measurement cell's acoustic resonator. The board driver itself is triggered by the frequency generator to simultaneously send a reference signal to the LIA for phase information. As the PA signal is directly proportional to the optical power, this laser diode is perfectly suited for the application due to its high output power($>$1W at operating point) compared its size. For wavelength and power stability the laser mount is cooled by a peltier element and a fan to remove the large waste heat of ca. 7W efficiently, regulated by the TEC driver. However, since the laser light can also interact with the metal of the measurement cell and resonator and therefore contribute to a PA background signal, a collimation lens system is placed directly in front of the diode, where the blue line in the figure starts. Ideally, as indicated by the line, the optical pathway only interacts with the windows and the sample gas.

B. Measurement cell

The used measurement cell includes a double open ended, cylindrical acoustic resonator with a differential microphone (InvenSense ICS40730) placed on the upper side in the middle for optimal signal detection, as the maximum amplitude of the 1st harmonic pressure wave is right in the middle. At both ends of the resonator, buffer volumes with much bigger diameter are designed to suppress external noise caused by e.g. the gas flow. To seal the measurement cell, the buffer volumes are closed by seals and antireflective (AR) wedged windows (Thorlabs WW40530-A). AR is needed to reduce absorption and stray light of the laser beam by the windows, that would also contribute to the background signal. Like the laser mount, the whole cell is temperature regulated by the TEC with a peltier element to 40°C. As the speed of sound changes with temperature the resonance frequency changes, following $f = c/\lambda$. So this constant regulation ensures a stable resonance frequency and therefore prevents signal changes caused by temperature variations. In order to heat up the whole sample gas to the correct temperature, it is passed through a spiral path over a long distance before entering the resonator.

C. Secondary sensors

With the aim of monitoring external influences on the measurement system, secondary sensors are implemented in the setup. A photodiode (PD) is attached to the upper part of the beam splitter, where a small part of the laser light is directed. The rest of the laser power is absorbed by a beam trap. The PD monitors the laser power. Measuring a drop in power can indicate dirt on one of the windows, aging or failure of the laser and damage to the collimation optics and therefore is important to distinct signal drops caused by these issues. Further, a temperature, pressure and humidity sensor is implemented in the gas flow of the measurement

cell (Bosch BME280). This gives the opportunity to surveil the temperature stability. For characterization, pure gas mixtures with set concentration, pressure and humidity are measured. The BME makes it possible to calibrate the setup towards these measurements in comparison with the sensors from the gas mixing system. Thus, this helps a later application in ambient air, where these values vary and so the measurement conditions need to be adjusted, following the calibration.

IV. Setup Characterization

Before the measurement system can be used in ambient air, several calibrations and characterizations need to be carried out regarding laser emission, resonance frequency, background signal, cross-sensitivities to other gases, humidity, temperature, pressure and ambient noise. The following section covers the first three issues.

A. Laser

Emission wavelength and power of semiconductor diode lasers depend on the temperature of the chip's junction. Obviously power increases towards lower temperature, as recombination processes in the active region are more efficient. Therefore, a low laser mount temperature and thus a lower

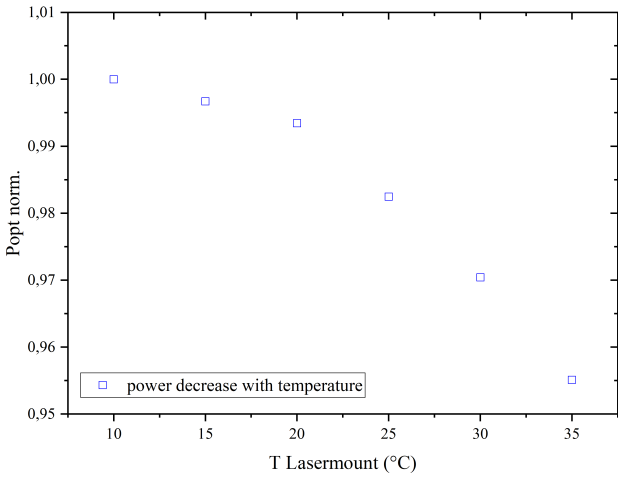


Fig. 3. Laser power decrease at different mount temperatures.

junction temperature is helpful for higher laser power, as Fig. 3 shows and due to the linear relation also increases the PA signal if collimated correctly.
The emission wavelength of the laser was measured using a spectrometer which showed, that at room temperature the laser emits roughly one nanometer below the datasheet value. It can be observed, that it undergoes a red shift with increasing temperature due to decrease of the band gap of the diode. This shift is linear in the measured temperature range with a dependency of $d\lambda/dT = 0,053nm/K$. In combination with the measured power, the emission spectrum overlap with the microstructure of the the absorption spectrum of NO_2 gives an indicator for the optimal operating temperature presented in Fig. 4. Estimating the highest PA signal is achieved at a

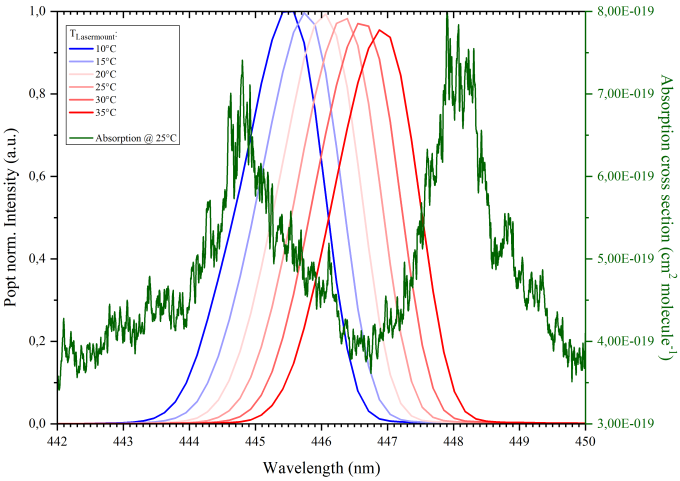


Fig. 4. Laser emission spectra for different mount temperatures normalized to the corresponding optical power and compared to the absorption cross section of NO_2 on the right axis. Absorption cross section extracted from [9].

mount temperature of 10°C. This is proven by computation and measurement. Multiplication of the emission spectra, normalized to the optical power, with the absorption cross section and integration carried out on it gives, referenced to the maximum value, a good prediction of the relative PA signal change. Comparing these results to a measurement of the PA amplitude at different laser mount temperatures, this confirms the previous guess that 10°C gives the highest signal, shown in Fig. 5. It turns out, that a laser emission as close as possible to one of the two absorption maxima is ideal. Since the laser degrades faster at higher temperatures, 10°C are used as operating point. Ideally, the laser emission given from the device would already be on one of the peaks. Shifting the

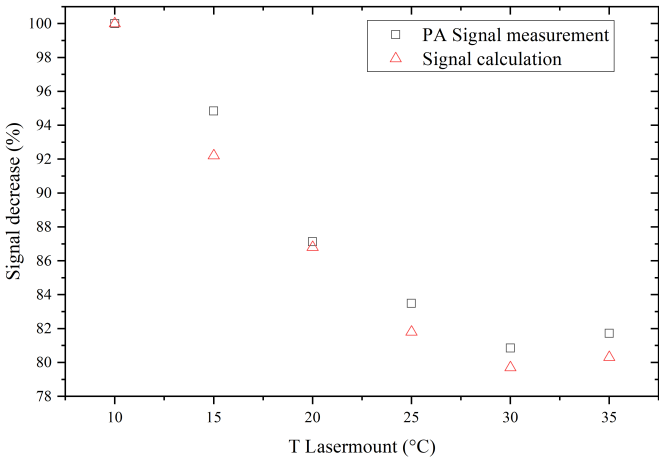


Fig. 5. PA Signal decrease on the one hand measured at different laser mount temperatures and the other computed by the calculation described in the text.

measured spectra to different wavelengths and calculating the PA amplitude referenced to 10°C gives a maximum PA signal at the right absorption peak using the power and full width at

half maximum (FWHM) of the 10°C spectrum. Approximately 10% PA signal benefit can be expected, if the laser diode would emit at 448nm compared to the 10°C emission. After setting to the ideal 10°C laser mount temperature, a power-current characteristic curve is measured with the purpose of calibrating the voltage signal from the photodiode to the actual laserpower.

B. Resonance profile

In order to find the resonance frequency of the acoustic resonator, a measurement of a dry gas sample at 40°C cell temperature with a mix of NO_2 and synthetic air (79,5% N_2 with 20,5% O_2) is carried out. The mix is set to a concentration of 20ppm (parts per million) NO_2. Theoretically with the speed of sound at 40°C ($c = 355,576m/s$) and the length of the resonator tube ($l = 3,1cm$) the resonance frequency can be calculated. As its a double open ended resonator, the wavelength of the 1st harmonic is $\lambda = 2l$. Also a shift of the resonance node points occurs in an open ended resonator, so the effective length is longer than the actual dimensions ($l_{eff} = 3,465cm$). Following $f = c/\lambda = c/2l_{eff}$, the theoretical resonance frequency is $5131Hz$.

The resonance frequency is determined by sweeping from low

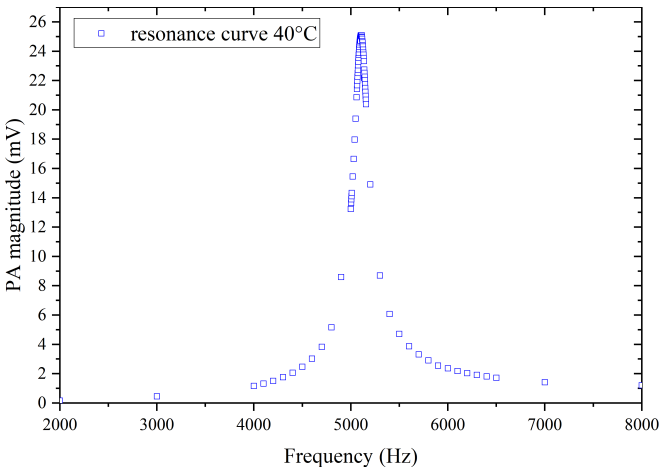


Fig. 6. Frequency sweep for characterizing the resonator of the measurement cell in Fig. 2.

to high frequencies and measuring the PA amplitude averaged for 100 data points, shown in Fig. 6. The resonance frequency is measured to be at $5110Hz$, the difference to the calculated value can be explained by the gas temperature being offset to 40°C and the node shifting estimation not being perfectly correct. As the cell is heated, the resonance frequency is constant for dry air. If the gas sample is humid, a shift occurs, and a new sweep needs to be carried out. For later application a sweep routine, that reacts to external influences measured by the secondary sensors need to be integrated into the software. A way to determine the quality of the resonance curve is the dimensionless Q factor. If the factor is low, that means the resonator rings better or is less damped. Its magnitude is computed by $Q = f_{res}/\Delta f$, with the resonance frequency f_{res} and Δf equaling the FWHM of the peak. In addition to the dimensions, the quality of the pipe's polish is one example of influence factors on a high Q factor. For this resonance profile the Q factor is 20,3 which matches to the other PAS measurement setups in the same laboratory with the same size, whose values are between 20 and 25.

C. Background signal

The background signal and the corresponding noise of a PAS setup restricts its limit of detection and signal stability. Different background signal sources can be distinguished:

- Sound coming from external sources or the gas supply and flow.
- Electromagnetic compatibility (EMC) generating crosstalk from electronic devices and cables to the signal line from the microphone to the LIA that therefor is shielded.
- Photons hitting the microphone membrane and producing a signal by impulse transmission to the membrane.
- Laser light interacting with the walls of the resonator tube caused by poor collimation and stray light.

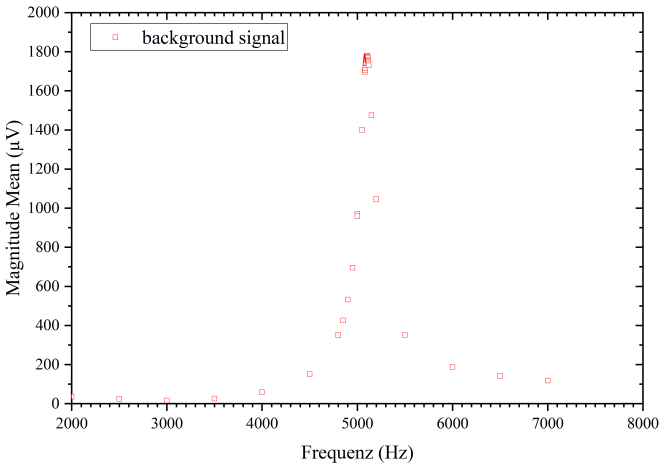


Fig. 7. Frequency dependent background signal caused by laser light interaction with resonator walls.

The last point is currently still a major problem in this setup due to the lack of flexibility in the adjustment of the collimation optics. This type of background signal depends on the modulation frequency of the laser and can therefore be measured by the same procedure as described in the previous section, just without any NO_2 analyte concentration and only synthetic air, with which the laser beam does not interact. The measurement results are displayed in Fig. 7.

V. Conclusion

This work shows the measurement principle of photoacoustic spectroscopy and realization in design and charakterization of a sensor setup. The connection between laser temperature, its emission spectra and the fine structure of the NO_2 absorption profile is presented. As optimal working point for this setup in regard of the PA signal obtained, 2,1A at 50% duty

cycle and 10°C mount temperature are determined. Studying the behavior of the acoustic resonator, a resonance frequency of $5110Hz$ is measured, which fits to theoretical calculations. The obtained Q factor of 20,3 is within the range of previous experiences in the laboratory with similar measurement cells and resonators. Due to the high background signal caused by stray light from poor collimation, further investigations are necessary to improve this issue and therefore enable the setup to measure towards the low concentrations called by the directives.

REFERENCES

[1] A. González Ortiz, C. Guerreiro, J. Soares, F. Antognazza, A. Gsella, M. Houssiau, A. Lükewille, E. Öztürk, Air quality in Europe 2019 report, Tech.Rep., Luxembourg, doi:10.2800/822355, 2019.

[2] WHO European Centre for Environment and Health, "Review of evidence on health aspects of air pollution – REVIHAAP Project," World Health Organization, 2013. [Online]. Available: http://www.euro.who.int/__data/assets/pdf_file/0004/193108/REVIHAAP-Final-technical-report-final-version.pdf. Last accessed: June. 2, 2020.

[3] World Health Organisation, Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide Global update 2005, WHO Press, Geneva, ISBN 92 890 2192 6, 2006.

[4] Umweltbundesamt, Informationen zu den Luftschadstoffen Stickstoffdioxid (NO2) und Stickstoffoxide (NOx). [Online] Available: https://www.umweltbundesamt.de/sites/default/files/medien/370/dokumente/infoblatt_stickstoffdioxid_stickstoffoxide_0.pdf. Last accessed: June. 2 2020.

[5] A. G. Bell, "On the production and reproduction of sound by light," American Journal of Science, vol. s3-20, no. 118, pp. 305–324, 1880.

[6] E. P. C. Lai, B. L. Chan, and M. Hadjmohammadi, "Use and Applications of Photoacoustic Spectroscopy," Applied Spectroscopy Reviews, vol. 21, no. 3, pp. 179–210, 1985.

[7] X. Yin, l. Dong et al., "Sub-ppb nitrogen dioxide detection with a large linear dynamic range by use of a differential photoacoustic cell and a 3.5 W blue multimode diode laser," Sensors and Actuators B: Chemical, vol. 247, pp. 329–335, 2017, doi: 10.1016/j.snb.2017.03.058.

[8] Z. Bozóki, A. Pogány, and G. Szabó, "Photoacoustic Instruments for Practical Applications: Present, Potentials, and Future Challenges," Applied Spectroscopy Reviews, vol. 46, no. 1, pp. 1–37, 2011, doi: 10.1080/05704928.2010.520178.

[9] K. Yoshino, J. R. Esmond, and W. H. Parkinson, "High-resolution absorption cross section measurements of NO2 in the UV and visible region," Chemical Physics, vol. 221, 1-2, pp. 169–174, 1997, doi: 10.1016/S0301-0104(97)00149-3.

Evaluation of Different Hardware Platforms for Real-Time Signal Processing

Mario Aicher
OTH Regensburg
Regensburg, Germany
Laboratory for Electroacoustics
mario1.aicher@st.oth-regensburg.de
Phone: +49 941 943 1163

***Abstract*—Real-time signal processing is part of many future-oriented technologies, e.g. interference cancellation in automatic speech recognition systems and digital image processing in autonomous driving. The fields of application are wide-ranging and the goal of keeping the development time short is in the mind of many stakeholders. In this research project, different hardware platforms are evaluated and a taxonomy for the systematic selection of the best-fitting platform, including guidelines for a purposeful engineering process, is developed.**
To this end, different strategies for increasing code efficiency are suggested in this paper and verified by implementing finite impulse response (FIR) filters on a Texas Instruments C6000 digital signal processor (DSP) and an ARM Cortex-M4 microcontroller. Thus, the maximum number of floating-point filter coefficients can be increased from 110, as achieved by the baseline DSP implementation, to more than 5200 coefficients at a sampling rate of 16 kHz without directly using processor-specific assembly instructions. When using the fixed-point implementation technique, even more than 8500 filter coefficients of data type INT16 are possible. The same filter implementation on the microcontroller achieved a maximum number of 270 single-precision floating-point and 320 INT16 filter coefficients.

I. Introduction

The processing part of real-time algorithms has to be accomplished in a given time to ensure proper functionality. Different hardware platforms fulfill this task more or less successfully. If one platform needs less time for calculation and/or data transfer than another platform, additional software features (e.g. safety functions, parity checks, more complex code, etc.) can be added, or a higher sampling rate can be chosen for more accurate results. Nowadays, there are already a number of voice-controlled internet-based assistants on the market, such as products from the Amazon Echo family or Apple Siri. In order to make these products accessible to the masses, cheap and less sophisticated audio hardware devices are used. Their deficits must be remedied by using increasingly complex signal processing software, which requires additional computing power to ensure the real-time conditions [1, 2].
Digital signal processors have been widely applied for decades and still constitute the most important device on the market, whereas microcontrollers are gaining more and more in importance. State-of-the-art microcontrollers (μCs) feature high clock rates and sometimes additional signal processing extensions with an instruction set optimized for computationally intensive signal processing tasks. This is already sufficient for many real-time applications (see Fig. 1). In addition, microcontrollers are offering distinct general-purpose possibilities. Today's field-programmable gate arrays (FPGAs) may also be utilized as specific and powerful signal processing platforms. They are particularly interesting for very computationally intensive tasks that are suitable for parallel processing.
Operating with various hardware platforms may lead to the same result, but in a significantly different workload and hence extremely varying development costs. Each of these platforms has its right to exist and offers unique characteristics for the usage in different application areas.

A. Content

The aim of this project is to make a performance comparison among the different hardware platforms, namely DSPs, microcontrollers, and FPGAs. As an example application, FIR-filters are used. These algorithms are optimized with platform-specific methods to achieve the best performance. For this reason, different scheduling methods (polling and interrupts) and different implementation techniques are compared and examples – implemented on fixed- and floating-point units – are contrasted. Moreover, different optimized DSP-libraries are used to reach a maximum of parallelization and therefore speed. The possibilities of FPGA hardware realization of the filter algorithms are demonstrated and evaluated, too. Apart from that, the needed development effort and the costs for the hardware and software are considered. The results are used to evolve guidelines and a taxonomy for various applications. This paper focuses on a comparison of DSPs and microcontrollers and how code efficiency can be increased on these platforms.

B. Digital Signal Processor - Introduction

Over the last decades, DSPs experienced a rapid development. The fixed-point units that were prevalent in earlier DSP models were – in some devices – expanded by floating-point units. Still, the fixed-point units are cheaper, faster

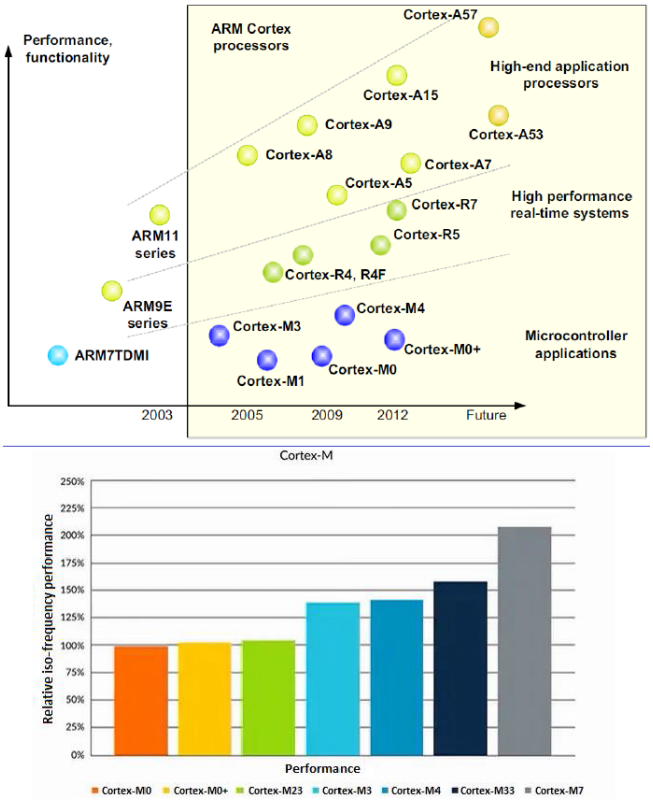


Fig. 1. Diversity and evolution of processor products from the Cortex processor family (above) [3] and performance comparison of Cortex-M processors (below) [4]

(but more complex to realize), and can deal with most of the technological conditions, which makes them widespread devices. A very important technical feature of DSPs is the MAC-operation (multiply and accumulate), which is predominantly applied in filter algorithms and allows to perform both instructions within one clock cycle. Furthermore, to reach the best performance, manufacturers often offer their customers free libraries of optimized and easy-to-use intrinsic functions that are programmed in assembler (see Fig. 2). As object of evaluation, the TMS320C6748 DSP from Texas Instruments is used.

```
int sadd(int a, int b)
{
        int result;
        result = a + b;
        if (((a ^ b) & 0x80000000) == 0)
        {
                if ((result ^ a) & 0x80000000)
                {
                        result = (a < 0) ? 0x80000000 : 0x7fffffff;
                }
        }
        return (result);
}
```

Technique	sadd() w/ -o3	_sadd() intrinsic operation
CPU Cycle Count	18	2

Fig. 2. Number of CPU Cycles: Intrinsic versus hand-coded C function [5]

C. Microcontroller - Introduction

Nowadays, microcontrollers often include special DSP units to deal with particular signal processing demands, e.g. performing MAC instructions very efficiently and offering fast interfaces to transfer the data to and from the audio codec. The Cortex-M4 cores include an optimized DSP unit and a floating-point unit, while being a cheap alternative to other platforms. Moreover, the usage of SIMD (single instruction multiple data) instructions and intrinsic functions (optimized assembler functions that can be used directly in the C code) is possible [3, 6, 7]. Fig. 3 shows the progress of the Cortex-M4 compared to the Cortex-M3 without DSP extensions.

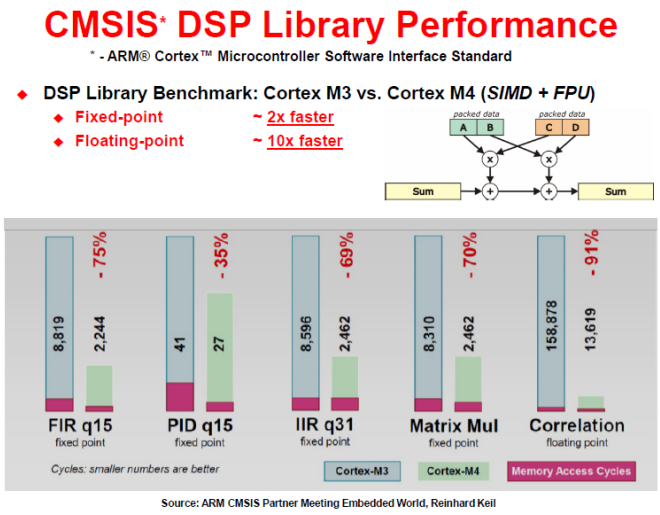


Fig. 3. Performance of Cortex-M3 and Cortex-M4 (with DSP extensions) [8]

D. Field-Programmable Gate Array - Introduction

FPGAs (programmable hardware devices) are very versatile and offer the ability to perform calculations in parallel. Especially the computation of often-used higher-order FIR-filters is a lot faster with that possibility. Therefore, FPGAs (with a high number of included DSP slices) are very powerful real-time platforms. However, the implementation of signal processing algorithms on FPGAs tends to be more costly and time-consuming than on DSPs and microcontrollers.

II. Methods

The following section describes the investigations on the respective platforms to achieve the best results in terms of processing power. For this purpose, different profiling techniques are applied to measure the performance and to improve the code. A typical, simplified optimization flow is shown in Fig. 4. The tests represent the execution time and clock cycles for a given filter order and the maximum number of filter coefficients that can be achieved in real-time for a given sampling rate. Different scheduling methods and various implementation techniques are compared in terms of performance. Moreover, different optimization levels are taken into account, and the resulting assembler code from the compiler(-optimization) is analyzed.

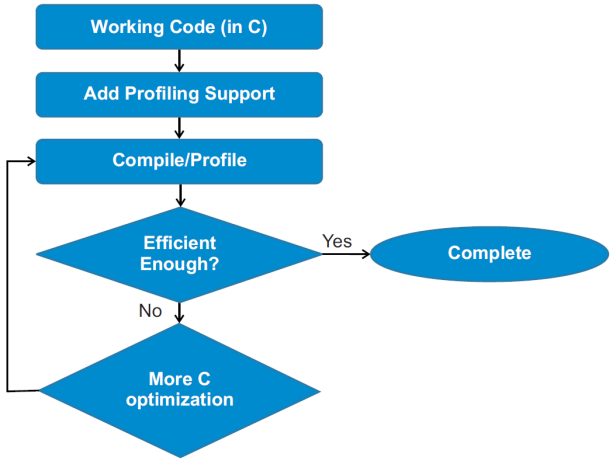


Fig. 4. How step by step code optimization works [5]

A. Digital Signal Processor - Methods

For initial investigations, the implementation takes place in C and the influence of compiler optimization and intrinsic functions is examined. The software is developed with the Code Composer Studio environment from Texas Instruments and the associated C6000 compiler. In the first steps the two scheduling methods polling and interrupt service routine (ISR), together with an FIR-filter, are implemented and contrasted. In addition to that, the brute-force filtering and circular buffering method are realized (see Fig. 5).

Brute-force filtering in real-time signal processing describes the shifting of filter data by one sample, before a new sample arrives. The process of shifting the individual samples forward takes some time, especially with filters of very high order. With the help of a circular buffer, the time consuming process of shifting all the samples is bypassed. However, the indexing of the samples in the circular buffer requires additional computations when implemented in C, since the hardware support for circular buffering cannot be utilized. Both fixed- and floating-point realizations are implemented.

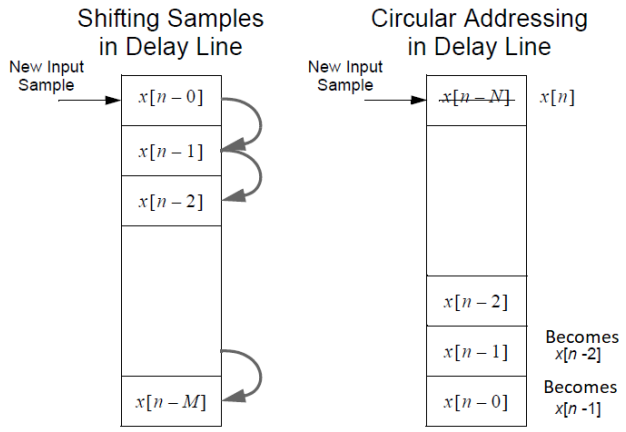


Fig. 5. Brute-force (left) and circular buffer (right) attempt [9]

B. Microcontroller - Methods

The investigations on the DSP are expanded to the microcontroller. Similar filter algorithms are implemented and the performance before and after optimization is measured. The software is developed in Keil μVision using the ARM compiler. The microcontroller includes a single-precision floating-point unit [8]. Calculations with data type double are not benefitting from that additional hardware and still take a lot of time.

C. Field-Programmable Gate Array - Methods

The FIR-filter is implemented in fixed-point for hardware realization. Due to the usage of the transposed direct form, each filter coefficient is allocated to one multiply adder block [10], that is able to perform the MAC operation. The multiply adder blocks can be calculated simultaneously and provide the result after just one clock cycle (see Fig. 6, example implementation). When using the multiply adder block from the IP catalog, the DSP48 slices are instantiated.

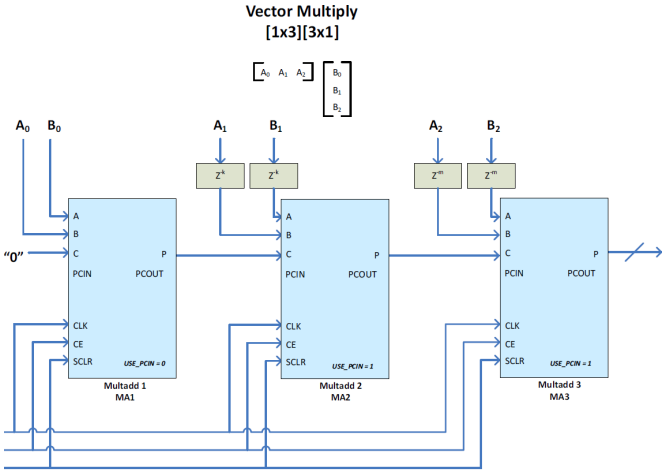


Fig. 6. Vector multiply - multiple DSP slice implementation [10]

III. Results

The following section sums up the results from the DSP and microcontroller implementations and contrasts them.

A. Digital Signal Processor - Results

In Fig. 7, the influence of compiler optimization is shown for the different scheduling methods. In the upper diagram, no optimization is applied at all, whereas in the lower diagram, the highest compiler optimization is applied.

The maximum filter order in the ISR circular buffer version is significantly lower than in the ISR brute-force version (see figure 8, below), since the compiler is not capable of ideally optimizing the circular buffer in the ISR. A look at the assembler code confirms this thesis. The implementation of the circular buffer in assembler would solve this problem, since the optimization itself is thus inserted.

The execution times in ISR mode for the DSP are significantly shorter than in polling mode (see Fig. 8, upper panel).

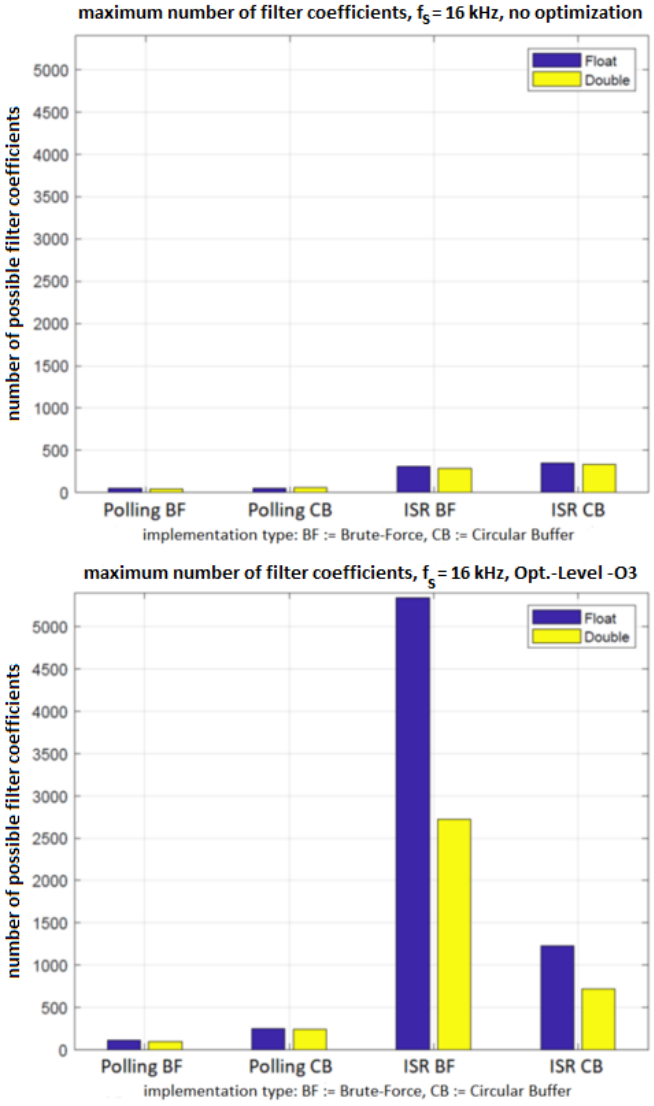


Fig. 7. Comparison of the brute-force and circular buffer implementation in polling and ISR operation without (above) and with compiler optimization (below) in relation to the maximum possible filter order [11]

In addition the execution time, when using the data type float is shorter in comparison to the data type double (results from Fig. 7).

B. Microcontroller - Results

In Fig. 9, the impact of the single-precision floating-point unit is shown. With the DSP, the execution time for single- and double-precision floating-point data and therefore the number of filter coefficients were approximately in the same order of magnitude (see Fig. 7). Instead, the gap between the maximum number of filter coefficients for the microcontroller is significantly larger for both data types, due to the fact that only single-precision floating-point data is supported by the floating-point unit.

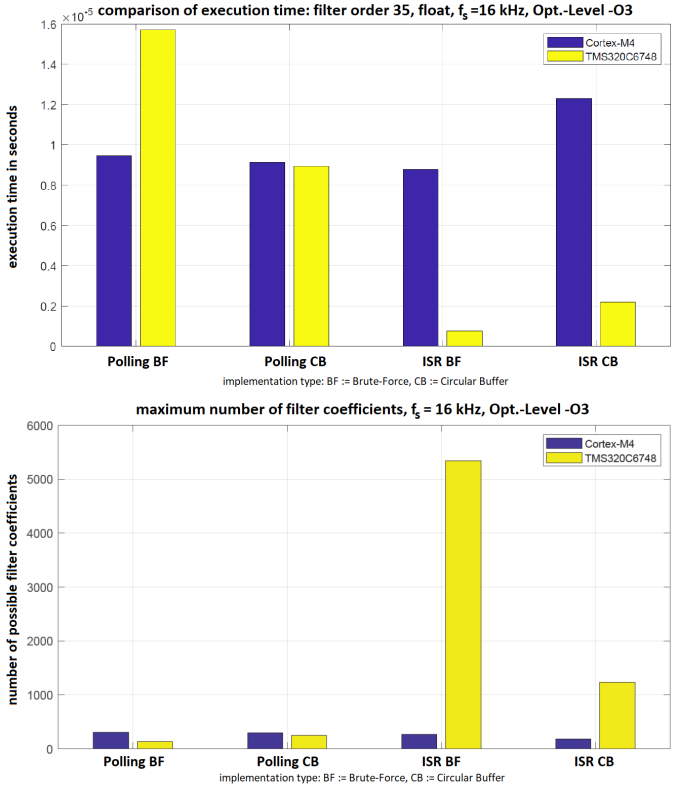


Fig. 8. Comparison of DSP and μC in relation to execution time (above) and maximum number of filter coefficients (below) with data type float [11]

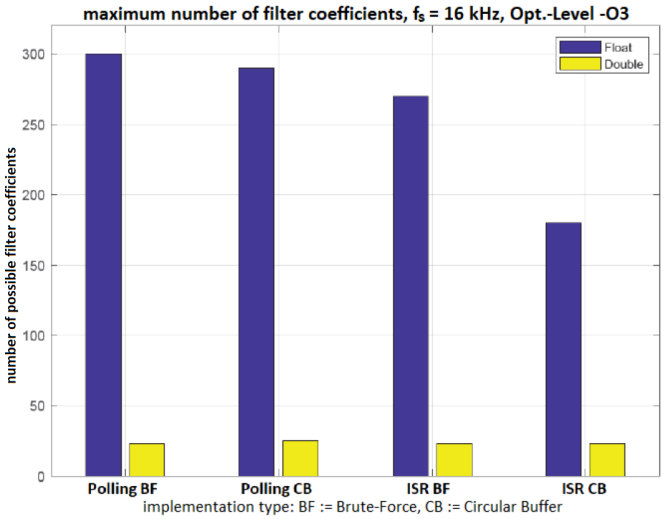


Fig. 9. Maximum number of single- and double-precision floating-point coefficients [11]

Figure 8 shows the potential of DSP optimization. When using the interrupt service routine, the number of possible coefficients in the DSP increases significantly. In contrast, the microcontroller in ISR operation in combination with the brute-force method is by far not that effective.
Fig. 10 summarizes the knowledge gained about the hardware

platforms DSP and μC. The computing speed of the DSP is clearly superior to that of the μC, as expected. Both platforms achieve the best performance when using the brute-force variant together with fixed-point implementation (data type INT16, see Fig. 10). The maximum number of possible filter coefficients for data type float (32 bit) and INT16 (16 bit) is very similar, when using the μC, despite the data size is different and floating-point calculations are more time consuming in comparison to the fixed-point calculations. This once again results from the single-precision floating-point unit.

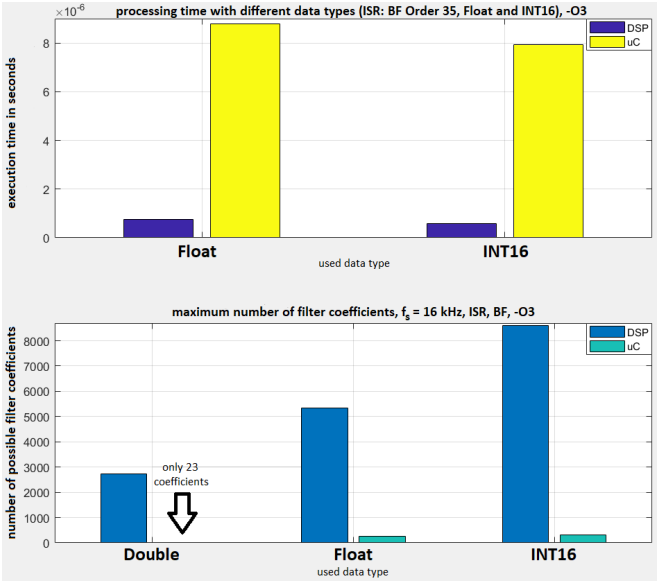


Fig. 10. Comparison of execution time (above) and maximum number of filter coefficients (below) with different data types for DSP and μC [11]

IV. Discussion

With comparatively little effort, the number of filter coefficients could be increased and the processing time could be crucially decreased. Intrinsic functions proved to be very powerful and simple to use. Still, to get the maximum performance, linear or C6000 assembler for the DSP is the first choice. Unfortunately, the programming effort increases as the optimization progresses (see Fig. 11). In general, DSP optimization follows the 80/20 rule, which states that 20% of the software in a typical application requires 80% of the processing time. This is particularly true for DSP applications, where a large part of the calculation time is required for the inner loops of the DSP algorithms (e.g. with high-order FIR-filters). The first step is not about how to optimize the code, but which code section to optimize. To do this, first different profiling methods should be utilized to find out where most clock cycles are needed and so-called bottlenecks occur. An important prerequisite for successful optimization is sound knowledge of the hardware architecture, the compiler used, and the algorithm implemented. Each processor and each compiler has different strengths and weaknesses, which the developer should know for the necessary optimization tasks [12, 13].

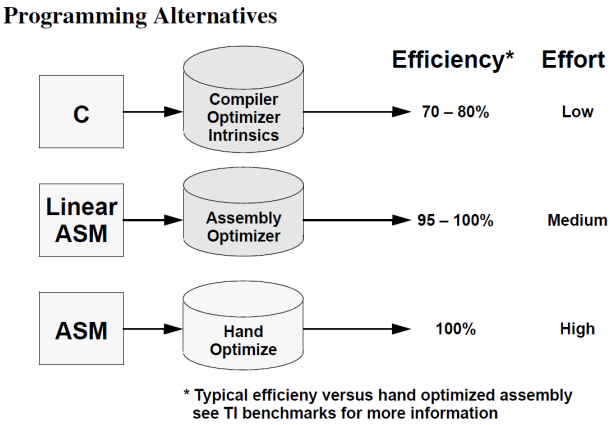


Fig. 11. Optimization level and the required effort [9]

V. Conclusion and Future Work

Concluding this contribution, DSPs are still essential in today's market, but good alternatives exist and may replace DSPs in some applications.
On the one hand, the microcontroller's computing power is already sufficient for a lot of general-purpose real-time applications. On the other hand, with the C6000 assembler, the DSP offers additional optimization opportunities for experts to achieve the maximum power for very computationally intensive situations. This is mainly used, where high efficiency or low energy consumption is more important than a short time to market.

In future work, the FPGA-design is expanded by the integration of an audio codec for real-time applications. To this end, the onboard audio codec has to be configured via an I2C (Inter-Integrated Circuit) master controller.
The former investigations on the DSP and microcontroller are expanded to the FPGA. The possibilities that arise with the use of the FPGA are very versatile and interesting for very computational-intensive and time-critical applications. We will see to what extent the computing power differs from DSP and FPGA. An overall comparison between the platforms is carried out and the guidelines are developed from this. FPGAs as part of heterogenous computer architectures are furthermore a powerful option which should be considered.
Apart from that, a complex time-critical example, an adaptive filter for interference cancellation in real-time, will be implemented on the mentioned platforms using the determined guidelines and a comparison is done. Adaptive filtering is playing an increasingly important role in modern means of communication. It is mainly applied to eliminate interference or echos, especially in the increasingly emerging so-called hands-free applications. Examples of this are hands-free telephony and voice control in vehicles and video conference systems. Applications that are currently very relevant are speech

assistants, which are based on automatic speech recognition (ASR) systems [14].

ACKNOWLEDGMENT

Thanks to OTH Regensburg and the support from the Laboratory for Electroacoustics. Special thanks go to my professor Armin Sehr and to my colleagues from the laboratory.

REFERENCES

[1] A. Stenger and R. Rabenstein, *An Acoustic Echo Canceller with Compensation of Nonlinearities*. European Signal Processing Conference, 1998.

[2] C. Huemmer, C. Hofmann, R. Maas and W. Kellermann, *Estimating Parameters of Nonlinear Systems Using the Elitist Partical Filer Based on Evolutionary Strategies*. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2018.

[3] J. Yiu, *The definitive guide to ARM Cortex-M3 and Cortex-M4 processors*. 3.ed. Amsterdam : Elsevier Newnes, 2014.

[4] J. Beningo, *Improving Embedded Security with the Armv8-M Architecture and TrustZone*. Digi-Key ELECTRONICS, November 06, 2018. [Online]. Available: https://www.digikey.com/en/articles/improving-embedded-security-with-the-armv8-m-architecture-and-trustzone.

[5] P. Yin, *Texas Instruments Incorporated: SPRABF2 Application Report: Introduction to TMS320C6000 DSP Optimization*. Version: October 2011.

[6] Texas Instruments Incorporated, *SPMS376E Tiva TM4C123GH6PM Microcontroller Datasheet*. Version: 2007.

[7] J. Yiu, *ELSEVIER : Appendices https://booksite.elsevier.com/9780124080829/appendices.php*. Version: 2014

[8] Texas Instruments Incorporated, *Getting Started with the Tiva TM4C123G LaunchPad Workshop Student Guide and Lab Manual*. Version: November 2013.

[9] M. Wickert, *Real-Time DSP ECE 5655/4655 (OMAP-L138 and TMS320C6748)*. Version: 2010.

[10] XILINX, *Multiply Adder v3.0, LogiCORE IP Product Guide, PG192)*. November 18, 2015.

[11] M. Aicher, *Evaluation verschiedener Hardware-Plattformen für die Echtzeit-Signalverarbeitung. Projektbericht 1: Vergleich von DSP und Mikrocontroller*. OTH Regensburg, October, 2019.

[12] R. Ohsana, *DSP Software Development Techniques for Embedded and Real-Time Systems*. Newnes, 2006.

[13] C. Roppel, *Grundlagen der digitalen Kommunikationstechnik: Übertragungstechnik - Signalverarbeitung - Netze ; mit 42 Tabellen und 62 Beispielen*. München : Fachbuchverl. Leipzig im Hanser Verl., 2006.

[14] J. Yang Senior Member IEEE, *MULTILAYER ADAPTATION BASED COMPLEX ECHOCANCELLATION AND VOICE ENHANCEMENT.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

Measurement of kinematics and muscle activity of athletes under stress

Lukas Reinker
Laboratory for Biomechanics
Ostbayerische Technische Hochschule
Regensburg, Germany
lukas2.reinker@st.oth-regensburg.de

***Abstract*— Muscle injuries represent almost one third of the total injuries soccer players suffer per season, which especially occur in the thigh muscles. The model of stress and athletic injury shows that an injury may follow stress responses triggered by cognitive or somatic interventions. Several studies show the effect of mental stress on upper extremities, but only few studies investigate postural changes resulting from a state of increased activation of the human body nor focus on the vulnerable lower extremities. Therefore this work analyses the influence of mental stress on the muscle activity (EMG) and the kinematic changes (motion capture) of the lower extremities using the example of highly dynamic exercise. Five male participants had to run five times a 10 meter distance as fast as possible twice, one time just focused on the physical task and the second run with an additional cognitive task. EMG signals were used to see differences in muscular behaviour, time of performance serves seeing a difference in speed, step lengths were analysed to see differences in kinematics and a NASA-TLX gave a self-assessment. The comparison of EMG values (D2/Base) showed that there was no even difference between the two sprints over all subjects. The comparison of the times of performance showed that two of the subjects were slower, two were faster and one was equally fast in the second run. The change of left and right step lengths over the sprint showed slight differences between both sprints, which seems to be caused by varying starting and turning points. Even the form of self-assessment provides the information that the subjects were mentally challenged by the additional task but did not fell to have any difference in the physical demand and performance. As the number of participants is small, it is difficult to give a valid statement. The stressor does not seem to be effective. For future work the number of subjects needs to be increased and other validated stressors must be used. In total the outcome of this study under the current circumstances does not show any correlation concerning the additional mental task and the physical demand they tried to fulfil the sprints with.**

Keywords—EMG, stress, motion capture, highly dynamic exercises, kinematics

I. INTRODUCTION

The influence of stress on sports injuries has been demonstrated in certain empirical studies. Stress factors affect human performance in various ways, both positive and negative [1]. There are different kinds of stress classified as emotional, cognitive and physical stress [2]. Cognitive stress is the most common stressor involved when humans try to accomplish real-life tasks because most jobs demand the coordination of multifaceted task aspects [3]. Andersen and Williams (1998) developed a model based on stress theory which can be seen in Fig. 1. It shows that an injury may follow stress responses influenced by personality, history of stressors and coping resources and triggered by cognitive or somatic interventions [4].

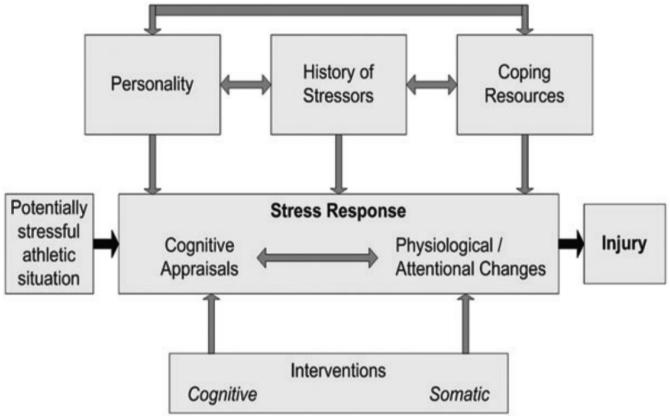


Fig. 1: The model of stress and athletic injury. Factors which may cause stress response are shown. [4]

Muscle injuries are very common to occur to athletes in highly dynamic sports. They represent almost one third of the total injuries soccer players suffer per season. Specifically, these injuries occur in the lower limbs (92%) and concern the thigh muscles, where 37% apply to the hamstrings and 19% to the quadriceps [5]. Participating in such competitive sports as soccer sets high demands on athletes' physical skills. Consequently, injury frequency is rather high [6]. In the past years, there have already been studies which showed the effect of mental stress on upper extremities. Higuchi et al. (2002) figured out that, under stress, movement strategies tend to lead to more constrained trajectories in a computer-simulated batting task., as is seen under conditions of high accuracy demand, even though the difficulty of the task did not change [7]. This study mainly concentrates on the general performance and does not investigate possible changes in the muscle activities and recruitment during stressful situations. These activities could provide insights if the muscles react in a different way under stress and if the potential risk for injuries is higher. Van Loon et al. (2001) had a look at the changes in limb stiffness under conditions of mental stress. Two experiments showed a more precise performance of the tasks by increasing the limb stiffness. Nevertheless no differences in Electromyography (EMG) activity were observed [8]. That contradicts with the findings of Lacquaniti et al. (1991). They argued that reflex coactivation results in a transient increase in joint stiffness. Antagonist muscles produce joint torques with opposite signs but cooperate to increase joint angular stiffness [9]. But surprisingly few studies investigate postural changes resulting from a state of increased activation of the human body nor focus on the vulnerable lower extremities. Furthermore solely aiming tasks are the main topic of previous studies, not high intense motions

as sprints or fast changing directions during which most muscle injuries occur.

Hence, the present work analyses the influence of mental stress on the muscle activity (EMG) and the kinematic changes (motion capture) of the lower extremities using the example of highly dynamic soccer related movements.

II. Materials and Methods

Five male participants between 21 and 25 years and an average workout time of 1 to 12 hours per week took part in this study. They had to fulfil three different exercises with and without external stressors. Anthropometric data was notated for further processing of kinematic measurements.

	Anthropometric data		
Subjects	**Age** [Years]	**Height** [mm]	**Weight** [kg]
	23 ± 2	1816 ± 38	76 ± 5

Tab. 1: Anthropometric data of the five male subjects

A. Motion Capture

Kinematic data was recorded with a motion capture system (MVN Link, Xsens Technologies B.V., Netherlands). The system uses 17 sensors which are fitted on the body with adjustable straps and a Lycra suit. This allows accurate data recording with an output rate of 240 Hz. In this study motion capture data was used to figure out the exact time of exercise and for further simulation purposes. Step lengths of each subject were evaluated.

B. Electromygraphy (EMG)

For measuring muscle activity 10 surface EMG-sensors (Trigno EMG, Delsys Inc., UK) have been attached to the main muscles in the participants' thighs according to the scheme which can be seen in Figure 2. These recommendations for electrode application are working with an anatomical landmark system which is based on dominant bone areas and prominences or other structures that can easily be palpated [10].

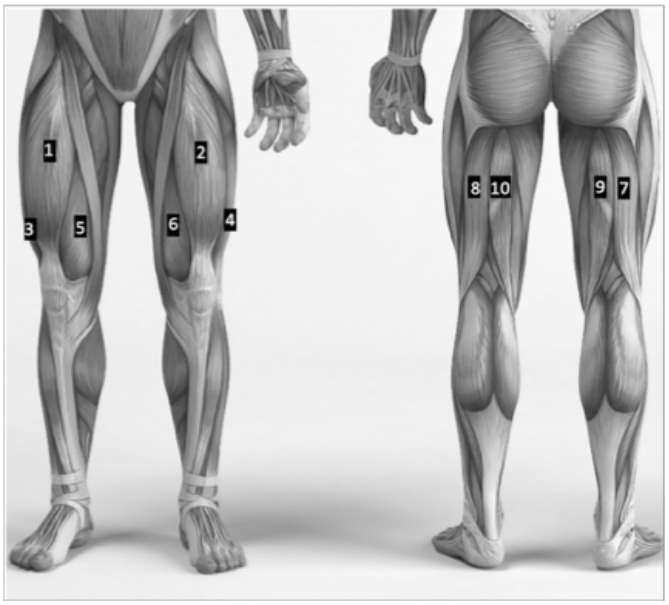


Fig. 2: Placement of EMG sensors according to SENIAM project [10]

The sensors were positioned in the same way on the left and right leg at the suggested spot of M. rectus femoris (Sensor 1, Sensor 2), M. vastus lateralis (3, 4), M. vastus medialis (5, 6), M. biceps femoris (7, 8) and M. semitendinosus (9, 10). Normalization is an important component of the electromyography process to enable valid and reliable interpretations and comparisons of muscle activity. Therefore, the subjects were instructed to perform a maximum voluntary contraction (MVC) task for the anterior and posterior groups of muscles. For muscles concerning the extension of the leg, the participants had to push their legs away from load cells fixed to their heel. For muscles concerning the flexion of the leg, the participants had to push their legs against a load cells fixed to their heel. Each movement for each leg was repeated twice with a short break of one minute between the repetitions to gain recovery of the muscles.

C. Exercises

The first task which had to be performed by the subjects was a knee-flexion at maximum speed while lying prone on a divan bed to get data of biceps femoris and semitendinosus. They were given acoustic signals when they had to do this task. A total of five repetitions was recorded. A knee-extension sitting upright with maximum speed helps receiving data of their antagonists rectus femoris, vastus lateralis and vastus medialis. The procedure resembles the way knee-flexion was measured. To capture the interaction between agonists and antagonists the participants had to run five times between two marks (distance of 10m). This paper focuses on this highly dynamic exercise. Knee-flexion and extension will be evaluated in a later work. Artificial turf was used as floor cover to give the subjects more grip. These tasks were performed twice, starting without stressor followed by an external stressor (d2-test). The setup of the experiment can be seen in Figure 3. Two screens facing the running area are used for presenting the stressor. External factors as type of sport, sportswear and weather conditions are excluded in this work.

Fig. 3: Experimental setup for performing the sprint over 50 meters

D. External Stressor

Cognitive stress is applied by the modified d2-test. The d2-test measures selective and sustained attention. The test contains the letters d and p in combination with a different number of stripes which are shown after each other. The participant had to determine all ds with two stripes and tell the adviser the correct answer while doing the physiological task [11]. For getting an idea of the subject's demands during the experiment, a NASA-TLX, which is a form for self-assessment, was handed over after each exercise [12]. The main points of the form were mental, physical and temporal demand, performance, effort and frustration. Each question could be answered with a maximum of 100 points with 5-point steps. All questions were compared between the exercise with and without stressor to see if the subjects noticed a difference.

E. Reprocessing/Musculoskeletal model

Kinematic data was imported into the AnyBody Modelling System (AMS) and applied to a motion capture model. Foot nodes are used to calculate ground reaction forces in this model. These reaction forces will be discussed in prospective works. The foot nodes were also used for determining the step lengths during sprint.

III. Results

A. EMG

After calculating the root mean square of any EMG signal, a mean was found for each muscle. For better comparison the mean of each muscle of the run with D2 stressor was divided by the corresponding Base value. The quotients for each subject and its muscles can be seen in Fig. 4. In general, the quotients are around 100 percent. That means there is no serious different between the sprint with and without the stressor. Even if there are 20 percent more or less to the 100 percent, there is no trend for any specific muscle to be more or less on average. Subject 104 shows the highest deflections from around 0 percent to 220 percent. This is justified by issues with the EMG sensors. Two of them (Rectus Femoris Right and Vastus Medial Right) fell off during the second sprint which caused a lower mean and a smaller quotient.

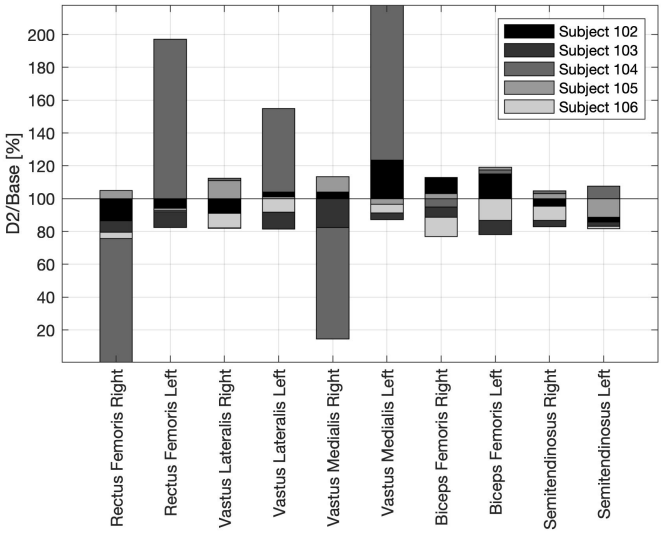


Fig. 4: Quotient (D2/EMG) of EMG signals for every subject and analysed muscle. The ordinate is scaled in percent. The single bar graphs overlay.

B. Performance

Beside muscle activity the EMG signals also provide information about the duration of the exercise, which is the time of performance. This is the time between the subjects starts moving until it stands still again. The times of performance for each subject can be found in Fig. 5.

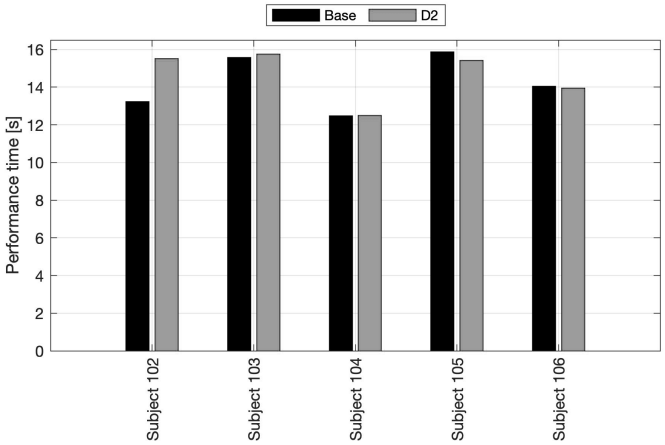


Fig. 5: Times of performance of each sprint without **(black)** and with stressor **(grey)** for every subject

Except Subject 102 with a delta in performance of 2.3 s, there is no big difference between the two sprints of the subjects.

C. Kinematics

Contact with the ground was basis for calculating step length. The norm of the vector between two steps offered the value. Figure 6 shown the changing step lengths during sprint with and without stressor for Subject 103. The abscissa depicts the number of the step. In the first subplot the difference between Base and D2 of the left foot are shown, the second one shows the difference of the right foot. The course of the graphs is similar. Because the subject started running with different starting position the graph of the base sprint looks a bit delayed.

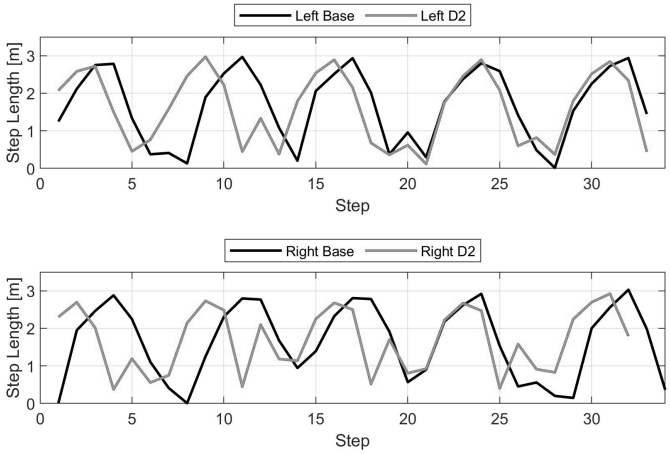


Fig. 6: Comparison of changing left **(top)** and right **(bottom)** step lengths during sprint with and without stressor for Subject 103. The ordinate is scaled in meters,

Peaks are showing the maximum step length of each segment. These peaks were averaged for each subject, sprint and foot and outlined in Table 2. There are slight differences between the runs with and without stressor but devoid of consistent trend.

Subject Nr.	max step length [m]			
	Base		D2	
	Left	Right	Left	Right
Subject 102	2.87	3.01	2.72	2.73
Subject 103	2.88	2.74	2.86	2.74
Subject 104	2.49	2.63	2.45	2.64
Subject 105	2.63	2.56	2.82	2.65
Subject 106	3.02	2.84	2.90	2.96

Tab. 2: Averaged maximum step length for each foot and sprint per subject over 50 m

D. NASA-TLX

The self-assessment according to the NASA-TLX showed in almost every category a higher score for the sprint with stressor than without. In average the highest delta of 39 points can be seen in the mental demand. The physical demand is 4 points less with stressor. The temporal demand reaches a delta of 24 points. The performance with stressor is 4 points higher than without, the effort is equal and the frustration is 16 points higher with stressor.

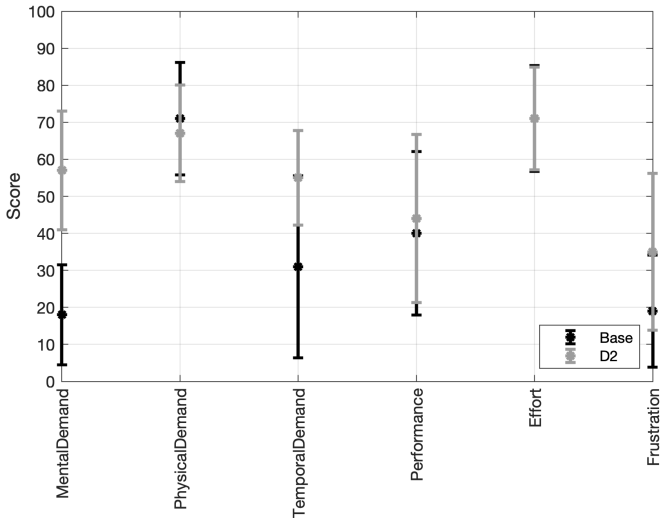


Fig. 7: Reached score with standard deviation of NASA-TLX assessed by the subjects for each sprint. Abscissa shows all assessed categories

IV. Discussion

The aim of this study was to investigate the influence of mental stress on muscle activity and kinematic changes of the lower extremities using the example of highly dynamic soccer related movements. Two sprints, one time without stressor and one time with stressor (d2), over 50 metres with change of direction after every 10 metres, were the basis of the data.

Down to the present day six subjects have taken part in this study. After measuring the first subject some main adjustments had to be done to the measurement environment. The result of this was that only five subjects offered valid data. That makes the size of the experiment very small to give solidified statements. Nevertheless, every kind of collected data was analysed to find indications for an influence of the stressor.

Having a total of 27 sensors on the subject's body, with even 16 of them partly wired on the lower extremities, caused some issues. Some of the EMG sensors fell off during the sprint which led to lost insights in the corresponding muscles. After the Subject 104 the sensors were additionally fixed by tape, which kept the sensors at their place. Anyway, the available results of the EMG data showed no even difference between the two sprints. Neither one subject showed an equal deviation over all muscles to the first sprint nor one specific muscle showed an equal deviation over all subjects. When looking back to previous studies, differences in measured EMG signals were found after applying a stressor. On top of that injuries were more likely to occur under stress. J. Ekstrand and M. Hägglund showed that the muscle health is affected by stress, so that in this study differences in the EMG signal should be seen, which come from higher performances or a higher muscle tonus affecting antagonists during the tasks. That leads to the conclusion that either simple sprinting task may not be affected by stressors or the kind of stressor is not effective.

Also, the time of performance did not make a uniform statement possible. Two of the subjects showed an increased time of performance, one subject was as fast as in its first sprint and two of the subjects showed a decreased time of performance in the second sprint with stressor. Having a look at the step lengths shows that there are slight differences between both sprints, but that does not seem to be caused by mental stress but rather by different starting positions and turning points. Though based on the small number of participants, currently no clear predication can be given, if the difference in the graph of step lengths comes from the stressor. The results of the times of performance and step lengths as they are, support the assumption that the stressor is not effective.

The self-assessment NASA-TLX provides the information that the subjects were mentally challenged by the additional task and got frustrated by not ticking every answer right. They did not fell to have any difference in the physical demand and performance nor the effort they gave. The kind of mental demand caused by the modified D2 is obviously not enough to stress the physical behaviour of the subjects.

Next to the pending analysis of the extension on flexion of the knee, it must be thought about validating and including other types of stressors in future work. Even the number of subjects needs to be increased to be able to validate outliers in the results. For gaining more insights different viewing conditions and investigation methods may be considered.

In total the outcome of this study under the current circumstances and viewing conditions does not show any correlation concerning the additional mental task and the physical demand they tried to fulfil the two sprints with. The modified d2 test serving as stressor does not cause a difference in the physical and muscular way of sprinting.

References

[1] S. Fisher, Stress and strategy, London: Erlbaum, 1986.

[2] J. J. Adam and P. C. W. Van Wieringen, "Worry and emotionality: Its influence of the performance of a throwing task," *International Journal of Sport Psychology,* vol. 19, pp. 211-225, 1988.

[3] A. W. A. Van Gemmert and G. P. Van Galen, "Stress, Neuromotor Noise, and Human Performance: A Theoretical Perspective," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 23, no. 5, pp. 1299-113, 1997.

[4] M. B. Andersen and J. M. Williams, "Psykosocial antecedents of sport injury: Review and critique of the stress and injury model," *Journal of Applied Sport Psychology,* vol. 10, pp. 5-25, 1998.

[5] J. Ekstrand, M. Hägglund and M. Waldén, "Epidemiology of muscle injuries in professional football (soccer)," *The American journal of sports medicine,* vol. 39, no. 6, p. 1226–1232, 2011.

[6] U. Johnson and A. Ivarsson, "Psychological predictors of sport injuries among junior soccer players," *Scandinavian journal of medicine & science in sports,* vol. 21, no. 1, pp. 129-136, 2001.

[7] T. Higuchi, K. Imanaka and T. Hatayama, "Freezing degrees of freedom under stress: Kinematic evidence of constrained movement strategies," *Human Movement Science,* vol. 21, pp. 831-846, 2002.

[8] E. M. van Loon, R. S. W. Masters, C. Ring and D. McIntyre, "Changes in limb stiffness under conditions of mental stress," *Journal of Motor Behavior,* vol. 33, no. 2, pp. 153-164, 2001.

[9] F. Lacquaniti, N. A. Borghese and M. Carozzo, "Transient reversal of the stretch reflex in human arm muscles," *Journal of Neurophysiology,* vol. 66, pp. 939-954, 1991.

[10] Seniam, "Recommendations for sensor locations on individual muscles [Online]," 2019. [Online]. Available: http://www.seniam.org. [Accessed 15 03 2020].

[11] R. Brickenkamp, Test d2 - Revision : Aufmerksamkeits- und Konzentrationstest, Göttingen: Hogref, 2010.

[12] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Human Mental Workload,* vol. 52, pp. 139-183, 1988.

Development of Automatized Procedures for the Generation of Complete Measurement Time Series

Ludwig Brey
Ostbayerische Technische Hochschule Regensburg
Forschungsstelle für Energienetze und Energiespeicher (FENES)
Seybothstr. 2, 93053 Regensburg, Deutschland
Email: ludwig.brey@st.oth-regensburg.de

***Abstract*—As part of the project *neos* (Netzentwicklungs-offensive Strom) a general selection procedure for control algorithms for regulated distribution transformers is composed based on different technical and economical evaluation criteria. This procedure is intended to support the grid operator like a guideline to classify his grids and to achieve a more effective usage of regulated distribution transformers by using the optimal control concept. A simulation model is being developed to investigate various control algorithms. Time series of a whole year are stored in a resolution of one minute for consumers and producers of the respective grid model in order to simulate the network behavior in a realistic manner. The used data are raised by various measurements on households, photovoltaic systems and companies in the trade, commerce and service sector at the low-voltage level as well as some measurements of different nodes at the medium-voltage level. Due to measurement errors and failures, it is sometimes necessary to replace missing data points in the time series with different interpolation methods or to replace them completely with comparable data sets. This paper presents an exemplary selection of the developed solution methods, which are implemented in an automated process to complete and adjust the real measurement data sets.**

Keywords—electrical distribution grid; emulation of distribution grid behavior; automated preparation of measurement time series

I. INTRODUCTION

Since the "Act on the Priority of Renewable Energies" (EEG) came into force in 2000, Germany's energy supply has undergone a major restructuring. As a result of the decided measures of the agreement, a steady expansion of generation plants has been recorded since then [1]. In addition, new types of consumers will be added in the future because of the upcoming electro mobility, which will consume large amounts of electric power for a limited period. The still increasing volatile feed-in of decentralized generation plants as well as the erratic consumer behavior result in bidirectional and strongly varying load flows in the distribution grid. In order to keep the resulting voltage fluctuations within the permissible tolerance limits in accordance with DIN EN 50160 [2], grid reinforcements are required to solve the voltage problems so the new supply structure can be realized. One possible grid reinforcement option for solving this task are regulated distribution transformers.

II. MOTIVATION AND OBJECTIVE

For regulated distribution transformers are various control algorithms available, which differ in their technical complexity and the associated costs. In accordance with the technical capabilities of the individual control algorithm, different integration potentials can be achieved depending on the area of application. The unequal nature of low-voltage grids and their different future development in the sectors load, feed-in and electro mobility makes the selection process of a suitable control algorithm for the best possible utilization of the existing grid capacities difficult. Therefore, the selection of the optimal control algorithm for a particular low-voltage grid results in a high planning effort for network planning.

As part of the research project *neos* ("Netzentwicklungs-offensive Strom") a selection of control algorithms for regulated distribution transformers will be investigated based on technical evaluation criteria for various low-voltage grid structures. The aim is to design a general selection procedure for control algorithms for regulated distribution transformers by using network parameters. This planning assistance makes the use of regulated distribution transformers technically more effective and helps the grid planer to evaluate the measure compared to other grid expansion options in an easier way. Due to the urgency of the expansion measures in practice, the aim is to accelerate the decision-making process in grid planning.

Several simulations investigate the selected control algorithms of regulated distribution transformers for different low-voltage grid structures. Real measurement time series are stored in the simulation grid models for a realistic reproduction of the grid behavior. Due to measurement errors and failures, it is necessary to replace missing data points in the real measurement time series by appropriate interpolation methods or to replace them completely by comparable data sets. Furthermore, in view of the simulation requirements, a further effort is required to make the available data sets usable for the simulations. The generation of complete measurement time series as well as their adaptation to the set conditions of the simulation is carried out by generally valid and automated methods. The focus of this paper is not the implementation of these automated processes, but the presentation of the developed solution concepts for the treatment of the different problems of the used measurement data sets. At the beginning an explanation of the simulation model settings and the associated requirements

for the measurement time series is given. In the following, the processing of the data sets for modeling the medium voltage fluctuations and the photovoltaic feed-in is described exemplarily.

III. Framework Conditions of the Simultion Model

As already mentioned in the introduction, a selection of control algorithms for regulated distribution transformers will be investigated with regard to their suitability for different low-voltage grid structures. The simulations are carried out with the power grid calculation software *PowerFactory* from *DIgSILENT*. In order to generate a large number of realistic and dimensioning relevant grid conditions, time series of one year, which are used to emulate:

- voltage fluctuations at different medium-voltage nodes,
- photovoltaic plants in the low-voltage level,
- household loads,
- commercial, trade and service companies in the low-voltage level and
- Charging stations for electric mobility

are included in the simulation model. Depending on the intended objectives of grid planning, various parameters and criteria are used to analyze and evaluate a control algorithm for regulated distribution transformers. Among other aspects, the focus is on determining the number of switching operations performed by the regulated distribution transformer within the observed simulation period. The control behavior of a regulated distribution transformer is determined by various design and setting parameters, which are summarized in [3]. If the regulated voltage violates the switching limits, the on-load tap-changer will not initiate a changeover until the delay time has passed. For more details on the switching process see [3]. The chosen delay time of the transformer controller plays a significant role for the observation of steps within a specified time interval. According to [4] and [5], delay times range from at least 10 s to 90 s. In order to simulate voltage peaks and drops in the simulation model which are within the range of the delay time, a corresponding choice of the resolution accuracy of the simulation time steps and the time series is necessary. Due to the fact that one load flow calculation is performed per simulation time step, the total duration of the simulation correlates directly proportional to the resolution accuracy of the simulation steps. Taking into account a required simulation period of a complete year as well as the consideration of different low-voltage grid structures with different development scenarios, the simulation effort has to be kept as limited as possible. In view of a typical delay time of 60 s in practice, a one-minute resolution of the simulation time steps and the time series is therefore regarded as sufficient for the investigations of the control algorithms for regulated distribution transformers.

Real measurement data with one-minute resolution is available to emulate the voltage fluctuation of medium-voltage nodes, photovoltaic systems and household loads by time series. There are no real measurement data sets with one-minute resolution to describe load profiles of different commercial, trade and service companies as well as of electro mobility charging profiles. For this reason, synthetic time series are used in these cases. The development procedures of appropriate synthetic load curves are not part of this paper and will not be discussed further in the following. See [6] and [7] for a detailed description of the procedure for generating electro mobility charging profiles.

When integrating the real measurement time series into the simulation model, attention must be paid not only to the plausibility of the measurement data to be used but also to their completeness for the chosen year 2017 in the required one-minute resolution. All existing real measurement time series have sporadic points in time as well as longer periods with missing or invalid measurement data, making it necessary to process the real measurement time series to complete time series for the usage in the simulation model.

IV. Modeling Medium-Voltage Fluctations

This section describes within the first part the measurement data used to emulate different characteristic medium-voltage variations. The second part describes the procedure for completing the original measurement data.

A. Characterization of Medium-Voltage Fluctuations

A total of nine different medium-voltage nodes are used to map different medium-voltage fluctuations. The measurements were taken within the framework of an already completed project at FENES. As already mentioned in chapter III, the evaluation of the measurements is carried out in a one-minute resolution for the year 2017.

To characterize the different medium-voltage fluctuations, the voltage distributions are analyzed by taking into account the topological conditions of the network area. Figure 1 depicts in several diagrams the voltage distributions of different medium-voltage nodes. The setpoint voltage U_{Set} of the considered grid area is 20.6 kV. Therefore, the medium-voltage voltage of 1.03 p. u. has a higher level than the nominal voltage U_N of 20.0 kV. The higher voltage level counteracts the voltage drop and reduces the occurring line losses of the distribution grid.

The blue histogram in Figure 1a shows the voltage distribution of the medium-voltage measured point *58*, which is located directly on the lower voltage side of the high-voltage/medium-voltage transformer of the substation. In order to adapt the voltage level to the various grid conditions, a continuous, load flow-dependent setpoint adjustment is done by the voltage regulator of the high-voltage/medium-voltage transformer. For this reason, a wide voltage range of 7.57 % relative to the nominal voltage U_N results at the observed node. Due to the high installed power feed-in in the grid area and the resulting feedback into the overlaying grid level, voltage values below the setpoint voltage U_{Set} of 20.6 kV increasingly occur due to the dynamic setpoint adjustment.

The red distribution in Figure 1a shows the modified original measuring time series of the measured medium-voltage node *58 (adjusted)*, in which the voltage values are corrected by the adjustments made by the dynamic setpoint control. Compared to the original data of the measuring point, the distribution is almost symmetrical about the nominal voltage U_{Set} within the

voltage values of 20.25 kV and 20.95 kV. The resulting voltage range corresponds to the control bandwidth of 3.5 % of the nominal voltage U_N of the regulation of the high-voltage/medium-voltage transformer.

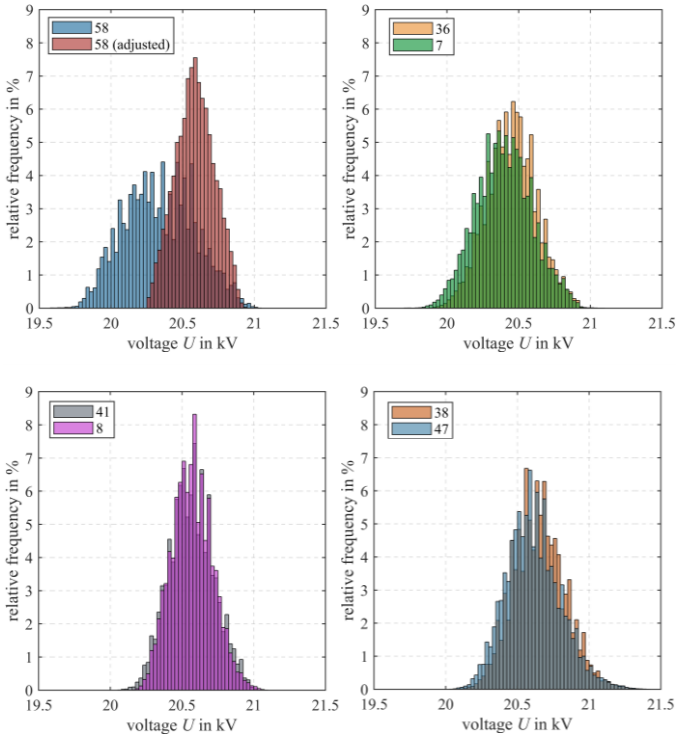


Fig. 1. Voltage distributions of the measured nodes *58* and *58 (adjusted)* with and without dynamic setpoint control at the transformer substation **(a)**, of the measured nodes *36* and *7* with short distance to the transformer substation **(b)**, of the measured nodes *41* and *8* with medium distance to the transformer substation **(c)** and of the meas-ured nodes *38* and *47* with long distance to the transformer substation **(d)** for the year 2017.

Figure 1b shows the voltage distributions for the medium-voltage nodes *36* (orange) and *7* (green). Both nodes are located at the beginning of two different lines of the observed grid area. The distributions have a similar voltage range with spans of 6.51 % and 6.71 % relative to the nominal voltage U_N. Due to the increased power consumption at both measured points, the majority of the occurring voltage values are below the setpoint voltage U_{Set}.

The voltage distributions of the measured points *41* (grey) and *8* (violet) are shown in figure 1c. The medium-voltage nodes are located in different strings with medium distance to the substation transformer. The centers of both distributions are above the setpoint voltage U_{Set} of 20.6 kV. Common cause are wind energy plants, which are installed next to the measured nodes *41* and *8*. Due to voltage-maintaining by reactive power regulation at the feed-in points, the measured nodes *41* with 5.45 % and *8* with 5.36 % of the nominal voltage U_N have narrower voltage spectrums compared to the measurements at other medium-voltage nodes.

Figure 1d shows the voltage distributions of the medium-voltage nodes *38* (orange) and *47* (blue), which are located in different strings with a great distance to the substation transformer. Both distributions are characterized by a center of gravity of the distribution above the setpoint voltage U_{Set} of 20.6 kV and a wide voltage range. With 7.06 % and 8.46 % of the nominal voltage U_N, the measured points *38* and *47* have the highest fluctuation margins of all regarded medium-voltage nodes.

B. *Adjustment and Completion of he Measurement Time Series*

Considering the requirements of the simulation model for testing the control algorithms for regulated distribution transformers as described in section III, the following problems arise in the original measurement time series for the evaluated period of the year 2017:

- sporadic points in time with missing voltage values,
- sporadic points in time with voltage values below 19.0 kV and
- long periods of missing voltage values, occurring simultaneously at all considerate medium-voltage nodes.

1) Replacement of Individual Points in Time with Missing or Implausible Measurement Values: With regard to the original measured time series of the selected medium-voltage nodes, short periods of one to three minutes with voltage values below 19 kV represent untypical outliers in the voltage curve. Voltage drops like this would lead to undesired effects when investigating the control algorithms for regulated distribution transformer. Therefore they cannot be used. The substitute values for these individual points in time are calculated by using linear interpolation approximation. The existing measured values of the two bordering points of time around the missing time interval serve as support points for the interpolation function.

2) Replacement of Large Time Periods with Missing Measurement Values: In addition to individual missing or incorrect points in time, failures of longer periods of time occur, ranging from one day to six consecutive missing days in the worst case. The cause is the failure of the superordinate measuring system. As a result, the identical time periods are missing at all measuring points considered. Due to the simultaneous failure of all measuring points in the same grid area, it is not possible to use a suitable procedure to use and incorporate the measurement data of another measuring point as a substitute for the gap. Instead, the measured values of the previous period are incorporated depending on the size of the missing time interval. If, for example, the measured values of two consecutive days are missing, the measured values of the previous two days are used, starting from the time with the last available measured value. The time of the last available measured value is always the 00:00 o'clock value of the first faulty day of a failure period due to the daily replacement. With the aim of not generating any artificial jumps in the voltage curve, an adjustment of the voltage values at the interfaces of the failure time interval is carried out when the measured values

of the replacement period are incorporated. The creation of the adjusted substitute values U_{fit} for the missing time period t_{A} is done by using equation

$$U_{\text{fit}}(t_{\text{A}}) = U_{\text{sub}}(t_{\text{A}}) \cdot g_1(t_{\text{A}}) + U_{\text{Start}} \cdot g_2(t_{\text{A}}) + U_{\text{End}} \cdot g_3(t_{\text{A}}). \quad (1)$$

The terms

$$g_1(t_{\text{A}}) = \begin{cases} \dfrac{t_{\text{A}}}{T_{\text{b}}} & \text{for } t_{\text{A}} < t_{\text{A,Start}} + T_{\text{b}} \\ 1 & \text{for } t_{\text{A,Start}} + T_{\text{b}} \le t_{\text{A}} \ge t_{\text{A,End}} - T_{\text{b}} \\ \dfrac{t_{\text{A,End}} - t_{\text{A}}}{T_{\text{b}}} & \text{for } t_{\text{A}} > t_{\text{A,End}} - T_{\text{b}} \end{cases} \quad (2)$$

and

$$g_2(t_{\text{A}}) = \begin{cases} 1 - \dfrac{t_{\text{A}}}{T_{\text{b}}} & \text{for } t_{\text{A}} \le t_{\text{A,Start}} + T_{\text{b}} \\ 0 & \text{for } t_{\text{A}} > t_{\text{A,Start}} + T_{\text{b}} \end{cases} \quad (3)$$

as well as

$$g_3(t_{\text{A}}) = \begin{cases} 0 & \text{for } t_{\text{A}} \le t_{\text{A,End}} - T_{\text{b}} \\ 1 - \dfrac{t_{\text{A,End}} - t_{\text{A}}}{T_{\text{b}}} & \text{for } t_{\text{A}} > t_{\text{A,End}} - T_{\text{b}} \end{cases} \quad (4)$$

with

$$T_{\text{b}} = 30 \text{ min} \quad (5)$$

indicate a piecewise linear weighting function for the voltage values of the replaced period U_{Sub} (t_{A}) and the real measured voltage values U_{Start} and U_{End} of the two bordering times around the failure time interval t_{A}. In order to keep the editing influence on the real measurement data as low as possible and to adequately reduce and equalize the amount of artificial voltage jumps at the interfaces, a length of 30 minutes is chosen for the adaptation time interval T_{b}. To illustrate the effect of the adjustment function according to equation 1, Figure 2 shows an exemplary time section of the time series of the measured medium-voltage node *38*.

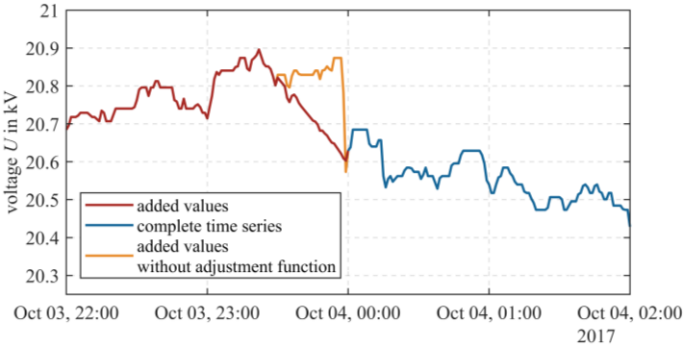


Fig. 2. Exemplary time section of the time series of the medium-voltage node *38* with adjustment of the substitute values (red) and non-adjustment of the substitute values (orange) at the intersection of a period with missing measured values

The completed time series with the adjusted substitute values U_{fit} is represented by the red curve for the missing period t_{A}. The completed time series is marked blue outside the failure period. In contrast, the orange curve indicates a completed time series without adjustment of the substitute voltage values. An artificial voltage jump would occur around 00:00 o'clock without an adjustment of the substitute voltage values. In contrast to the example shown in Figure 2, the artificial voltage jumps may be less pronounced at the transition points and may be below the level of real voltage jumps, whereby the substitute measured values would fit uncritically into the original measurement time series. The algorithm does not differentiate between these situations and applies an adjustment of the measured values in all cases.

V. Modeling Photovoltaic Feed-In

The first section of the section describes the data basis used to simulate the photovoltaic feed-in. The procedure for completing the incomplete original measurement data follows.

A. *Description of the Used Measurement Data*

When selecting an appropriate measurement time series to emulate the photovoltaic feed-in, the location of the photovoltaic system must be taken into account in addition to the requirement of a one-minute resolution of the year 2017. Next to the daily and seasonal influences photovoltaic feed-in also depend strongly on the location. To ensure that the occurring photovoltaic feed-in is sufficiently compatible with the overlaying medium-voltage fluctuation, it is practical to use the existing measurements on photovoltaic plants from the same grid area of the medium-voltage nodes. The recorded active and reactive power data (P and Q data) for the measured photovoltaic point *60* are taken from the same databank. For the later simulation to investigate the control algorithms for regulated distribution transformers, it is sufficient to consider a single photovoltaic feed-in profile for the modelled low-voltage grid. The photovoltaic park connected at the medium-voltage level has an installed capacity of 2.19 MW. For a simulation of the photovoltaic feed-in at the low-voltage level, it is necessary to normalize the P and Q time series of the photovoltaic park from the medium-voltage level to the maximum P or Q value and to scale it accordingly in the low-voltage grid model for the different plants.

B. *Adjustment and Completion of the Measurement Time Series*

In addition to the normalization of the P and Q values, the following problems have to be solved for using the measured photovoltaic time series:

- sporadic points in time with missing P and Q values and
- long periods of missing P and Q values.

1) Replacement of Individual Points in Time with Missing or Implausible Measurement Values: The original time series of the selected photovoltaic point *60* has a few time points with missing P and Q values. Up to an interval of ten consecutive points in time, a linear interpolation method is used to approximate the missing measured values. The existing P or Q measured values of the two adjacent time points around the failure period act as support points.

2) Replacement of Large Time Periods with Missing Measurement Values: For longer periods with more than ten successive missing measurement values of the original time series, a differentiation is made whether the missing data occur within the day or night hours. In both cases, the P and Q values are always replaced in pairs. If the measurement failure occurs within the day hours, the affected day is completely replaced by the measurement data of the directly preceding day. The background to this approach is the intention to maintain the seasonal trend of the photovoltaic feed-in. Consequently, a repetition of an identical PV feed-in profile is tolerated. Similar to the methodology used to complete the time series of the medium-voltage nodes, the 00:00 point of time of the first faulty day of a failure period represents the intersection to the existing original values of the time series. An adjustment of the active or reactive power values at the transitions is not necessary due to the insignificant fluctuation of the measured P and Q values during the night hours. For this reason, periods with missing measured values that occur exclusively within the night hours from 22:00 to 05:00 are replaced by the measured values of the previous night. Due to this differentiation there is no unnecessary exchange of existing measured values of relevant daily hours. Figure 3 shows a section of the normalized P time series of the measured photovoltaic point *60* to illustrate this procedure.

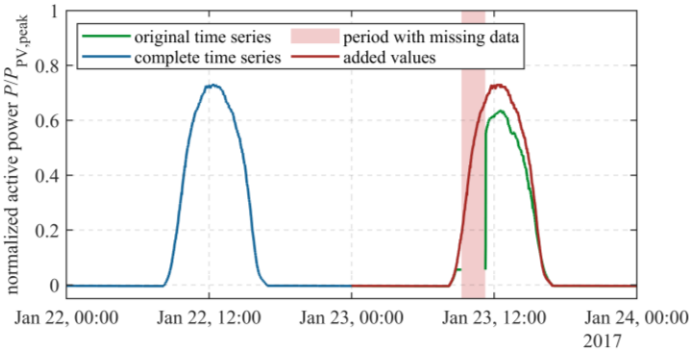


Fig. 3. Exemplary time section of the P time series of the measured photovoltaic point *60* with the missing time period (red background) of the original time series (green) and the completed time series (blue) with the substitute values (red graph)

VI. Conclusion and Next Steps

To investigate the control algorithms for regulated distribution transformers, time series of one year are stored in the simulation model to emulate the grid behavior (medium-voltage fluctuation, photovoltaic feed-in, household and commercial, trade and service companies loads). Due to measurement errors and failures it is necessary to modify the real measurement data sets. In case of one to a maximum of ten consecutive missing data points, an approximation of the missing values by interpolation is made. If the missing period is larger, it is necessary to incorporate measurement data from a comparable data set. The developed solution methods for the selection and incorporation of the substitute data for the completion of the annual time series for the year 2017 differ depending on the considered data set. Due to the modular structure of the automated procedure, all data sets are able to use different program routines, to solve common problems. An example would be the routine to determine all missing points of time of the measured time series. Furthermore, the integration of extensions to treat new issues of other data sets can be implemented in a simple way.

In the next step, the various low-voltage grid models will be supplemented with realistic temporal development scenarios for generation and consumer capacity as well as electric mobility. The results of the various simulations are assessed using a systematic evaluation methodology. Based on the evaluation results, the simplified procedure for selecting a suitable control algorithm for regulated distribution transformers is derived.

References

[1] UBA, Umweltbundesamt: Erneuerbare Energien in Deutschland. Daten zur Entwicklung im Jahr *2018*. [Online] Available: https://www.umweltbundesamt.de/publikationen/erneuerbare-energien-in-deutschland-2018. Accessed: 19.07.2019.

[2] DIN, Deutsches Institut für Normung: DIN EN 50160: Merkmale der Spannung in öffentlichen Elektrizitätsversorgungsnetzen, Berlin, 2011.

[3] L. Brey: "Kategorisierung von Regelalgorithmen für regelbare Ortsnetztransformatoren und deren Beschreibung als Regelkreise" Projektarbeit 1, Forschungsstelle für Energienetze und Energiespeicher (FENES), Ostbayrische Technische Hochschule, Regensburg, 2019. unpublished.

[4] Forum Netztechnik / Netzbetrieb im VDE (FNN): rONT - Einsatz in Netzplanung und Netz-betrieb: Technischer Hinweis, Berlin, 2016.

[5] E.-M. Königsheim: "Optimized Tap Changer Operation and Control in Distribution Grids with a High Penetration of Distributed Generation", Dissertation, Technische Universität München, München, 2017.

[6] S. Schwarz: "Generation von Ladeprofilen für Elektrofahrzeuge über Bewegungsprofile und die Analyse der Auswirkung auf die Spannungs- und Auslastungsverhältnisse in einem ausgewählten Niederspannungsnetz" Bachelorarbeit, Forschungsstelle für Energienetze und Energiespeicher (FENES), Ostbayrische Technische Hochschule, Regensburg, 2019. unpublished.

[7] F. Adler: "Entwicklung eines Modells zur Abschätzung der Lastgänge von Schnellladestationen für Elektroautos basierend auf Mobilitätsdaten" Bachelorarbeit, Forschungsstelle für Energienetze und Energiespeicher (FENES), Ostbayrische Technische Hochschule, Regensburg, 2020. unpublished.

SESSION C2

Sebastian Peller
Ultrasound beamforming with phased capacitive micromachined ultrasonic transducer arrays for the application flow rate measurement

Johannes Schächinger
Suitability of biogas plants for congestion management in distribution grids

Matthias Götz
Realistic case study for the comparison of two production processes

Andreas Arnold
Machine learning methods for creating personality profiles from data in social networks

Ultrasound beamforming with phased capacitive micromachined ultrasonic transducer arrays for the application flow rate measurement

Sebastian Peller
Faculty of Applied Natural and Cultural Sciences
OTH Regensburg
D-93053, Regensburg, Germany
sebastian.peller@st.oth-regensburg.de

***Abstract*—Capacitive micromachined ultrasonic transducers (CMUTs) offer many benefits in comparison to commonly used piezoelectric micromachined ultrasonic transducers (PMUTs). Besides the reproducibility, especially in terms of temperature stability and biocompatibility. The latter is of great importance in medical applications. For instance, CMUTs can satisfy the requirement of lead-free systems for respiratory monitoring unlike PMUTs. In recent projects, our institution developed an applicable system for flow rate measurement including a pair of CMUTs (emitter and sensor) mounted on opposing ends of a diagonal path with the fluid flowing in between. One can determine the flow rate by the principle of ultrasound runtime difference caused by the flowing medium, while alternating the roles of the CMUTs as a sender or receiver. Henceforth, a new generation of CMUTs with a larger number of cells on a single chip ensures the usability of the system in applications with flow velocities not significantly smaller than the speed of sound in the respective media. Whilst applying a certain phase shift pattern to the numerous cells of the new CMUT, one can achieve a beamforming effect to equalize the drift of the ultrasound propagation path away from the receiver. This maintains the signal level on the receiving end.**

Keywords—CMUTs, beamforming, ultrasound, flow rate measurement, phased array

I. Motivation

Many technical applications require monitoring systems for a fluid flowing through pipes. Along other measurands like the temperature or the density of the flowing medium the flow rate is very often of great importance as its manipulation is a common way to control e.g. the fuel feed of a system. Generally, the determination of a measurand comes along with a perturbation of the system. Like the simple measurement of an electrical current requires an ammeter that has a non-zero resistance and effectively reduces the current conducted by the node on which the ammeter is inserted, so you do when you place a simple Prandtl-probe into a pipe with air flowing through [1]. On the one hand you get the dynamic pressure and thereby the flow rate but on the other hand you arouse turbulences in the gas flow because of the solid probe in it. Therefore, you have manipulated the flow conditions and thus the measured value is not the same as it was before placing the probe. That is why it is desirable to establish a measurement application that does not interfere too much to the flow conditions.

II. Introduction

In terms of fluid flow measurement existing solutions usually offer either a direct determination of the wanted measurand or a turbulent-free way of obtaining it. For instance, there is a principle called laser-doppler-anemometry (LDA) which uses small particles fed into the fluid flow that periodically scatter light in an interference pattern created by the presence of two crossed laser beams [2]. According to the frequency of the detected light reflexes one can determine the speed of the particles that – if they are small enough – equals the flow rate. The big advantage of LDA is the absence of any solid probes or similar but adversely one uses an indirect way of flow rate measurement and on top of that the use of lasers results in high costs. The other way round – turbulent but direct and cheap measurement of the flow rate – is executed using the aforesaid Prandtl-probes. A better solution is to use ultrasound emitters and sensors [3]. In recent projects, our institution developed an applicable system for flow rate measurement including a pair of CMUTs (emitter and sensor) mounted on opposing ends of a diagonal path with the fluid flowing in between (fig. 1). One can determine the flow rate by the principle of ultrasound runtime difference caused by the flowing medium, while alternating the roles of the CMUTs as a sender or receiver. This method yields positive characteristics like the independency to density, temperature or viscosity of the fluid. In our latest research project, we work on the development of advanced CMUTs that are capable of beamforming. This effect is achieved by implementing several independent cells on a single CMUT-chip – a so called phased array of ultrasound emitters or sensors. Alongside with this evolution another advantage comes into account, namely the possibility to mount the whole sensorics and electronics on one printed circuit board when using a reflector in the pipe, so that long coaxial conductors are no longer necessary which leads to a better responsiveness of the CMUTs and also reduced parasitic effects on the signal-level. Furthermore, the costs of such a system can be decreased drastically. Besides, the measured

absorption of the acoustic wave intensity yields a proportionality to the gas concentration [4]. This makes it possible to completely analyze the gas flowing through a pipe – its flow rate as well as its blend and the concentration of its compounds.

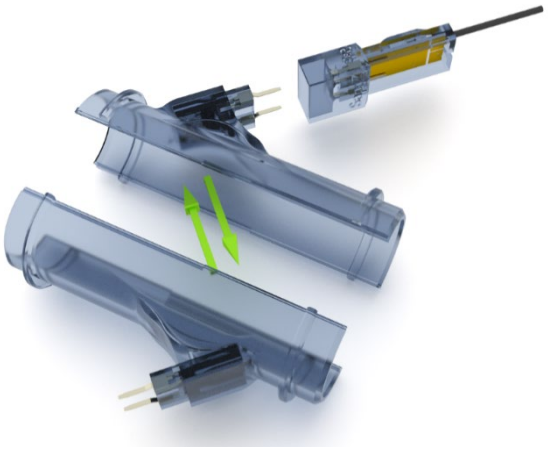


Fig. 1: Measurement setup for the determination of the flow rate by using the ultrasound runtime difference principle

III. CMUTs

Capacitive micromachined ultrasonic transducer (CMUTs) are part of an upcoming technology in the ultrasound generation. They are a potential candidate to replace commonly used piezoelectric micromachined ultrasonic transducers (PMUTs) [5]. The latter are inferior to CMUTs in terms of temperature stability and biocompatibility. These two advantages can be very helpful in medical applications since the requirement for lead-free components and the ability to sterilize them in an autoclave cannot be satisfied by PMUTs but CMUTs. Furthermore, the simple transducer element can be fabricated cheap when emerged from a microtechnological fabrication process for high volume production. This makes it interesting for the use in medical home-care products to be used several times. Additionally, CMUTs are not subject to structure-borne sound, which makes it possible to mount the transducers on a PCB with a small distance between them whilst not arousing crosstalk.

Basically, a CMUT comprises a static and a dynamic electrode [6]. More specifically, the static electrode is embedded into a cavity within a substrate (e.g. glass) and the dynamic electrode is represented by a thin (i.e. a few microns thick) silicon membrane bonded to the substrate (see fig. 2).

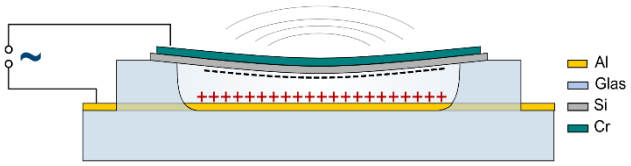


Fig. 2: Cross section of a single-cell CMUT

The CMUTs is then operated with an electrical signal applied. The voltage between the two electrodes has an optional DC as well as an AC part with the first forcing the CMUT into a state called ‘static deflection’ – this decreases the minimum distance between the electrodes and thus increases the sensitivity of the CMUT in receiving mode – and the latter causing a periodic oscillation of the membrane. That is why a CMUT is also categorized as an electro-mechanical or acusto-electrical transducer with the vibrational stimulus to the medium above the membrane (e. g. air) creating a propagating pressure oscillation (i.e. sound wave).

IV. Acoustic Beamforming with phased Arrays

Analog to a convex optical lens with given (numeric) aperture there are ‘acoustic lenses’ in the form of a multi-cell CMUT or phased CMUT array that is also capable of focusing and steering ultrasound waves like a lens does it with optical waves. The aperture of such an array has a similar meaning like it has in optical applications [7].

But first one needs to specify the different forms of phased arrays. They are commonly distinguished in 3 different main types: 1D, 1.5D and 2D arrays (see fig. 3 and 4). There are some literatures that even introduce 1.25D and 1.75D arrays for a more thorough classification [8].

A 1D array has several elements along one direction. The whole length of the row is named ‘total active aperture’ *A*, whereas *w* is the ‘width’ of a single element and e its ‘length’ or ‘elevation’. The array increment is also called ‘element pitch’ *p*. Additionally, the gap *k* between two single elements is denoted by the word ‘kerf’.

1.5D arrays show a few elements across the secondary dimension with the ‘total passive aperture’ equivalent to the active one mentioned above.

Eventually the active and passive aperture of 2D arrays are of the same magnitude with a similar number of elements in across each direction. Therefore, a distinction between active and passive plane is no longer necessary. All functionalities in terms of beamforming are available in both dimensions.

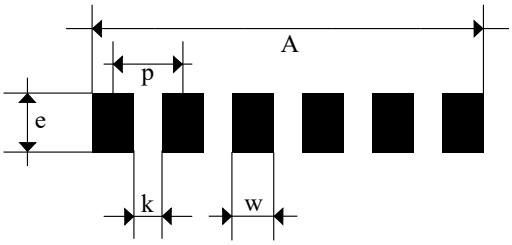


Fig. 3: 1D array with corresponding geometric terminations

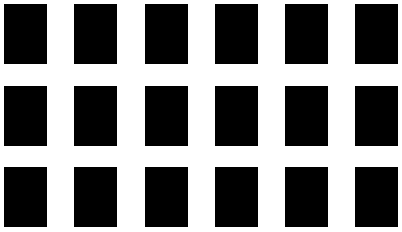


Fig. 4: 1.5D array with fewer elements across the passive plane; 2D array would be correspond to a similar number of elements across both planes (not shown in this figure)

So now we can classify the different types of phased arrays according to their beamforming capability. Basically, there are two important impacts that can be aroused on an ultrasonic beam: focusing and steering (see fig. 5).

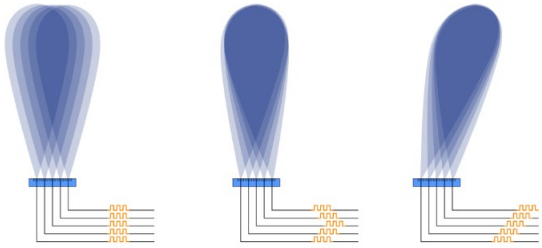


Fig. 5: 1D array of CMUTs with different phase shifts to single cells. Left: no phase shift (normal mode), middle: focusing, right: beamsteering

Focusing is already possible with little number of cells across one direction whereas steering is solely feasible with a bigger number of cells. So beamsteering in both directions is a unique selling point for 2D arrays. Another effect that can be used with full 2D arrays is apodization [9]. It is equivalent to a gaussian filter with which one can focus through holes that would create diffractive effects and thus decreasing the contrast in imaging ultrasonic applications in medical institutions.

These effects can be achieved by delaying the excitation of the single cells in a certain manner (see fig. 5). Here we come back to the analogy to optical lenses. A convex lens is thicker in the middle than at the edges. That causes different runtimes of the light through the lens and therefore leads to a curved wave front when it was planar before. These annular wave fronts meet at the focus point likewise the ultrasound gets focused when a phase shift pattern is applied to the CMUT array.

V. Air flow Measurement

The determination of the flow rate in gaseous media (e.g. air) can be done by the ultrasound runtime difference method. This principle is based on the effect that a soundwave has a relative propagation velocity – the speed of sound. Depending on how the medium moves, the soundwave may take different times to cover a fixed distance [10].

Referring to the situation in fig. 1 the ultrasound alternatingly runs in and against the flow direction resulting in two different runtimes or time of flights (TOF) (1).

$$c = c_0 \pm v \cdot cos\varphi \tag{1}$$

These runtimes are used to calculate the delta of the inverse runtimes that is used to calculate the flow rate v according to equation (2).

$$v = \frac{L}{2 \cdot cos\varphi} \cdot \Delta t_{invers} \tag{2}$$

The inverse runtime difference is defined in equation (3) as follows:

$$\Delta t_{invers} = \frac{1}{t_2} - \frac{1}{t_1} \tag{3}$$

One can see that the flow rate can be determined completely without knowledge of any other sizes but two geometric constants – the length L and the angle φ to the flow direction of the diagonal path. Not even the speed of sound comes into account but the premise of being a constant along the path which is usually satisfied in homogenous media.

The application of beamforming and -steering with phased CMUT arrays can be helpful when used in air flow measurement. So, it gets possible to circumvent the given condition that the two CMUTs must be mounted on opposing ends of a diagonal path and thus demands the presence of cavities in the pipe that cause turbulences (see fig. 6).

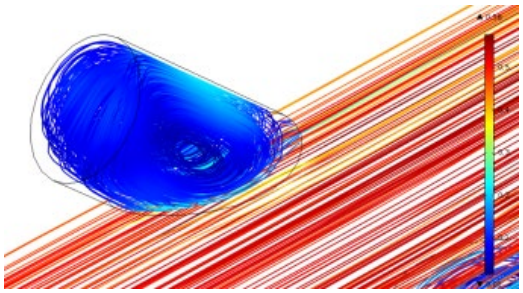


Fig. 6: Turbulences caused by the cavities for the diagonal path with a CMUT at each end

A better way is to use a reflector opposed to the now plane integrated CMUTs and to provide the ultrasonic beam to hit the receiving CMUTs by beamsteering (see fig. 7).

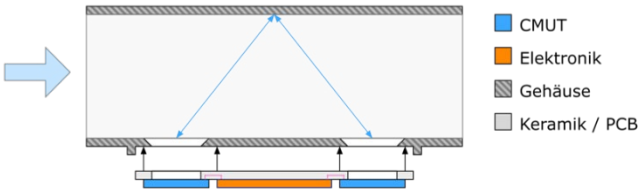


Fig. 7: Planar integration of both CMUTs into the pipe for flow rate measurement in air

According to the guideline VDI/VDE 2642 *Ultrasonic flow rate measurement of fluids in pipes under capacity flow conditions* [10] the reflector in the installation depicted in fig. 7 may simply be the inner wall of the pipe (see fig. 8).

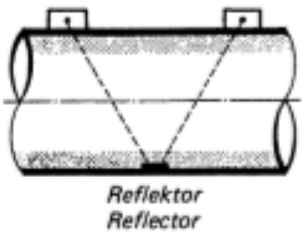


Fig. 8: Installation guideline for ultrasound flow rate measurement

Another advantage of this installation is the possibility to react to changing flow rates that would cause a drift of the ultrasound beam resulting in a reduced signal level on the sensing end. By adjusting the angle by which the ultrasound beam gets steered one can readjust the soundwave to the receiving CMUT to maintain the signal level. The data processing could be carried out using a field programmable gate array (FPGA).

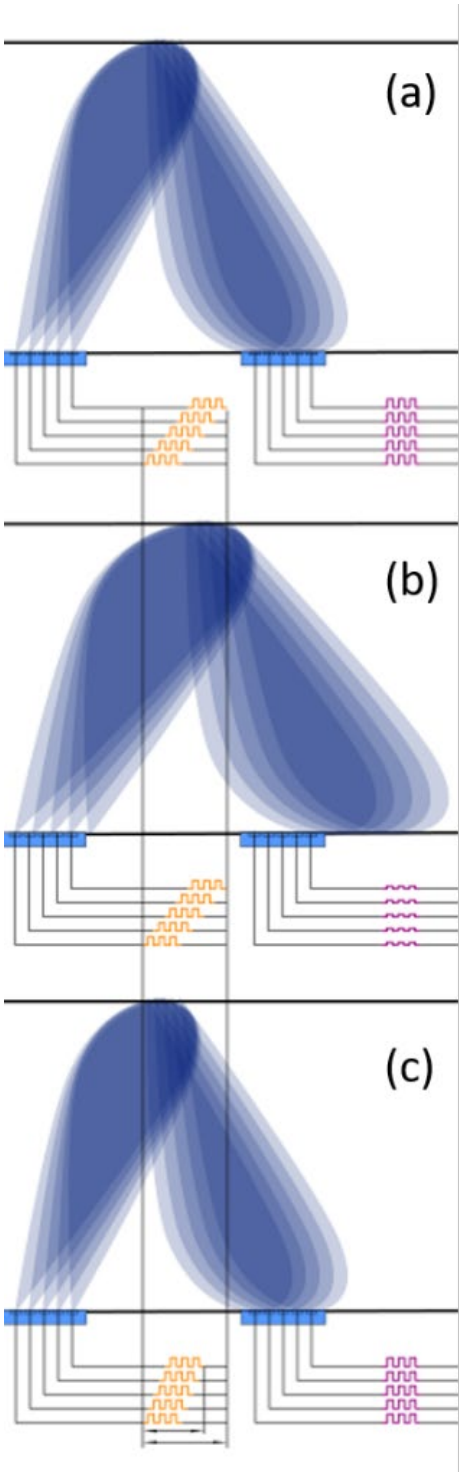


Fig. 9: Planar integration of both CMUTs into the pipe for flow rate measurement in air

This procedure is schematically portrayed in fig. 9. In case (a) the speed of sound c is much bigger than the flow rate v, thus no effect can be seen. In case (b) there is a perceivable drift because the flow rate is no longer several magnitudes smaller than the speed of sound. And in case (c) a readjustment action realigned the ultrasound beam to the receiving CMUT.

CONCLUSIONS

A cheap turbulence-free flow rate measurement system for air flows will be of huge interest in the future. A phased-CMUT-array-based solution can satisfy these requirements and is a new technological solution that currently does not exist. In the end one will be able to use this application in numerous field applications such as the spirometry in which the respiratory volume as well as the consumed oxygen can be observed [11]. Graphene-based CMUTs that are transparent may also be very interesting if one wants to perform spectroscopic measurements through the CMUT. Another application might be in the field of materials testing or in automotive in means of an air mass meter. Especially in motorsports, where common air mass meters are replaced by a restrictive α-n-map to avoid turbulences respectively flow detachments in the intake of a combustion engine, such a system could dominate future solutions.

REFERENCES

[1] Tietjens, O.G. (1934). Applied Hydro- and Aeromechanics, based on lectures of L. Prandtl, Ph.D. Dove Publications, Inc. pp. 226–239. ISBN 0-486-60375-X.

[2] Durst, F.: Principles of laser Doppler anemometers in Von Karman Inst. of Fluid Dyn. Meas. and Predictions of Complex Turbulent Flows, Vol. 1 11 p (SEE N81-15263 06-34)

[3] Lynnworth, L.C.: Ultrasonic Measurements for Process Control. Academic Press, Inc. San Diego. ISBN 0-12-460585-0.

[4] A. B. Bhatia (1985): Ultrasonic Absorption, Dover Publications Inc., New York, ISBN 0-486-64917-2

[5] S. Akhbari, A. Voie, Z. Li, B. Eovino and L. Lin, "Dual-electrode bimorph pmut arrays for handheld therapeutic medical devices," 2016 IEEE 29th International Conference on Micro Electro Mechanical Systems (MEMS), Shanghai, 2016, pp. 1102-1105. doi: 10.1109/MEMSYS.2016.7421827

[6] A. S. Savoia, G. Caliano and M. Pappalardo, "A CMUT probe for medical ultrasonography: from microfabrication to system integration," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 59, no. 6, pp. 1127-1138, June 2012, doi: 10.1109/TUFFC.2012.2303

[7] A. Chahbaz and R. Sicard: Comparitive Evaluation between Ultrasonic Phased Array and Synthetic Aperture Focusing Techniques, AIP Conference Proceedings 657, 769 (2003); doi: 10.1063/1.1570213

[8] Douglas G. Wildes: Elevation Performance of 1.25D and 1.5D Transducer Arrays, IEEE transactions on ultrasonics, ferroelectrics, and frequency control, vol. 44, no. 5, september 1997

[9] J. A. Jensen and N. B. Svendsen, "Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 39, no. 2, pp. 262-267, March 1992, doi: 10.1109/58.139123

[10] VDI/VDE 2642 (1994): Measurement of fluid flow by means of ultrasonic flowmeters in fully floated circular cross-section closed conduits, AC-Code: DE18960284

[11] Altalag, Ali (2009): Pulmonary function tests in clinical practice. Springer, London. ISBN 9781848822306

Suitability of biogas plants for congestion management in distribution grids

Johannes Schächinger
Research centre for electrical grids and energy storage systems
OTH Regensburg
Regensburg, Germany
Email: johannes.schaechinger@st.oth-regensburg.de

***Abstract*—Biogas plants have the ability to serve the electrical grid with several ancillary services. Congestion management is one of this possible contributions. For further investigation of a possible contribution of biogas plants to congestion management, simulations on a grid model of a distribution grid are done. Therefore a reference biogas plant is choosen, which is used in the simulations. Also a congestion generation strategy is defined. From the generated congestions a situation is selected for further investigation. Several simulations with different congestion situations show the ability of the reference biogas plant to serve the grid in this concrete situation. Based on the results, further necessary investigations are derived.**

***Index Terms*—biogas plant, distribution grid, ancillary services, congestion management, power flow simulations**

I. Introduction

Biogas plants have the advantage over other renewable generation plants that they have an integrated energy storage system with their gas storage tank. Thus they are in a position to shift their electricity production away from gas production. With this feature biogas plants can provide ancillary services in many different ways for the electrical grid and thus make an important contribution to gradually replacing the conventional power plant park in Germany with renewable energies. [1], [2]

Until 2012, payment due to the Renewable Energy Sources Act (EEG) only focused on the amount of produced energy. Therefore a huge incentive was given to reach high utilisation of the generators [1, p.18]. Despite efforts to support flexibilisation of biogas plants since 2012 [1, p.64], the capabilities of these units are not generally known among plant and grid operators. Thus, one aim of the research project OPTIBIOSY is to figure out situations in which biogas plants can serve the grid.

In this paper possible contributions of biogas plants to congestion management in distribution grids are discussed. Therefore simulations on a grid model of a real grid from the project partner Lechwerke Verteilnetz GmbH (LVN) are done. This requires a generation strategy for congestions, because in the original grid there are no congestions. The selection of a reference biogas plant on which the investigations will be carried out is briefly discussed.

II. Methods

To figure out the capabilites of biogas plants in congestion management, load flow simulations with real net-, generation- and consumption data are done. LVN therefore provided a grid model of an existing grid. Also time series of feed-in power for an entire year for some generation units, which are measured due to regulatory requirements, in this grid model are available. The remaining feed-in and consumption time series were created using different modelling strategies [3]. The time series are necessary for the simulations.

Defined reference biogas plants are placed in the grid model. Congestions are forced through a congestion generation strategy, which is based on the future expansion of generation plants and consumers and thus causes problems that will arise in the distribution grid in the coming years. With a simulation of an entire year single congestion events are filtered which will be inspected further via simulations of the day they occur. By manually getting congestions worse, the ability of the additional reference biogas plant to solve the congestion is examined. For this paper, only one grid model and only one congestion event is examined.

A. Used reference plant

For this paper, one biogas plant out of a pool of reference plants is used for simulations. This plant consists of two combined heat and power units, one with 250 kW, the other with 500 kW electrical power. The power coefficient is one, i.e. the heat output corresponds to the electrical output. The volume of the gas storage tank is 1500m^3. The volume is considered remaining constant over the year, environmental impacts like changes in temperature or atmosphere pressure will receive no consideration. A tank with 100 t water is used as a heat storage tank. The operating temperature lies between $50°C$ and $90°C$. The biogas plant supplies heat to its own fermenter, a drying plant and a residential area. As analysis of existing biogas plants show [4], plants like this are usually found in southern Germany.

To get a realistic feed-in behaviour of the reference plant, project partner Lechwerke AG (LEW) calculates spot-market oriented feed-in schedules based on market prices in 2018 and on estimated prices in 2035. This is necessary because flexibilised biogas plants often use the so called "Marktprämienmodell", based on §§ 20 and 23a EEG to get revenues beyond the regular sponsorship [1, p.62].

B. Definition of congestion

For the congestion generation strategy to apply, at first a congestion has to be defined. There are two types of congestions in electrical grid: violation of voltage limits and overloading of elements.

1) Voltage limits: In medium- and low-voltage grids voltage must remain within a band of ± 10 % of agreed supply voltage [5]. Generation units an storage systems are allowed to raise voltage in low voltage grids 3 % at maximum [6]. Since only non-regulable local network transformers are present in the grids under consideration, medium and low voltage levels are rigidly coupled here. Thus voltage in medium voltage grid must not exceed 107 % of agreed supply voltage.

In consultation with the project partner LVN, using 21.2 kV, i.e. 106 % of agreed supply voltage, is a practice-oriented value for the upper voltage limit. This value is used in the following as a definition for voltage problems.

2) Overloading: Overloads can permanently damage or destroy equipment and must therefore be prevented. Medium-voltage grids are usually planned for (n-1)-security which means that in the event of a failure of one piece of equipment, no other must be overloaded. Since generation plants can be remotely controlled, this rule does not apply to their connection to the grid. [7, p.195]

Thus, in the following, overloading of elements due to feed-in of generation units is considered as utilisation above 100 %.

C. congestion generation strategy

Congestions must be generated in the congestion-free network model for further investigations. This is basically possible by increasing the feed-in power from PV and wind power plants or by increasing the connected consumer power. To select a procedure, data of the Grid Development Plan Electricity 2030 [8] just as the feed-in time series of elements in the grid and the feed-in schedule of the reference biogas plant are analysed.

In scenario B 2035 of the Grid Development Plan, an installed PV capacity of 97.4 GW in 2035 is assumed for the whole of Germany. Compared with the system master data of the transmission system operators [9], this results in an increase of 55.0 GW in relation to 2017. 90.8 GW of installed capacity is assumed for wind onshore in the same scenario, which corresponds to an increase of 40.6 GW. The development will be distributed differently across the regions. While southern states will have a large installed capacity of photovoltaic systems, the north will be dominated by capacity from wind turbines. Scenario B 2035 reflects an ambitious expansion of electricity generation from renewable energies, moderate sector coupling and greater flexibility for consumers.

For electricity demand by consumers, the NEP identifies both demand-increasing and demand-reducing factors. Regional differences are particularly marked in the development of consumption. Depending on the scenario, net electricity demand in some districts will fall by 25 % or more, while in other areas it will rise by more than 25 %.

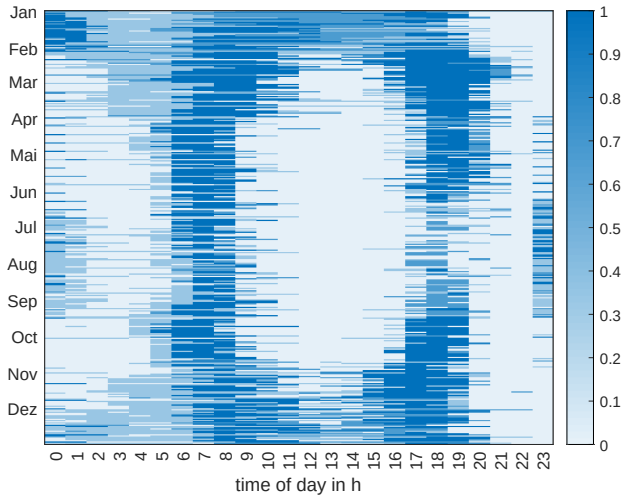


Fig. 1. Generation schedule of the reference biogas plant

On the basis of these forecasts, it is decided to generate congestions in the grid model by changing the feed-in power. The feed-in power will increase throughout Germany and only differ regionally in the technologies. Moreover, the increase in feed-in power in the simulations allows the use of the 100 % capacity utilization limit, as described in the previous section.

Increasing the feed-in power can be accomplished through either scaling photovoltaic or wind power plants. A comparsion between the biogas generation schedule, the photovoltaic power generation and the wind power generation shows, that increasing wind power is most suitable for generating congestions which can be handled by the biogas plant. In figure 1 one can see the generation schedule for the whole year 2018 as a heatmap, scaled on maximum power output from the biogas plant. It is immediately apparent that the biogas plant is usually, except in winter, not in operation from late morning to late afternoon. One can compare this to the usual time photovoltaic plants feed in, which is shown in figure 2. This figure shows the feed in power of photovoltaic plants in the grid model used for the examination. Comparing these two figures, it is apparent, that the biogas plant usually is in operation at times, when the photovoltaic infeed is low. As a result, PV-related congestions usually occur outside the operating hours of the flexibilised biogas plant.

This behaviour can be explained with the spot market prices. In [10] the spot market prices in 2018 are pointed over the photovoltaic feed-in power in germany. A raising amount of photovoltaic power in the grid results in a lower price. This is to be expected thus photovoltaic plants today don't sell their power at the spot market, but get a fixed sponsorship which is granted through EEG. At the spot market, photovoltaic power is sold with a price of 0 EUR. Because the feed-in schedule of the biogas plant is spot market optimized, i.e. it is attempt to earn the most, the plant is in operation when prices are high and therefore usually not, when photovoltaic infeed is high.

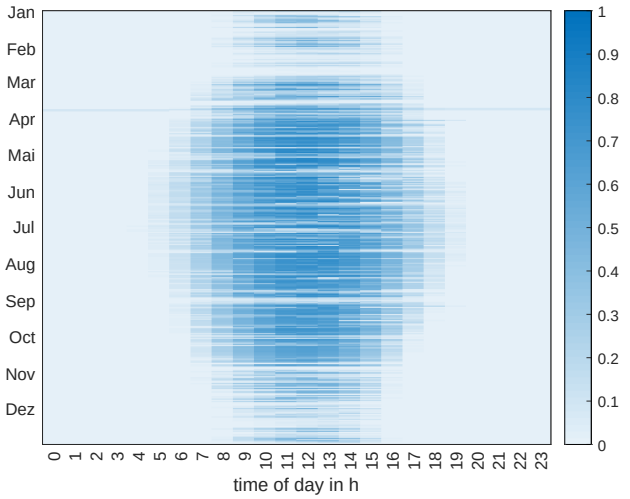


Fig. 2. Feed-in power of photovoltaic plants in the grid model, scaled to installed photovoltaic power

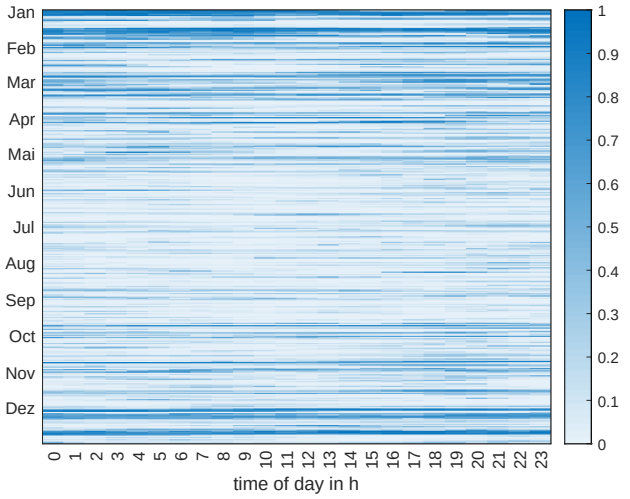


Fig. 3. Feed-in power of wind power plants in the grid model, scaled to installed wind power

Biogas plants are only able to handle overload situations in the grid caused by high infeed when they are in operation at this times. Then they can reduce power. Otherwise, the only possibility was to switch on and feed in additional power, which is contradictory to the aim of reducing current on overloaded elements. Due to the relations mentioned above, the choosen biogas plant with the calculated generation schedule cannot serve the net in situations, where photovoltaic driven congestions appear.

For congestion generation infeed of wind turbines can be scaled, too. Figure 3 shows the generation characteristics of the wind power plants in the grid model used for the examination.

One can see, that infeed power of wind power plants shows no characteristic daily behaviour like photovoltaic. At most

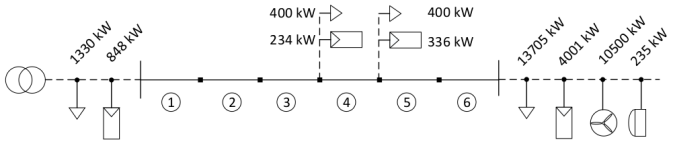


Fig. 4. Schematical drawing of critical overhead line section

figure 3 shows a seasonal difference in power generation between summer and winter. In winter the wind turbines are high utilised for longer periods of time then in summer. Because phases with high infeed-power of wind turbines can last many hours up to more than a day, there are times where wind infeed is high as well as the reference biogas plant is in operation. The biogas plant isn't able to shut down for such a long time, because gas production is a continous process and storage capacity will be exceeded after several hours. This makes it possible to generate congestions through scaling the installed wind power in the grid model which can be handled through the biogas plant.

D. Choosing a congestion situation for examination

After some simulations with the grid model, an overhead line section shows up as a critical point for overload when the feed-in power is increased. A schematical drawing of the section is shown in figure 4.

The medium-voltage string starts from the transformer substation with an almost 4 km long NA2XS2Y cable section in which consumers with a rated load of 1330 kW as well as PV systems with a rated output of 848 kW are connected. The affected overhead line section connects to the cable section. The overhead line section extends over a total length of 1.78 km with a cross-section of 95 mm^2 and is divided into six sections by five nodes. Within the section, two short spur lines supply two hamlets with 400 kW rated consumer power each and 234 kW and 336 kW of installed PV power. The remaining string after the affected overhead line section branches out several times. Connected to it are in total 13705 kW consumers, 4001 kW PV systems, 10500 kW wind turbines and 235 kW biogas plants.

The mentioned reference biogas plant is set at the very end of the medium-voltage string, which means it is geograpically a few kilometres away from the overhead line section. To analyse the suitability of biogas plants for congestion management, a simulation model is used, which monitors the load at section 1 of the overhead line. If the utilisation is to high, feed-in power from the biogas plant will be reduced. Another simulation model is used, with witch the given generation schedule is followed, no matter how high utilisation of the overhead line is.

Simulations with a scaled wind power output show that with a scaling factor of 1.2 an overload occurs on 23.09. at 10:30 am[1]. At this time the biogas plant is in operation. Therefore, this day will now be examined more closely.

[1]times are normalized to winter time

III. FIRST RESULTS

With a scaling factor of 1.2 for the feed-in power from wind energy plants and a scheduled operation of the biogas plant, the simulations show an overload of the complete overhead transmission line in the quarter hour from 10:30 am to 10:45 am. Sections 1 to 3 show a utilisation of 101.69 %, section 4 101.41 % and sections 5 and 6 100.82 %.

If the operation of the biogas plant is adapted to the congestion, no more congestion occurs. The biogas plant is therefore able to completely eliminate the congestion by adapting the operation mode. The model reduces the feed-in power from the original 750 kW to 521.4 kW between 10:30 am and 10:45 am. The utilization in section 1, whose utilization is used as control parameter, is thus reduced to 99.98 %.

In this specific case, the utilisation of the overhead line in section 1 has a sensitivity to the change in active power of $\frac{\delta c}{\delta P} = 7.48 \frac{\%}{\mathrm{MW}}$. The reaction of the biogas plant to the congestion also cause a change in filling levels of the storage tanks at the end of the day. The gas storage tank filling level shows 126 kWh more stored energy due to power reduction to scheduled operation. In contrast, there is no change in the heat storage tank filling level. This is due to the fact that in operation with spot market optimized schedule the heat storage tank is full from 9:30 a.m. to 12:00 p.m. and the thermal energy produced is discharged via the emergency cooler. Due to the congestion reaction, the amount of energy removed by the emergency cooler drops from 152.5 kWh to 103.06 kWh within the fifteen minutes in which the control is carried out. The use of a substitute heat supply as well as the gas flare is not necessary either when operating according to the schedule or after the congestion reaction within the day under consideration.

With a scaling factor for wind turbines of 1.225 and a scheduled operation of the biogas plant, the considered overhead line section is overloaded in three quarters of an hour. From 08:00 a.m. to 08:15 a.m., only section 4 is affected with 100.08 % and sections 5 and 6 with 100.21 % load. In the quarter hour from 09:00 a.m., the load in sections 1 to 3 is 101.35 %, in section 4 100.98 % and in sections 5 and 6 100.27 %. The third quarter of an hour affected begins at 10:30 a.m. At that time, the load factor is 104.06 % in sections 1 to 3, 103.78 % in section 4 and 103.19 % in sections 5 and 6.

If the operation of the biogas plant is adapted to the congestion situation, the congestion is completely cleared at 09:00 a.m. and 10:30 a.m. At 09:00 a.m. the simulation model reduces the output of the biogas plant from 750 kW to 540 kW, at 10:30 a.m. from 750 kW to 190 kW. The overload of sections 4 to 6 at 08:00 a.m. is not detected by the model, because the load in section 1 is used as the control variable.

The sensitivity of the load of the overhead line to a change in active power of the biogas plant is $7.47 \frac{\%}{\mathrm{MW}}$ at 09:00 a.m. and $7.49 \frac{\%}{\mathrm{MW}}$ at 10:30. Assuming that the sensitivity at 08:00 a.m. also has these values, a power reduction of 28 kW would be necessary to eliminate the overload at this time.

The gas storage is 454.79 kWh more filled at the end of the day after the power reductions to eliminate the congestion than during scheduled operation. The gas flare will not be used. At the end of the day, the heat storage tank filling level does not differ from the schedule operation compared to the congestion regulation. Only 181.9 kWh less heat energy is dissipated to the environment via the cooler. The use of a substitute heat supply is not necessary.

With a scaling factor of 1.25 for the output from wind turbines and pure scheduled operation of the biogas plant, overloads of the overhead line section occur in five quarters of an hour. From 08:00 a.m. to 08:15 a.m. the utilisation rate is between 102.25 % in section 1 and 102.60 % in section 6. From 09:00 a.m. to 09:15 a.m. the utilisation rate is slightly higher at 102.63 % in section 1 and 103.72 % in section 6. From 10:15 a.m. there is a utilisation rate between 100.33 % and 100.71 %, which rises significantly from 10:30 a.m. to values between 105.57 % and 106.44 % before falling back to values below 100 % at 10:45 a.m. From 15:15 to 15:30, sections 1 to 3 are overloaded with 101.19 %, section 4 with 100.54 % and sections 5 and 6 show no overloads.

If the feed-in capacity of the biogas plant is regulated according to the load in section 1, the picture is improved. At 08:00 o'clock the capacity is reduced from 750 kW to 420 kW. Thus, the congestion in sections 1 to 4 can be eliminated, in sections 5 and 6 an overload of 100.12 % remains due to the model. At 09:00 a.m. the congestion can be completely removed by reducing the output from 750 kW to 220 kW, and at 10:15 a.m. by reducing it to 630 kW. At 10:30 a.m., despite a complete shutdown of the biogas plant, a capacity utilisation of 100.83 % remains in sections 1 to 3 and 100.56 % in section 4. The sections 5 and 6 are free of congestions. At 15:15 there is no change in the situation, as the biogas plant is not in operation at this time.

The sensitivity of the capacity utilisation in the overhead line sections to a change in the active power feed-in of the biogas plant lies between $7.46 \frac{\%}{\mathrm{MW}}$ and $7.51 \frac{\%}{\mathrm{MW}}$ at the times of the congestions. A concrete analysis of the situation at 08:00 a.m. shows that an additional reduction of 16 kW of active power would be necessary to sufficiently relieve sections 5 and 6 as well. This is possible without any problems, but was not carried out by the model due to the specification of the load in section 1 as control variable. At 10:30 a.m., the power would have to be reduced by a further 111 kW to eliminate congestions. This can no longer be achieved by the biogas plant alone. Due to the power reductions for congestion control, there is 1094.27 kWh more energy in the gas storage tank at the end of the day than in pure scheduled operation. However, in both cases the gas flare is not used. The heat storage tank filling level at the end of the day is identical comparing congestion regulation operation to scheduled operation. The less produced heat energy results only in a lower use of the emergency cooler. In the case of bottleneck control, this has to dissipate 1247.9 kWh of heat,

a reduction of 423.6 kWh compared to scheduled operation. The use of a replacement heat supply is not necessary.

IV. DISCUSSION

Because of Corona-crisis, creation of spot-market optimized generation schedules through an external partner of the research project has become delayed, so that not enough simulations have been done yet to enable reliable statements.

Nevertheless, results in III show that in this concrete situation (given grid, given supply and demand situation, given network connection point of the biogas plant), the biogas plant can lower utilisation of the overhead line with around $7.5\frac{\%}{\mathrm{MW}}$. The effects on gas storage levels are small enough to delay energy losses into the next day. This allows the changed levels to be taken into account in day-ahead planning. The reduction in heat energy losses are positive for environmental reasons.

V. FUTURE WORK

In further simulations, it has to be figured out, how generally applicable these results are. Therefore different types of spot-market oriented generation schedules have to be simulated and other network connection points for the reference biogas plant have to be choosen. It has to be examined, how often biogas plants can handle grid congestions in specific net constellations of wind, PV and biogas plants and how much biogas plants can lower the forced reduce of power from wind and PV plants. The results must be validated at different grids. All this comes especially for the sensitivity values of the biogas plant. These values will be used in an optimisation model which is intended to get the spot-market optimised generation schedule as an input and calculates the best generation schedule taking into account several ancillary services for the grid. One of this ancillary services will be congestion management.

REFERENCES

[1] G. Häring, K. Bär, M. Sonnleitner, W. Zörner, and T. Braun, "Steuerbare Stromerzeugung mit Biogasanlagen," Institut für neue Energie-Systeme, Technische Hochschule Ingolstadt, Schlussbericht, Apr. 2015.

[2] U. Holzhammer, B. Krautkremer, M. Jentsch, and J. Kasten, "Beitrag von Biogas zu einer verlässlichen erneuerbaren Stromversorgung," Fraunhofer-Institut Windenergie und Energiesystemtechnik, Tech. Rep., 2016.

[3] O. Brückl and J. Schächinger, "Forschungsprojekt OPTIBIOSY - Bericht: Analyse der Netzmodelle," 2020, unpublished.

[4] M. Wildfeuer, M. Becker, O. Schmidt, and J. Schächinger, "Forschungsprojekt OPTIBIOSY - 2ter Zwischenbericht," 2020, unpublished.

[5] *Merkmale der Spannung in öffentlichen Elektrizitätsversorgungsnetzen*, DIN Std. 50 160, Rev. 2010, 2010.

[6] *Erzeugungsanlagen am Niederspannungsnetz - Technische Mindestanforderungen für Anschluss und Parallelbetrieb von Erzeugungsanlagen am Niederspannungsnetz*, VDE Std. AR-N 4105, Rev. 2011, 2011.

[7] A. Sillaber, *Leitfaden zur Verteilnetzplanung und Systemgestaltung*. Wiesbaden, Germany: Springer Fachmedien, 2016.

[8] 50Hertz Transmission GmbH, Amprion GmbH, Tennet TSO GmbH, and TransnetBW GmbH, "Netzentwicklungsplan Strom 2030: Version 2019, 2.Entwurf der Übertragungsnetzbetreiber," Apr. 2019.

[9] 50Hertz Transmission GmbH, Amprion GmbH, Tennet TSO GmbH and TransnetBW GmbH. (2019, Sep.) EEG-Anlagenstammdaten. [Online]. Available: https://www.netztransparenz.de/EEG/Anlagenstammdaten

[10] Fraunhofer ISE. (2020, Jun.) Energy charts. [Online]. Available: \url{https://energy-charts.de/price_scatter_de.htm?source=priceVSSolar&year=2018}

Realistic case study for the comparison of two production processes

Matthias Götz
OTH Regensburg
Prüfeninger Straße 58
93049 Regensburg, Germany
Email: matthias2.goetz@st.oth-regensburg.de

***Abstract*—Each manufacturing company has to decide which production process they use. The selection is complicated, because these processes have different advantages and disadvantages as well as an impact on other parts of the company. Therefore, in this paper two of the most common processes are compared. These processes are the job shop and the flow shop. In the case study it is assumed that the exemplary company has a defined amount of stations. The result of the case study is that the flow shop has better results if some conditions are fulfilled. These conditions are a short transport time for the flow shop as well as grouped work steps which have a processing time close to the cycle time.**

I. Introduction

This paper was written in the context of the master thesis at the Ostbayerische Technische Hochschule (OTH) in the laboratory for business informatics, SAP and production logistics under the direction of Professor Dr.-Ing. Frank Herrmann. The primary topic of the project is scheduling. In the case study the production processes job shop and flow shop are compared.

A. Goal of the project

The aim of this thesis is to obtain a realistic comparison of the production processes job shop and flow job. The production methods are used to determine which work steps of different orders are processed at which time at a station. The target criterion for the comparison is the average delay of all orders.

B. Requirements for the case study

To achieve a realistic result, the transport times between stations must be taken into account. In addition, the arrangement of the stations, the processing times of the work steps, and the transport times between stations must be chosen carefully. To achieve a better result, the specified orders can be repeated over several iterations, which helps to show different effects of the two production processes. In doing so, orders that have not yet been completed at the end of an iteration have a preferred priority over new orders in the following period. These basic conditions must be fulfilled for a realistic and fair comparison.

C. Tool support

The calculation of the algorithms for the workshop and assembly line production is done by a tool programmed for this purpose, which can also display several iterations. It offers the possibility to change the algorithms and add new algorithms.

D. Structure of the paper

The outline of this paper is as follows. In order to understand the comparison, in section II the algorithms used for the two production processes are presented and explained. Then in section III the case study used to compare the production processes is described. Following in section IV with the application and results of the case study. The final chapter is about the conclusion and analysis of the comparison.

II. Algorithms of the production processes

A. Job shop

For job shop the priority rule of shortest operation time is used but in a modified version. The reason for this priority rule is, that the rule leads to a relatively small medium delay for very long delays for a small part of the orders, which suits to the defined goal of the project. [1]

This modified version of the priority rule pays attention to the transport times, too. It is important because otherwise it would have considerably worse results. It allocates the work steps the following way. Each work step is checked at a specific time, starting with the first time unit to the end of the iteration, when at least one station is available. If a work step is available, that includes all necessary previous work steps are finished, it will be compared with the other available work steps with the best rating for all available stations. The rating is calculated in the following way. The rating of an available work step at an available station equals the process time of the work step plus the remaining transport time for the work step at an available station. This has to be repeated for all available stations and all available work steps. The work step with the smallest rating will be selected and the steps for a specific time are repeated until there are no available stations. The exception is when orders have delays and still have to be finished in the next iteration. Then these work steps have priority over the other work steps of the iteration, but will be selected with the smallest rating based on the description above. [2]

B. Flow shop

The first step is the belt adjustment. First of all the description of the product variants in the form of process graphs (precedence graphs) with the respective processing time per work step p_{vj} (v=work step, j=product variant) as well as the

number of products to producing units per product variant (dj) and the cycle time (T) for all stations is needed. Using the precedence graphs and the number of units to be produced per variant an aggregated graph can be created. This provides the basis for the belt adjustment and the station occupancy. To determine the cumulative processing time the equation 1 is applied.

$$p = \sum_{j=1}^{n} p_{vj} \cdot d_j \tag{1}$$

The assignment of the work steps to the stations can be done according to the remaining processing time or the priority rules. In this case, the Helgeson and Birnie algorithm allocates the work steps over the remaining processing time. After the assignment of the operations to the respective stations, the scheduling function creates a sequence for the for orders to be produced are determined. The sequence planning for the individual variants is carried out using the Nearest-Neighbor algorithm. Starting from a start node an adjacent edge with the minimum weight to a next node visited. A sequence is to be determined in which the sum of the directly successive orders a and b becomes minimal. All orders are used as locations and the load measure as the distance of the orders or places a and b. The aim of this procedure is a so-called round trip, in which each place is visited only once. As a result, the search is for a shortest Hamiltonian circle within the complete graph G. This means that a closed edge sequence of smallest length is found, that each node of the graph G only once. Thus an optimal order sequence (according to the selected order size) can be determined. [3]

III. Case study for comparison

A. Exemplary company Eder & Sohn GmbH

The company of this case study was founded in 1970 by the siblings Franz and Xaver Eder was founded in Regensburg. At that time it was a small carpentry workshop with five employees. Very soon after the foundation the specialization in desks, as the demand for them was constantly increasing. As a result, in a short time time a larger quantity of tables can be produced, which leads to an increased setting new employees. Thus the company continued to grow over the years. In With regard to the continuation of the company, one of the sons joined the carpentry one. In 2000, the company was converted into a limited liability company, the ”Eder & Sohn GmbH”. From The small family business became a medium-sized company with 200 employees. Thanks to the many years of experience of the Eder family in wood processing decided that the product range, which at that time consisted of a desk variant existed, is extended by a second variant. The desks produced are delivered to furniture stores in the region. Eder & Sohn GmbH is exclusively engaged in the production of desks. The raw materials, auxiliary materials and the packaging material are supplied by external related. Value is placed on sustainable produced and ecologically degradable materials. Environmentally conscious and sustainable production as well as continuous internal and external training flow into the daily work routine, optimise production and provide the is the mission statement of Eder & Sohn GmbH. The company consists of the management, Accounting, quality management, purchasing, production and sales. [4]

B. Requirements for production at Eder & Sohn GmbH

The Eder & Sohn GmbH has a 2000 square meter production area and additionally a warehouse. This makes it possible to operate the production machines both in a row, as well as in a circle. This enables the company to handle large orders comfortably and quickly. In addition, pre-produced products can be be stored in the hall provided for this purpose. The location is in an industrial area and is therefore very conveniently located and can be easily reached by suppliers and customers will be. The fictitious company described is certified according to DIN EN ISO 9001. It is therefore obliged to maintain all production facilities at fixed intervals and to check. The quality standard can thus remain at a constant level or even be improved. The use of a quality management system has led to The consequence is that the quality standard can be ensured in the long term. The company guarantees all employees regular instructions and training in the production techniques currently used. This serves to protect the employees and should serve to continuously increase productivity. Most modern Production conditions ensure a unique and outstanding design of the tables. Great importance is attached to a sustainable production of the tables. The The most important raw material in the production of the tables is wood. This building material has a neutral CO2 balance. The wood products used are certified according to the FSC (Forest Stewardship Council) and provided with a seal of approval. This means that the wood comes from responsible forestry. The company also places great value on suppliers from the region. So long transport routes and delivery times can be avoided and costs saved. Furthermore just-in-time production can be guaranteed, since the goods are delivered when they are needed. Short transport routes and times also have a positive effect on the climate and the environment. [4]

C. Arrangement of stations

1) Job shop: An arrangement of the stations in a circle is preferred in this case study in order to reduce the bearing deliveries to the machines and also the transport routes between the stations to be kept as low as possible. The following figure 1 represents the structure of the workshop production in this case study. Five machines are available to produce the six orders (S1 - S5). The arrangement is in a circle and the respective neighbouring stations require a transport time of 5 TU (time units) and opposite stations 8 TU. The warehouse is located between S1 and S2 and has a delivery time of 5 TU to these stations. The move time to S3 and S5 is 13 TU and to S4 12 TU.

2) Flow shop: In comparison to the job shop the flow shop has a less complex structure. The arrangement of the stations is in a line as seen in figure 2. Every order has to be transported

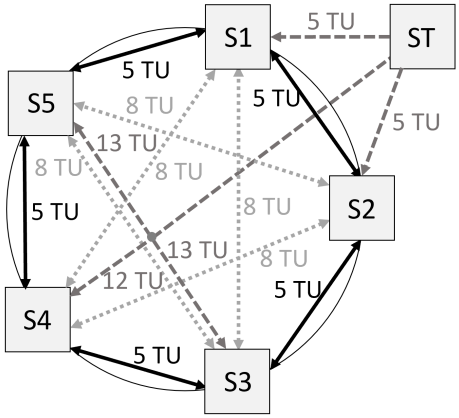


Fig. 1. Arrangement of stations for job shop

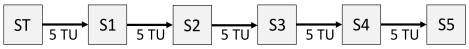


Fig. 2. Arrangement of stations for flow shop

from the storage to station 1, station 2, station 3, station 4 and to station 5. The transport time is 5 TU for each transportation.

D. Orders

For this case study six orders are used. These orders are divided in two variants, where the first three orders are variant one and the other three orders are variant two. These variants have different work steps. The following table shows all the work steps and their processing time for the variants. The processing time can differ at certain work steps, because some of them can be done faster or slower due to the variants.

TABLE I
PROCESSING TIME FOR WORK STEPS OF THE VARIANTS

ID	Description	Variant 1	Variant 2
1	Material cutting	40	20
2	Turning table legs	50	-
3	Gluing tabletop	20	30
4	Varnish table legs	50	50
5	Installing cable duct	40	-
6	Intarsia	40	60
7	Varnish tabletop	75	75
8	Preassembly of fittings	25	25
9	Quality Control	30	50
10	Packaging	35	35

The details of the variants like the sequence of work steps are described in the following subsection. The variants are shown in figure 3 and 4, which are described as general precedence graphs. [5]

1) Variant 1: Variant 1 consists of 10 work steps. First the required material is cut to size. Subsequently, work is carried out on the table top and the table legs in parallel. On the one hand, the table legs are turned, then painted and finally pre-assembled the fittings. On the other hand the table top is glued, a cable duct the inlay and the entire panel was painted, were all of these parts installed and after the successful completion of the following steps, the quality control of the individual parts is carried out and finally the packaging for shipping. In the figure 3 the sequence of the work steps is shown. The description for the IDs can be found at table I.

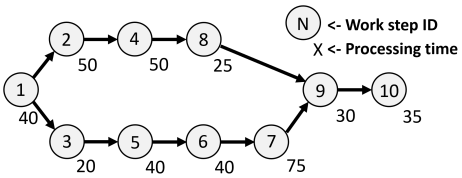


Fig. 3. Precedence graph of variant 1

2) Variant 2: This product variant consists of eight work steps. As with variant 1, first the material is cut to size and then the table legs are painted in parallel and the table top is glued. Between gluing and varnishing the table top, another inlay is inserted. The pre-assembly of the fittings can only be started after successful completion of steps 4 and 7. The quality control represents the penultimate operation before the product is packaged and is followed by the varnishing of the table top and the pre-assembly of the fittings. In the figure 4 the sequence of the work steps is shown. The description for the IDs can be found at table I.

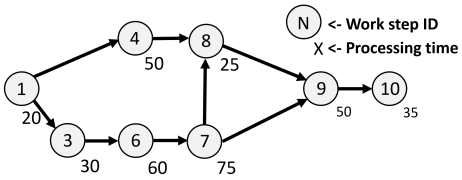


Fig. 4. Precedence graph of variant 2

E. Iterations

As stated in the introduction in the section requirements for the case study, it is helpful to have multiple iterations. In this case study there are 10 iterations. The reason is, that delays can have an huge impact on the following iterations. Since the customer requirements only arrive over time, planning always needs to be updated. Therefore the planning distance is set to 450 TU. The total time for all iterations is 4500 TU. For each iteration all available work steps are assigned to a station with a specific time for the production. [6]

IV. APPLICATION AND RESULTS

In this case study for ten iterations there are in total 60 orders and 540 work steps. This is a large sum. Therefore the results are shown either as a part or as a summary, because it offers a better overview. The table II shows the average and total delay for all six orders of an iteration. In the table js

TABLE II
TOTAL AND AVERAGE DELAYS OF THE TWO PRODUCTION PROCESSES

Iteration	js total	js average	fs total	fs average
1	147	24,5	0	0
2	209	34,8	0	0
3	224	37,3	0	0
4	249	41,5	0	0
5	270	45	0	0
6	280	46,7	0	0
7	295	49,2	0	0
8	298	50	0	0
9	306	51	0	0
10	321	53,5	0	0
Total	2599	43,3	0	0

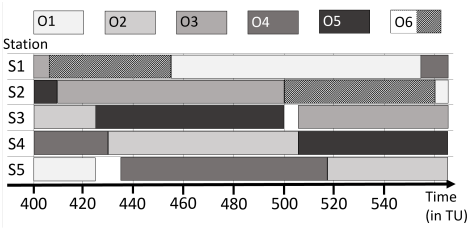


Fig. 5. Allocation of work steps for flow shop

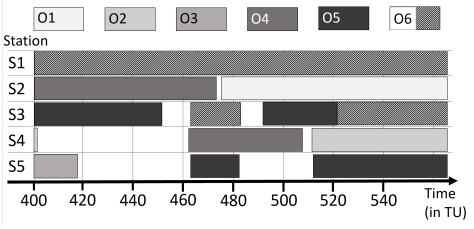


Fig. 6. Allocation of work steps for job shop

stands for job shop and fs for flow shop. While there is no delay with the flow shop, there is a growing delay for each iteration with the job shop. The total delay of 2599 TU is very high in comparison to the total time for the case study.

While there is a average delay of 43.3 TU, in most cases there is an order with a huge delay while up to three orders (at the first iteration) are in time. The highest delay for an order is 170 TU, which can be seen in the table III. For a better understanding of the different methods, there are the figures 5 and 6. The observed range is from time unit 400 to time unit 550. This is an interesting range, because there the change between iteration one and two takes place. In figure 6 there is the simple procedure of the flow shop. Each order starts at station 1 and is transported to the next station when the work step package for a station is finished. In figure 5 there is no noticeable structure for the job shop. While some of the orders are already finished, there are two orders that impact the next iteration, because it tends to keep orders at a single station. Also, there is a huge idle time at station four and five.

TABLE III
DELAY OF ORDERS FOR EACH ITERATION FOR JOB SHOP

Iteration	Order	Delay	Iteration	Order	Delay
1	1	0	6	1	0
1	2	0	6	2	30
1	3	0	6	3	0
1	4	8	6	4	142
1	5	27	6	5	60
1	6	112	6	6	48
2	1	0	7	1	15
2	2	29	7	2	33
2	3	0	7	3	0
2	4	7	7	4	37
2	5	43	7	5	65
2	6	130	7	6	145
3	1	0	8	1	20
3	2	22	8	2	25
3	3	20	8	3	0
3	4	25	8	4	40
3	5	157	8	5	65
3	6	0	8	6	148
4	1	0	9	1	20
4	2	20	9	2	35
4	3	0	9	3	0
4	4	135	9	4	43
4	5	52	9	5	148
4	6	42	9	6	60
5	1	7	10	1	15
5	2	35	10	2	33
5	3	28	10	3	50
5	4	30	10	4	43
5	5	165	10	5	170
5	6	5	10	6	10

V. CONCLUSION

There are some key factors that can have an impact on the two production processes. Short transport times are beneficial for the flow shop in comparison to the job shop, because there will be exactly the amount of stations (under consideration of the storage) as the total amount of transports for every order. Therefore a high transport time would result in much idle time for the stations while the job shop tends to accept short transport times when they are much smaller then the processing time of work steps which results in more idle time. Another key factor is the selection and amount of work steps. If there are only few work steps the work packages of the flow shop can have very different processing times while it does not have an impact on the job shop.

ACKNOWLEDGMENT

The author would like to thank Prof. Herrmann for the support on the project as well as the colleagues who designed the products and the colleague who supported in programming the tool for the production processes.

REFERENCES

[1] F. Herrmann, *Operative Planung in IT-Systemen für die Produktionsplanung und -steuerung*, 1st ed. Wiesbaden: Vieweg+Teubner, 2011.

[2] T. Nebl, *Produktionswirtschaft*, 7th ed. Munich, Vienna: Oldenbourg, 2011.

[3] K. Neumann, *Produktions- und Operations- Management*, 1st ed. Berlin, Heidelberg: Springer, 1996.

[4] T. Niebler,J. Urmann *Fallstudie im Rahmen des Projektstudiums*, Regensburg: Internal Report, 2020.

[5] H.O. Günther,H. Tempelmeier, *Produktion und Logistik*,9th ed Berlin, Heidelberg : Springer, 2012.

[6] F. Herrmann, M. Manitz *Materialbedarfsplanung und Ressourcenbelegungsplan*, 1st ed. Wiesbaden: Springer, 2017.

Machine learning methods for creating personality profiles from data in social networks

Andreas Arnold
Laboratory for Information Security
OTH Regensburg
Email: andreas.arnold@st.oth-regensburg.de

***Abstract*—Although spear phishing is more effective, it is used much less than classic phishing. The reason for this is the effort to collect the required prior knowledge. Methods of machine learning, such as neural networks, support vector machines and decision trees can reduce this effort using social network data, such as Facebook. This project deals with creating personality profiles, phishing and methods of machine learning and describes a model for automation in spear phishing. This model generates, among other things, a directory of personality traits like name, gender and CV details, and varying e-mail modules such as introduction, main part and conclusion. These modules then are randomly combined to create email drafts and passed on to artificial intelligence.**

***Index Terms*—Social engineering, Phishing, Personality profiles, Neural networks, Support vector machine, Decision trees**

I. Introduction

The theft of sensitive user data by means of misleading messages is known as phishing. [cf. 1] Back in 1994, fraudsters posed as employees of America Online (AOL) and asked AOL users to verify their accounts or confirm their billing information. [cf. 2] Little has changed in this approach up to date.

According to the International Coalition against Cyber-Crime, the number of phishing attacks discovered each year has increased rapidly since 2012 (320.081 attacks), peaking at 1.413.978 attacks in 2015. However, a decline has been seen in recent years. In 2018 the figure was at 1.040.654 attacks. [cf. 3]

In contrast to classic phishing, in spear phishing the attacker attempts to specify and personalize his e-mails. The more successful this is, the higher the chances of success. In 2016, 28% of phishing attacks worldwide were spear phishing attacks [cf. 4, S. 34].

Spear phishing attacks were the most common targeted attack method worldwide in 2017 with 71.4%. Only 28.6% were carried out by other means, such as trojans. [cf. 5, S. 76] In Japan, a leading industrial nation, the number of spear-phishing attacks has increased from 3828 in 2015 to 6740 in 2018, according to the National Police Agency. [cf. 6, S. 3]

Although spear phishing is one of the most effective targeted attack methods, it is used much less than classic phishing. The reason for this is the additional effort required for researching prior knowledge and collecting personality traits. Methods of machine learning, such as neural networks, support vector machines and decision trees can be a solution to this problem using data from social networks such as Facebook. These can automate the complex process of spear phishing and thus reduce the effort for the attacker.

In this paper, the basics of phishing, methods of machine learning (Supervised learning algorithms) and personality profiles are discussed. Subsequently, a model for automation in spear phishing using machine learning is explained.

II. Methods of machine learning

In the next section, the basics of machine learning will be discussed. This includes the categorization of learning algorithms and the presentation of three common methods. These form the core of the personality profiles used to create the Personality Factor Profile.

A. Supervised learning algorithms

Supervised learning requires a sufficiently large number of function arguments X, which already have a function value Y. These data records are labeled or marked. The model is then calculated from this basis. An example is a data set which contains information about which animal is depicted from a series of dog and cat pictures. This can be used to train a learning algorithm to distinguish between dog and cat images. If the learner has to decide which of the two animals is depicted in an unlabeled image, he or she can draw on the experience gained from learning the labeled data sets and make a decision. In this procedure, two problem categories are to be differentiated. The classification problem and the regression problem. [cf. 7, S. 12]

1) Classification: In classification, the target set Y is discrete and therefore each element must be considered individually. An example for this would be the example mentioned above with its target set Y = {cat, dog}. [cf. 7, S. 13 to 14]

2) Regression: In regression, on the other hand, the target quantity Y is built up continuously. If you move 0.5 units from the 3 to the left or right at the target quantity Y = [0.10] (interval of 0 - 10), the limit of the numbers 2 or 4 is located there. The target quantity of a regression problem therefore covers spaces and not elements to be considered individually as in the classification problem. A classic example of a regression problem would be the forecast of a stock price. [cf. 7, S. 14 to 15]

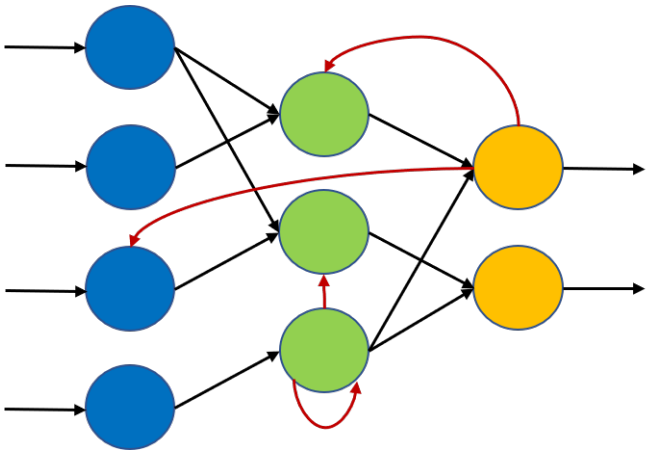


Fig. 1: Schematic neural network with feedback [cf. 8, S. 5 und 29]

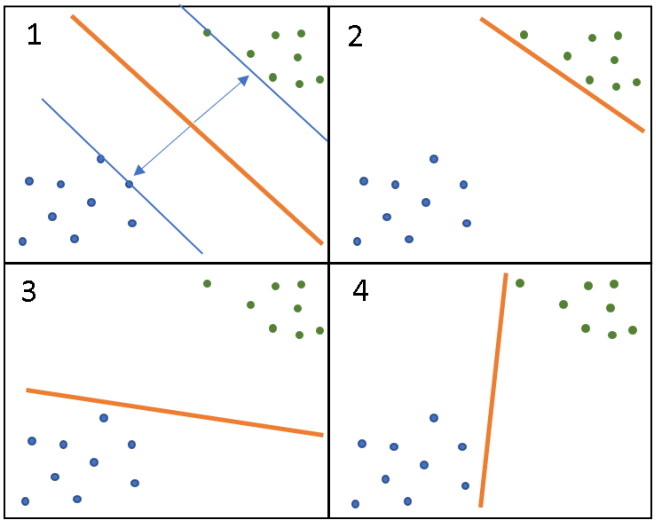


Fig. 2: Possible positioning of the hyperplane [vgl. 10, S. 129]

3) Eager learning and lazy learning: In the context of supervised learning, the terms Lazy and Eager Learning play an important role. Both describe procedures for modelling. With Eager Learning, more time is invested in training. This is because a model trained using Eager Learning, which is intended to distinguish between dog and cat images, for example, will deliver fast results due to its global concept and its complex training phase. In contrast to this is the learning process Lazy Learning. This method invests less time in training. But the time and computing costs for each query are significantly increased. The reason for this is the creation of a local model, which is designed for each query. Depending on the problem, the use of the different learning processes can achieve better results. [cf. 7, S. 15]

B. Unsupervised learning

In contrast to supervised learning, there is no defined target quantity Y for unsupervised learning. Unsupervised learning attempts to find structures and dependencies from an appropriately unmarked set of data X. If we give such a model the images of dogs and cats as function arguments X, it is possible that this model will give the same results as a supervised model. The algorithm has independently detected dependencies and structures between the images and has divided them into groups 1 and 2. The developer has to decide which animal species contains group 1 or 2. However, if the learning algorithm has, for example, structured according to the size of the animal depicted in the photo, it is possible that both group 1 and 2 contain cats and dogs. For example, a Chiwawa could fall into the group of cats due to its small size. [cf. 7, S. 16 to 18]

C. Neural networks

Artificial neural networks are modeled on the human brain. These networks consist of neurons, the so-called units (fig 1). There are three categories of units: Input units (marked in blue), hidden units (marked in green, optionally) and output units (orange). Input units receive input data and forward it to the subsequent layers. There is usually one input unit for each function argument. Hidden units are located between input and output units and are therefore called hidden. There can be zero or more hidden layers in a neural network. A layer is a series of units. Output units, of which there must be at least one, display the result. Units are connected with edges, where each edge has a weight. The weight represents the knowledge of the neural network and can be positive, negative and zero. If the weight is positive, the unit exerts an exciting influence on the connected unit, if it is negative, it exerts an inhibiting influence, and if the weight is zero, it has no influence on the next unit. Each unit receives an input from at least one predecessor. The higher the weight (w) and the higher the output of the predecessor (also called activity a), the greater the influence on the receiving unit. If one of the two values is zero, no influence is exerted. It applies to the input of unit i:

$$\text{input} = a_j w_{ij} \tag{1}$$

Where i is the receiving and j the sending unit. [cf. 8, S. 5 - 15]

1) Recurrent networks: Recurrent edges (fig 1, marked in red) direct the output of a unit back to a past, side or own unit and provide it with a memory. [cf. 8, S. 29]

2) Deep learning: A method that has hidden units is often called Deep Learning. [cf. 9, S. 15 - 16]

D. Support vector machines

The Support Vector Machine (SVM) as a method of machine learning divides a set of objects (vectors) into classes. This is done by a linear dividing line (hyperplanes). The basis for this is a set of labeled training data that is assigned to a class. The aim of SVM is to select the hyperlevel so that the area around the class boundaries is as free as possible. New vectors are assigned (predicted) to one of these classes depending on their location. [cf. 10, S.123 - 127] As shown

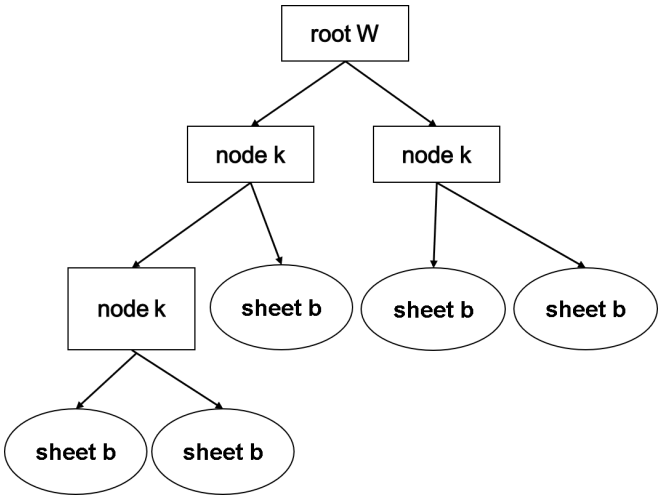


Fig. 3: Simple decision tree [cf. 7, S. 109]

in Figure 2, the hyperplane can be positioned differently. To increase the probability of a correct class assignment, the margin between the two classes should be as large as possible. Not all training vectors are considered. The margin depends only on the closest vectors (support vectors). The prerequisite for this is a linear separability, which is usually not available. In this case the kernel trick or the soft margin is used. [cf. 11, S. 196 - 201]

1) Soft margin: When using soft margin, the SVM tolerates a few points that can be misclassified and tries to balance the compromise between finding a hyperplane that maximizes distance and minimizing misclassification. [cf. 11, S. 204]

2) Kernel trick: The kernel trick uses the fact that non-linearly classifiable data in a higher dimensional space are linearly separable. [cf. 11, S. 200 - 201]

E. Decision trees

Decision trees as a method of machine learning are based on the data structure of a tree (fig 3). Each tree has a root W where the evaluation starts, a decision is made. From this root follow nodes k. Each node represents a function argument X (e.g. [gender, age, size]). If no decision is made at node k, or no further nodes follow, the logical unit is called leaf b. A state Y is classified at this node (for example, {dog, cat}) or a regression value is output (for example, [0, 10]). [cf. 7, S. 109] Two well-known learning algorithms/base models for decision trees are the ID3 and the CART algorithm.

1) ID3: The ID3 (Iterative Dichotomiser 3) calculates its node structure by arranging the function arguments, i.e. nodes according to the highest information gain, until all leaves have been classified. It is not guaranteed that the tree determined is the best possible tree. The information gain (formula 2) of a function argument X_1 is calculated from the difference of the entropy E(D) of data set D and the sum of the entropies of the function argument and its decisions. [cf. 12, S. 89 and 90]

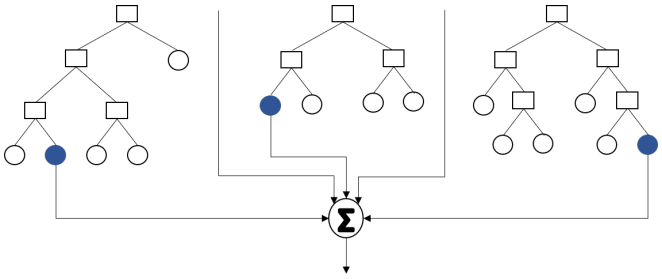


Fig. 4: Random Forest Bagging [vgl. 17]

$$\text{gain}(X) = E(D) - \sum_{i=0}^{n} \frac{X_\text{i}}{X} * E(X_\text{i}) \tag{2}$$

Entropy is a measure of the impurity of data. In machine learning, pure data is preferred, i.e. function arguments with decisions that are clearly distinguishable from each other. [cf. 13]

2) CART: The CART (Classification and Regression Trees) algorithm can be used for both classification and regression. For each decision there are exactly two subsequent nodes. The goal of the algorithm is to find an optimal binary separation for each decision. Similar to the ID3 algorithm, the function argument with the highest information content is chosen. [cf. 14, S. 73]

3) Ensemble methods: Ensemble methods are used in decision trees to achieve better results. They use a combination of several learning algorithms/base models. There are two types of ensemble methods: Bagging and Boosting. In bagging, several predictions from basic models are weighted equally and the average is calculated. A well-known example of this would be random forest. Boosting, on the other hand, combines many weak to strong predictions and thus gives more weight to misclassified data from previous rounds. Thus, each new tree corrects the one previously created. This procedure is used for gradient boosting (Boosted Trees). The basic model is the CART algorithm. [cf. 15, S. 551, 554, 580]

4) Random forest: Random Forest is used for regressions and classification problems. A random number of features k (where k is smaller X) is selected from the set of function arguments X and generated for these decision trees. Each of these decision trees now classifies the function arguments differently. When Random Forest Bagging is used, the average of all predictions is calculated and a prediction is made (fig 4). [cf. 16, S. 10]

F. The choice of a classifier

The choice of a classifier depends on the nature of the problem to be solved. In general, it is more important to find better functional arguments and improve the training data set than to focus on the choice of algorithm. According to the ”No free lunch theorem” for machine learning there is no algorithm that works best in all cases. To find the most suitable model, it is best to try them all. [cf. 18, S. 1343]

III. Personal profiles

According to Spektrum, a personality profile is the

> ”combination of personality traits that are meaningful or desirable for certain situations.” [19]

The creation of this overall picture is called profiling and takes place through analysis and evaluation of personal data. The most important application area for the work is psychology. [cf. 20]

A. Psychology

Some personality profiles are used in psychology. One of the best known is the five-factor model. It is named after its five main dimensions, which classify the personality of a person. These include open-mindedness, conscientiousness, sociability, tolerance and neuroticism (emotional instability). These are usually recorded by means of questionnaires. [cf. 21] A research study by the University of Maryland has shown that it is possible to predict the personality of a user through their publicly available Facebook profile information. In the thesis, this information was evaluated using machine learning methods and compared with the data collected by questionnaire. Thus, a model was developed which can predict the personality for each of the five personality factors with an accuracy of 11% of the actual values. [cf. 22]

IV. Application of machine learning and Personality profiles in context Spear phishing

In the following, the planned application of machine learning in the field of spear phishing by means of personality profiles is discussed. First the terms Phishing and Spear-Phishing will be explained.

A. Phishing

Phishing is a form of Internet crime. According to the Duden dictionary, it refers to the

> ”obtaining of personal data of other persons (such as password, credit card number or similar) with fake e-mails or websites”. [1]

Phishing is not just a technical process. It tries to manipulate victims through social engineering in such a way that they independently disclose the data or access to it. [cf. 23, S. 23] Etymologically, the term phishing results from the combination of the words phreaks, an early term for hackers, and fishing. [cf. 24, S. 560] As with fishing, a bait is put out and waited for a ’victim’ to bite.

B. Spear-Phishing

In contrast to classic phishing, the perpetrator of spear phishing tries to specify and personalize his e-mails. The more successful this is, the higher the chances of success. The attacker can, for example, use the company’s corporate design, forge the sender’s address, specify the recipient group or adapt the content to a given problem. In addition, he can adapt the appearance of the phishing website the victim was linked to in the e-mail as closely as possible to the original and only slightly change the URL of the original address. [cf. 25]

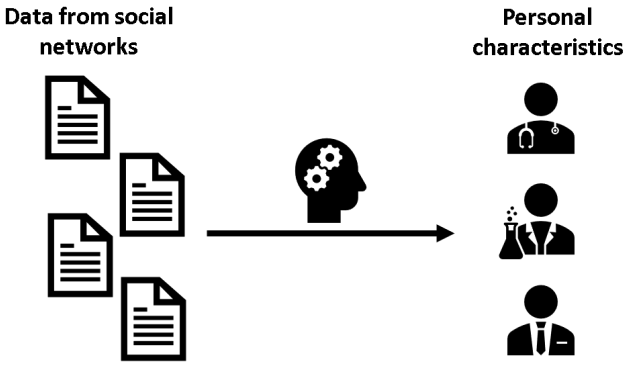


Fig. 5: Machine Learning and Spear-Phishing

C. Machine learning and spear phishing

Although spear phishing is more effective, it is used much less. The reason for this is the additional cost of acquiring the prior knowledge required. Machine learning and data from social networks, such as Facebook, could be a solution to this problem (fig 5). Data such as name, gender, age, list of friends, hobbies, taste in music, movies, ’like information’ and CV data can be collected in a file with personality traits. In addition, an artificial intelligence (AI) can improve the results by inferring more complex statements about the user’s personality from simpler basic data (fig 6). From this personality profile, mail modules such as subject, greeting, introduction, main part, conclusion and greeting formula can be generated. Several variants are created for each module. These are then randomly combined to form a e-mail. The combination of modules (function arguments) are passed on to another AI. The AI then evaluates whether the combination of building blocks is credible or not. If this is the case, a credible personality profile is created in the context of Spear-Phishing. Afterwards, a theoretical attack can take place. Even before the mail modules are handed over to the AI, the respective combination is assigned (labelled) to a function value Y (credible, not credible) by a person in the learning process (marked blue). With the help of this marking the AI can train its model.

V. Conclusion

In this paper, the basics of personality profiles, phishing and methods of machine learning were discussed and a model for automation in spear phishing was explained. This model generates a file with personality traits from data such as name, gender, age and CV details. Using this data, a program generates e-mail modules such as introduction, main part and conclusion. Several variants are created for each module. These are combined randomly to one mail and the modules are handed over to the AI. The AI then evaluates whether the combination of modules is credible or not. If this is the case, a credible personality profile is created in the context of Spear-Phishing. Afterwards a theoretical attack can be carried out with this e-mail. With the help of this model, a proof-of-concept tool for refining spear phishing attacks is to be

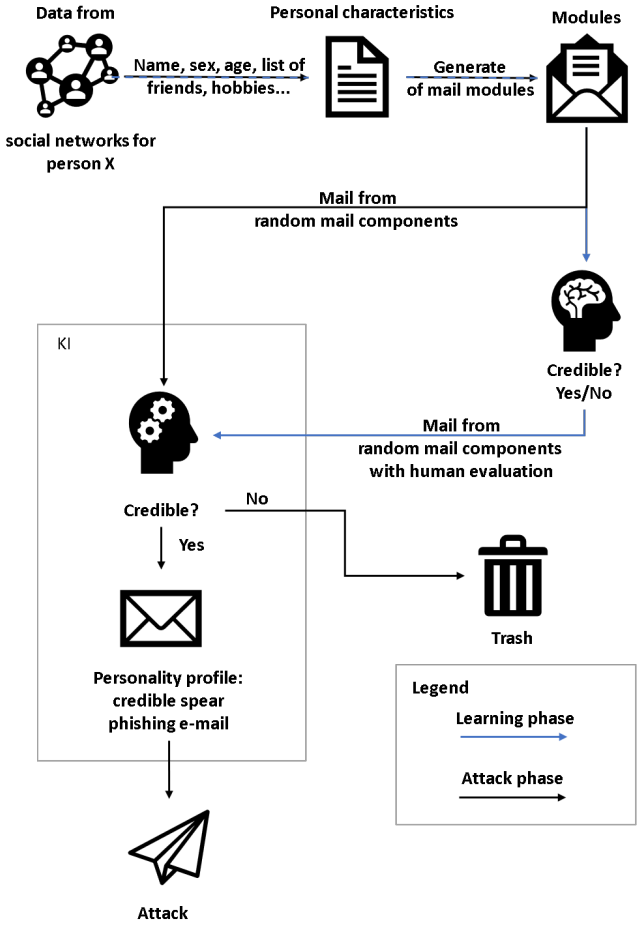


Fig. 6: Machine learning and personality traits

created in the coming project work. Subsequently, suitable countermeasures will be analysed and examined for their effectiveness. The goal is to achieve an automation of attack and defense in social engineering.

References

[1] Bibliographisches Institut GmbH, *Duden — phishing — rechtschreibung, bedeutung, definition, herkunft*. [Online]. Available: https://www.duden.de/rechtschreibung/Phishing.

[2] K. Rekouche, *Early phishing*. [Online]. Available: http://arxiv.org/pdf/1106.4692v1.

[3] APWG, *Phishing activity trends report*. [Online]. Available: https://www.antiphishing.org//resources/apwg-reports/.

[4] Verizon, *2017 data breach investigations report*, 2017. [Online]. Available: https://www.phishingbox.com/downloads / Verizon - Data - Breach - Investigations - Report-DBIR-2017.pdf.

[5] Symantec Corporation, *Istr internet security threat report volume 23*, 2018. [Online]. Available: https://www.phishingbox.com/assets/files/images/Symantec-Internet-Security-Threat-Report-2018.pdf.

[6] National Police Agency (Japan), *Number of spear phishing e-mail attacks in japan from 2013 to 2018*, 2019. [Online]. Available: http://www.npa.go.jp/publications/statistics/cybersecurity/data/H30_cyber_jousei.pdf.

[7] J. Frochte, *Maschinelles Lernen: Grundlagen und Algorithmen in Python*. München: Hanser, 2018, ISBN: 9783446457058.

[8] G. D. Rey and K. F. Wender, *Neuronale Netze: Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung*, 1. Aufl., ser. Programm Verlag Hans Huber Psychologie Lehrbuch. Bern: Huber, 2008, ISBN: 9783456845135. [Online]. Available: http://sub-hh.ciando.com/book/?bok_id=15205.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003. [Online]. Available: http://arxiv.org/pdf/1404.7828v4.

[10] C. J. Burges, *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998, ISSN: 13845810. DOI: 10.1023/A:1009715923555.

[11] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, ser. Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2002, ISBN: 9780262194754. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=78092.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, 1986, ISSN: 1573-0565. DOI: 10.1007/BF00116251. [Online]. Available: https://doi.org/10.1007/BF00116251.

[13] B. Aunkofer, *Entropie – und andere maße für unreinheit in daten*, 2019. [Online]. Available: https://data-science-blog.com/blog/2017/05/02/entropie-und-andere-masse-fur-unreinheit-in-daten/.

[14] L. Breiman, *Classification and regression trees*, Repr. Boca Raton: Chapman & Hall, 1998, ISBN: 0412048418.

[15] K. P. Murphy, *Machine Learning - A Probabilistic Perspective*, 1. Aufl. Cambridge: MIT Press, 2012, ISBN: 978-0-262-01802-9.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[17] J. D'Souza, *A trip to random forest*, 2018. [Online]. Available: https://medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a.

[18] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, 1996. DOI: 10.1162/neco.1996.8.7.1341. eprint:

https://doi.org/10.1162/neco.1996.8.7.1341. [Online]. Available: https://doi.org/10.1162/neco.1996.8.7.1341.

[19] Spektrum Akademischer Verlag, *Lexikon der psychologie: Persönlichkeitsprofil*. [Online]. Available: https://www.spektrum.de/lexikon/psychologie/persoenlichkeitsprofil/11400.

[20] Bibliographisches Institut GmbH, *Duden — profiling— rechtschreibung, bedeutung, definition, herkunft*. [Online]. Available: https://www.duden.de/rechtschreibung/Profiling.

[21] O John, L. Naumann, and C Soto, "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues," in. Jan. 2008, pp. 85–117.

[22] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, 253–262, ISBN: 9781450302685. DOI: 10.1145/1979742.1979614. [Online]. Available: https://doi.org/10.1145/1979742.1979614.

[23] Bundeskriminalamt, *Cybercrime: Bundeslagebild 2017*, 2018. [Online]. Available: https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/JahresberichteUndLagebilder/Cybercrime/cybercrimeBundeslagebild2017.html;.

[24] H. F. Tipton, *Information Security Management Handbook, Fourth Edition* -, Subsequent. Boca Raton, Fla: CRC Press, 2001, ISBN: 978-0-849-31127-7.

[25] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going spear phishing: Exploring embedded training and awareness," *IEEE Security & Privacy*, vol. 12, no. 1, pp. 28–38, 2013.

INDEX OF AUTHORS