COMPARING DIFFERENT WORKLOAD AND STRESS ASSESSMENT METHODS IN AIR TRAFFIC CONTROL SIMULATIONS

M. Finke, T. H. Stelkens-Kobsch
Institute of Flight Guidance, German Aerospace Center (DLR)
Lilienthalplatz 7, 38108 Braunschweig
Germany

Abstract

Measuring the current level of workload is one of the key indicators to assess the benefit of new supporting tools or new procedures for air traffic controllers. In addition, it is well-known that the current stress level depends on the current level of workload and is furthermore directly related to the current traffic capacity of air traffic control.

Different workload assessment techniques are already available, but they are either subjective like the Instantaneous Self-Assessment Technique (ISA) or cannot be used in real-time during simulations like the NASA task load index (NASA-TLX).

In the frame of the European 7th framework project GAMMA, a new stress measurement prototype was developed which is based on voice recognition. This prototype uses vocal stress signals to determine the current stress level of any speaker within air-ground voice communication in air traffic control. The advantage of such a system is the availability of objective stress measurements in real-time.

In order to verify this new stress measurement prototype, several high-workload ATC simulations were conducted at DLR Braunschweig in January 2017. During this simulation campaign, the prototype was used in parallel with ISA and with an assessment by a third person using a modified Cooper-Harper Scale (MCH). Active air traffic controllers attended the simulation campaign, which contained a total number of 20 air traffic control simulations using DLR's radar simulator ATMOS.

The trials showed that even common workload measurement methods (ISA and MCH) showed just a clear correlation between each other in less than 50% of the trials, which underlines the need for a reliable and objective stress measurement technique. The mentioned prototype shows the potential to become such a tool in the midterm future.

This paper gives basic information about the used assessment / measurement techniques, describes the setup and the conduction of these trials, illustrates obtained results and provides a discussion and an outlook about enhancements in stress measurement techniques.

1. INTRODUCTION

Air Traffic Flow and Capacity Management (ATFCM) is one of the cornerstones of air traffic management (ATM) in highly-frequented airspaces. ATFCM, or formerly Air Traffic Flow Management (ATFM), is defined by the International Civil Aviation Organization (ICAO) in the following way:

ATFM is *"a service established with the objective of contributing to a safe, orderly and expeditious flow of air traffic by ensuring that Air Traffic Control* (ATC) *capacity is utilized to the maximum extent possible, and that the traffic volume is compatible with the capacities declared by the appropriate Air Traffic Services* (ATS) *authority."*

This service primarily has the purpose to ensure that the declared ATC capacity is not exceeded by the traffic demand. This is done by analyzing the expected traffic flow in three phases (strategic, pre-tactical, tactical phase) based on filed flight plans. As a consequence, calculated take-off times (CTOTs) are allocated to flights which might otherwise cause an overload situation with an impact on a safe, orderly and expeditious flow of traffic [1].

Naturally, one key factor of the declared ATC capacity is the (maximum) workload of the air traffic controller responsible for an ATC sector or ATC area of

responsibility. In turn, the level of workload is driven by several aspects, for example [2]:

- Complexity of the area of responsibility.
- Complexity of the ATC procedures.
- Complexity of the air traffic and its demands.
- Traffic load.
- Status and capabilities of the available equipment and necessary workarounds.
- Secondary tasks.
- Stage of familiarization of the controller with the working environment and procedures.

Usually, for ATFCM the ATC capacity is estimated and declared mainly based on experience or theoretical calculations. In the future, workload measurement techniques could be used to support or even improve the estimation of available ATC capacity in real time and to detect upcoming overload situations. This approach would enable the consideration of aspects and factors which are not yet taken into account (e.g. emergency flights).

Further applications for workload measurement in the future are imaginable besides supporting ATFCM:

- ATC training or competence assessment: to determine the degree of routine a person or a trainee has maintained or developed.
- Dynamic workload distribution: e.g. for new ATM concepts like 'dynamic sectorization' [3] or 'multiple remote tower' [4].

Currently, workload measurement techniques are just widely used in research and development, especially in the human factors domain or for validating new tools and systems. Workload measurement is used to compare the level of workload with and without a new tool or procedure or between different system configurations. Examples for recent research activities using workload measurement are [4][5].

2. WORKLOAD AND STRESS

In air traffic control, a certain level of workload may lead to a higher level of stress experienced by an air traffic controller; the level of stress depends on the level of workload. For the following study, the level of workload is considered as 'cause' while the level of stress is considered as one feature of the 'human reaction' to it [6]. A stress reaction usually leads to measureable changes or other measureable signals of the human body, such as a raised heart rate, blood pressure, skin conductivity, eye activity etc. Former studies used physiological stress signals as indicators for workload [7].

3. WORKLOAD MEASUREMENT METHODS

Several workload measurement methods of different state of maturity are currently known and used in ATC simulation. For the study on hand, these techniques were grouped into the following categories:

1) Self-Assessment Techniques
2) Workload assessment questionnaires
3) Third-Person-Assessment Techniques
4) Primary / Secondary Task Performance
5) Technical measurements of any kind

Self-Assessment is done by asking the person under consideration how he or she would describe his or her current level of workload in general or in a regular time interval. This can be done by a person or by technical means.

Workload assessment questionnaires use a set of standardized questions to assess the workload experienced during the last task. These questions are answered after this task is completed and can be paper-based or supported by electronic means.

Third-person assessment is done by a person who is not involved in working on the scenario or in any other tasks of a simulation. This person observes the situation and tries to estimate the current level of stress or workload according to his or her impressions, knowledge and competence.

Primary / Secondary Task Performance assessment estimates the current level of workload by measuring specific task performance indicators (e.g. the time needed to complete a task).

Technical measurements are not yet fully investigated for the purpose of measuring stress or workload in an ATC situation. However, this category summarizes all thinkable technical approaches like measuring heart rate or blood pressure or other physiological signals of the human body or of human behavior. Provided that a reliable asset of technical measurement is available, it would be completely free of subjectivity. Further, it would be based on specific technical settings and pre-conditions, which provides a high level of comparability and allows an accurate repeatability. A disadvantage would be that detailed knowledge is needed how these measured values are connected with the corresponding level of workload / stress and what are the determining factors of this relation.

In the following sections, specific examples of different methods of workload / stress measurement are given. These are either well-established or currently subject to research. Other examples which shall only be mentioned at this point are the Subjective Workload Assessment Technique (SWAT), Subjective Workload Dominance Technique (SWORD), Malvern Capacity Estimate (MACE) or the Cognitive Task Load Analysis (CTLA) [8].

3.1. Instantaneous Self-Assessment

Instantaneous Self-Assessment (ISA) consists of continuous interrogation of the operator of a system or a person who is working on a defined task. This interrogation is done in defined time intervals and delivers snapshots of the level of workload experienced since the last interrogation [9].

This workload assessment technique is very flexible and can be paper-based or supported by specific ISA tools, e.g. applications available for iPhone and Android [10].

ISA uses a 5-step-scale from 1 = “under-utilised” to 5 = “excessively busy” [9].

Practically, the ISA workload measurement has the following advantages:

- It provides workload measurement in short time intervals, creating a workload profile during the completion of a task
- Results are immediately available and do not need to be worked up before analysis
- It is very simple, flexible and suitable for almost all tasks

On the other hand, the ISA workload measurement has the following disadvantages:

- ISA needs the cooperation of the person under assessment and can therefore contribute to or raise the experienced level of workload
- ISA itself can constitute an additional task of less importance for the person under assessment and may tempt him to ignore it or to give lukewarm ratings if the primary task is too demanding
- It is a subjective measurement; the way workload is perceived may vary from day to day or from person to person; it is also thinkable that an experience made during the beginning of the task influences some or all of the following ratings although each rating shall only determine the level of workload since the last interrogation. As a result, the level of workload may be over- or underestimated.

[9][11].

3.2. Cooper-Harper Scale

The Cooper-Harper Scale (CHS) was introduced in 1969 as a usability rating scale for aircraft handling. The CHS uses a decision tree for a rough scaling while a ‘turnout’ of this decision tree leads to a detailed scaling. FIGURE 1 shows the basic structure of the original Cooper-Harper-Scale [12].

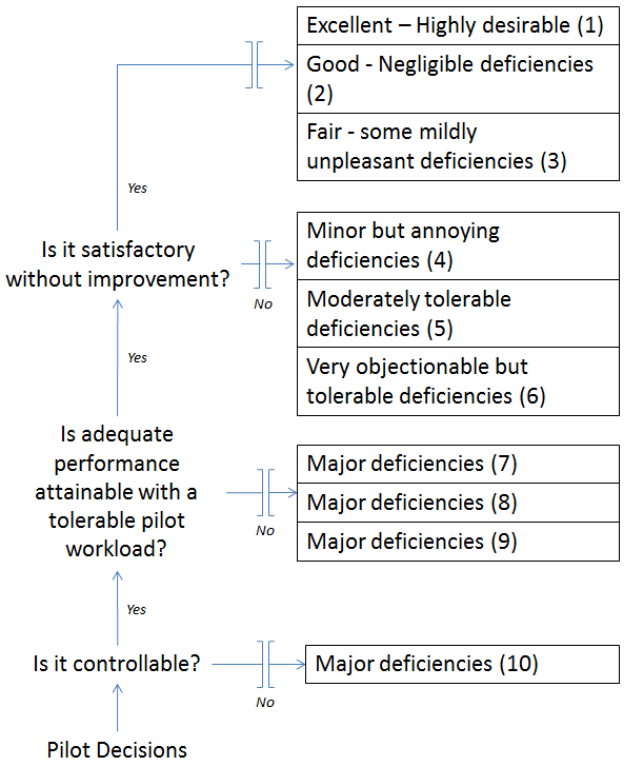

FIGURE 1. Basic structure of the original Cooper-Harper-Scale for aircraft handling

Although the Cooper-Harper Scale was developed as a scale for subjective aircraft handling ratings it can be used as basis and inspiration for workload rating scales in ATC simulations. Subjective rating scales as well as rating scales for a third-person assessment can be derived from it using the same decision-tree structure (in the following referred to as ‘modified Cooper-Harper Scale’ (MCH)). A third-person assessment using a MCH can be used to deliver a workload profile if done in a defined time interval.

Practically, when used in ATC simulations, a MCH has the following advantages:

- It is very flexible while providing the possibility to customize the scale to the experimental needs.
- It is more fine-graded than ISA (Scale ranges from 1 to 10 instead of 1 to 5).

On the other hand, the MCH has the following disadvantages:

- If used as a subjective scale, the MCH has disadvantages similar to ISA.
- The effort to customize the MCH to experimental needs is higher compared to other workload assessment methods such as ISA.

In recent DLR research activities, a MCH was used during the validation of a Multiple Remote Tower setup [4].

3.3. NASA TLX

The NASA Task Load Index (NASA-TLX) is a multi-dimensional scale which tries to measure several dimensions of workload separately and combines them to an overall workload score. The following aspects of workload are measured [13]:

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration Level

The NASA TLX can be determined with paper-based questionnaires or computer-based online questionnaires (FIGURE 2).

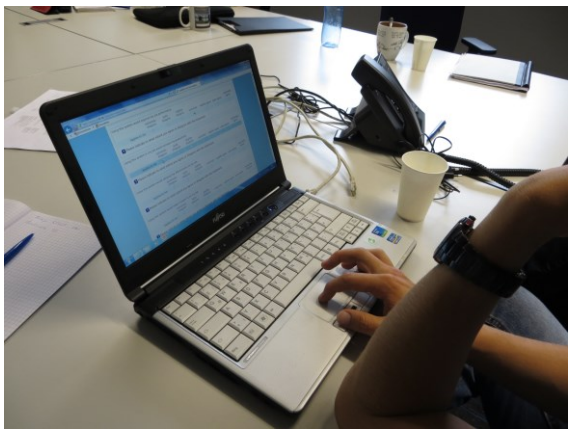

FIGURE 2. Online Questionnaire at DLR's ATM Validation Center

Practically, when used in ATC simulations, the NASA-TLX has the following advantages:

- It can be applied to all thinkable tasks or a specified group of tasks, to a specified part of or to a whole simulation
- It is usable for a broad field of tasks
- It uses multiple dimensions and eliminates subjective impressions to a certain extent.

On the other hand, the NASA-TLX has the following disadvantages:

- It takes up to several minutes to make one assessment; using it as a method to obtain a workload profile is possible but not appropriate for real-time ATC simulations as it would significantly interfere with the ATC tasks
- Therefore it is preferably used as post-simulation workload measurement; the obtained rating is valid for the whole task.

3.4. Secondary Task

During the validation campaign of the project AcListant, which was performed in DLR's ATM Validation Center in Braunschweig in 2015 and 2016, a secondary task was used to determine the overall workload in a simulation beside workload measurement with ISA. This secondary task was to sort cards of a card game. A third person measured the time needed until this secondary task was completed. The person under assessment was briefed that sorting the cards is a pure secondary task and shall only be continued if the workload of the primary task allows it [5].

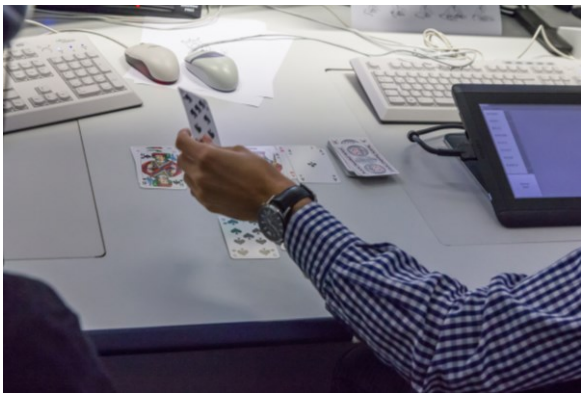

FIGURE 3. Sorting cards as secondary task to measure workload

Practically, when used in ATC simulations, this method has the following advantages:

- It is an objective method

On the other hand, this method has the following disadvantages:

- It delivers only very rough results
- It does not allow determining a workload profile.

3.5. EEG Measurements

In 2017 another trial campaign was conducted in DLR's ATM Validation Center with the goal to investigate if electroencephalography (EEG) can be used to directly measure the mental workload of an Air Traffic Controller during his work (FIGURE 4). It is expected that results are published soon by the Federal Institute for Occupational Safety and Health (BAuA), who led this trial campaign [14]. Similar studies for measuring the mental workload of pilots during flight using EEG were conducted in the past [15].

FIGURE 4. EEG Measurement in an ATC Simulation

It is expected that successful workload measurements based on EEG have the following advantages when used in ATC simulations:

- It is an objective workload measurement technique
- It doesn't need the direct cooperation of the person subject to measurements

On the other hand, this method has the following disadvantages:

- It requires a high technical effort
- The EEG signal is contaminated with other physiological signals such as muscle contractions
- The actual level of workload needs to be calculated from the obtained results; it is not directly visible.

4. STRESS MEASUREMENT IN GAMMA

Stress is considered a very complex human emotion which features several physiological reactions of the human body (e.g. heart rate or skin conductivity), which can easily be measured [16].

In the frame of the project Global ATM Security Management (GAMMA), seven security prototypes were developed. The 'Secure ATC Communications' prototype (SACom) combined several functionalities, such as speaker verification and stress detection by voice analysis, conformance monitoring, conflict detection and correlation. This prototype contained a so called stress detection module (SD), which was developed by the Slovak Academy of Sciences (SAV), Bratislava. The methods as well as the background of this approach are published in [17] and are not discussed further in this paper.

Stress measurement by voice characteristics using the SD is considered as another form of measuring physiological reactions of the human body in stressful situations.

This innovative method was subject to several validation activities in 2016 and 2017. The following work describes the approach and outcomes of a comparison of this method with selected workload measurement tools, because during normal operations the level of workload is seen as the most significant cause for stress in an ATC simulation (see section 2). Therefore it is expected that workload profiles show a similar behavior than stress profiles.

In the following, stress measurement using the SD of the SACom prototype is referred to as 'vocal stress measurement'.

5. APPROACH

The work described in this document had the objective to investigate the actual performance, the technical feasibility and also the influencing factors of the mentioned vocal stress measurement method when used in an ATC simulation.

To achieve this, a simulation campaign was conducted at DLR's ATM validation center in Braunschweig, Germany from 24th to 26th of January 2017. Four active air traffic controllers from the COOPANS cooperation (in detail: from Denmark, Sweden, Croatia and Ireland) were invited to take part in a series of five different approach control simulations per participant (20 simulation runs in total). All simulations were performed in DLR's Air Traffic Management and Operations Simulator (ATMOS).

The simulation scenarios modelled different levels of traffic from low to very high including changes of the level of workload in the same scenario. The simulated ATC environment was the approach control unit of Düsseldorf Airport while all approaching aircraft had to be picked up at the approach sector boundary and had to be guided to runway 23R for landing. Five different simulation scenarios were designed running 45 minutes to 65 minutes per scenario.

All aircraft were steered by so called pseudo-pilots who were sitting at prepared pseudo pilot working positions. In all simulations two pseudo pilots were in charge of all aircraft. The air-ground radio communication was simulated with the open-source radio communication simulator YADA, which is based on Voice-over-IP (VoIP).

In every simulation run, three selected workload assessment methods, which allow measuring the current level of workload or stress in real time, were applied in parallel:

- Instantaneous self-assessment (ISA) supported by an electronic input device
- Third-person assessment by using a modified Cooper-Harper-Scale (MCH)
- Vocal stress measurement (SD), which is in the focus of this research.

The ISA and the Third-person assessment with the MCH were applied at the same time interval of three minutes to have values at nearly the same timestamps of a simulation.

The MCH which was used in these trials was derived from another MCH which was used during recent Remote Tower Validation Trials of DLR [4]. The following ranking was defined for this MCH:

Normal ATC operations:

1) No problems: Desired performance can easily be achieved, the level of workload is low

2) Simple: Desired performance can be achieved, the level of workload is appropriate

3) Demanding: Increased effort is necessary to achieve the desired performance

Negative impact on ATC efficiency:

4) Little, but disturbing delays: Air Traffic Controller (re)acts with little delay

5) Medium loss of capacity: Situation leads to moderate delays (increased separation) in handling the traffic

6) Significant difficulties without a loss of safety: Situation leads to significant delays to work on pending tasks. This ranking is considered as 100% workload.

Negative impact on safety:

7) Problems in predicting the development of the situation: Air Traffic Controller starts to flounder, guides the air traffic infrequent and volatile, doesn't plan target-oriented

8) Problems in processing information: Air Traffic Controller is unable to establish a complete traffic picture, mixes up information, must correct himself very often

9) Problems in acquiring information: Air Traffic controller needs to ignore parts of his area of responsibility or information, Air Traffic controller forgets airplanes

10) Impossible: Traffic situation is no longer controllable

This MCH was filled every three minutes by an observer who had the possibility to monitor the traffic and to listen to the radio communication. All persons who filled the MCH had comprehensive ATC knowledge and operational experience.

Although the software application of the vocal stress measurement algorithms allows setting a sensitivity level, the default setting (1.00) was used because of a lack of experience with this new method.

6. RESULTS AND DISCUSSION

In 2 of 20 simulations the Instantaneous Self-Assessment failed due to technical problems, which means there is no data available for these runs. However, third-person assessments using the MCH were successfully completed and the vocal stress measurement was running for all simulation exercises. In 6 simulations the vocal stress measurement software did not detect any stress signals in the voice of any speaker, the stress detection score was always zero in the whole scenario.

Table TAB 1 gives an overview of conducted simulations, involved controllers and remarks.

Date	Run ID	ATCo ID	Scenario ID	Remarks
24th Jan	1	L33	PF2	
	2	L33	F2	No ISA data
	3	L34	PF2	
	4	L34	F2	
25th Jan	5	L33	F1	SD always 0
	6	L33	F3	
	7	L33	PF1	SD always 0
	8	L34	F1	
	9	L34	F3	
	10	L34	PF1	
	11	L35	PF2	
	12	L35	F2	SD always 0
	13	L36	PF2	
	14	L36	F2	
26th Jan	15	L35	F1	SD always 0
	16	L35	F3	SD always 0
	17	L35	PF1	SD always 0
	18	L36	F1	
	19	L36	F3	No ISA data
	20	L36	PF1	

TAB 1. Simulation Campaign Overview

In the following sections, at first the ISA results are compared with the results of the third-person assessment using the MCH. Thereafter, these results are further compared with the stress detection results. Selected detailed scores are presented later. Finally, a short assessment of the stress detection scores of pseudo pilots is done.

6.1. Comparing ISA and MCH-Assessment

In order to be able to directly correlate the 5-step ISA scale results with the 10-step MCH results, both scales have to be harmonized. As mentioned before, a workload of 100% is assumed when the ISA score is 5 while a MCH score is 6. Therefore, the following calculations are made to convert scale results into a level of workload (W):

(1) $W_{ISA} = V_{ISA} * 0.2$

(2) $W_{MCH} = V_{MCH} * 0.16$

Where

W_{ISA} is the level of workload according to ISA,

W_{MCH} is the level of workload according to the third-person assessment with the MCH,

V_{ISA} is the specific ISA value determined during the simulation,

V_{MCH} is the specific MCH value determined during the simulation.

For every simulation run the workload profiles according to ISA and MCH were analyzed manually to find a possible correlation between both graphs. The following categorization was used to roughly describe the degree of correlation for every simulation run:

- Unknown: A comparison cannot be made because one of both graphs was missing
- No: Both graphs mostly do not show a similar behavior
- Maybe: Major parts of the graphs show a similar behavior while there are nevertheless considerable parts not showing a similar behavior
- Clear: Both graphs show a similar behavior with no or little parts showing differences

Table TAB 2 gives an overview of the results of this analysis. Four controllers (L33-L36) took part; five different runs per controller were performed.

ATCo ID	Unknown	No	Maybe	Clear
L33	1	1	1	2
L34	0	1	1	3
L35	0	1	1	3
L36	1	0	4	0
Total	2	3	6	9
Total %	10%	15%	30%	45%

TAB 2. Comparison ISA with MCH

It can be seen that – when excluding the two runs which could not be rated – in 15 of 18 simulation runs (83.3%) there was at least a rough correlation between both graphs showing an increased or decreased level of workload or peaks or lows at the same timestamps. This basically confirms the usability of these methods, but this also shows that even well-established methods just reach a medium level of reliability and may even produce completely different results (clear correlation in less than 50% of the trials). The value of both methods for the purposes stated in the introduction chapter of this paper as well as for the following work can be seen as limited, which constitutes an additional challenge.

6.2. Comparing Vocal Stress Measurement with ISA and MCH-Assessment

As a next step, the results were compared with the output of the vocal stress measurement introduced in chapter 4.

The vocal stress measurement delivers a stress score for every utterance, so not only the scores by themselves but also their quantity of appearance play a role. An increasing level of stress would be visible through higher stress scores and / or more frequent stress measurements with a stress score >0.

The vocal stress measurement is monitoring all speakers taking part in the air traffic control simulation, which are one air traffic controller and 2-3 pseudo pilots. These speakers were monitored via separate audio channels, therefore it can be clearly identified which stress score belongs to which speaker. Nevertheless, the measured stress scores of the air traffic controller are at first of interest.

For every simulation the workload according to ISA, the workload according to MCH and the stress detection scores >0 were visualized in separate diagrams as a function of time. Every diagram was then analyzed manually if a correlation between the stress detection graph and any other graph is visible (see FIGURE 5, the comparison between ISA and MCH was done in section 6.1).

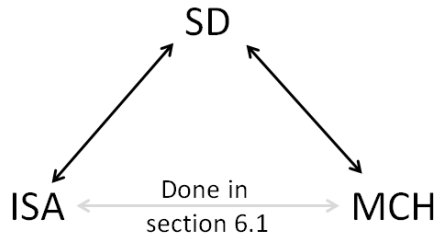

FIGURE 5. Principle of Comparison

The following categorization was used to roughly describe the degree of correlation for every simulation run (similar to the procedure used previously):

- No output: A comparison cannot be made because the vocal stress measurement always produced a stress score of zero for the air traffic controller in this exercise run
- No: the graph of the stress detection score does neither show a similar behavior with the ISA workload profile nor with the MCH workload profile
- Maybe: Major parts of the graphs show a similar behavior either with the ISA or with MCH or both while there are nevertheless considerable parts not showing a similar behavior
- Clear: SD graph shows a similar behavior either with the ISA or with the MCH or with both while no or little parts show deviations

Due to the low technology readiness level (TRL) of the vocal stress measurement this rough ranking shall allow to determine if this new approach shows similar profiles in relation to ISA or MCH or both. If so, further research is desirable.

Table TAB 3 gives an overview of the results of this analysis. As a reminder, four controllers (L33-L36) took part; five different runs per controller were performed.

ATCo ID	No output	No	Maybe	Clear
L33	5			
L34	3		2	
L35	5			
L36		2	2	1
Total	13	2	4	1
Total %	65%	10%	20%	5%

TAB 3. Comparison SD with ISA and MCH

At first it can be seen that the stress detection always produced no output for controller L33 and controller L35 while it produced output in 2 of 5 exercises for controller L34 and in 5 of 5 exercises for controller L36. A direct conclusion is that the stress detection does obviously not work with the same reliability and sensitivity for all persons under assessment. One reason for this may be that controllers are used to stressful situations, especially with growing experience, and therefore do not always show their stress in their voice. In order to achieve results which are comparable from person to person a way must be found to adapt the sensitivity of the vocal stress measurement continuously. If such a way cannot be found the stress detection can only be used to determine stress profiles showing an increase or decrease in stress without knowing the exact level.

In general, the vocal stress measurement did not produce scores >0 for the ATCo in 65% of all exercise runs, which clearly indicates that the sensitivity setting of the SD was too low.

Due to the sensitivity settings, all runs with no SD output are excluded for the following analysis. In 5 of 7 remaining exercise runs (71.4%) there was at least a rough correlation between SD and MCH or ISA graphs showing an increased or decreased level of workload or peaks or lows at the same timestamps. This clearly indicates the potential to use this new technique in ATC simulations, provided that the problem to properly adapt the sensitivity to the person under assessment is solved. In the following a textual description of the results is provided; selected graph examples are contained in section 6.3.

For controller L34, in two exercise runs the peaks of the stress detection graphs coincide with the peaks in the MCH graph while during the rest of the exercise the stress detection graph showed a score of zero.

For controller L36,

- In one exercise run (exercise run no. 14) the stress detection graph showed a quite clear correlation with the MCH graph. However, when comparing MCH and ISA it was found that they 'maybe' correlated but significant parts of the ISA and MCH graphs do not show a similar behavior. The graphs of this exercise run are displayed in section 6.3.

- In one exercise run (exercise run no. 13) the stress detection scores showed some correlation with the ISA and MCH graphs (rated as 'maybe'). The correlation between ISA and MCH was also rated as 'maybe'. This graph is also shown in section 6.3.

- In one exercise run, the stress detection score showed often, but not always correlation with MCH only (rated as 'maybe'). The correlation between ISA and MCH was also rated as 'maybe'.

- In one exercise run, the MCH graph showed a profile which can also be assumed to be present in the stress detection scores, but in a very vague way. Therefore the correlation between the stress detection graph and the MCH graph was rated as 'no'. Unfortunately, for this exercise run no ISA data is available due to a technical failure during the exercise ('unknown').

- In one exercise run, about one third of the stress detection graph seems to be correlated with the MCH graph while two third of the graphs don't show the same behavior. Therefore the total rating was 'no'. The correlation between ISA and MCH was rated as 'maybe'.

As a summary, it can be seen that the vocal stress measurement much better matches with the MCH than with the ISA graphs. Provided that the stress measurements are correct this could mean that the used third-person assessment using a MCH is more reliable and expressive than a self-assessment.

6.3. Selected Graph Examples

As a detailed description of all results would go beyond the constraints of this paper without additional value, the specific measured profiles of two selected exercises are published below.

FIGURE 6 shows the measured graphs of exercise run no. 14 (L36, scenario F2). The x-axis shows the simulation timestamp (seconds elapsed since midnight), while the interval between two marks is exactly 600 seconds / 10 minutes. The y-axis either shows the level of workload (ISA and MCH) or the measured stress detection score.

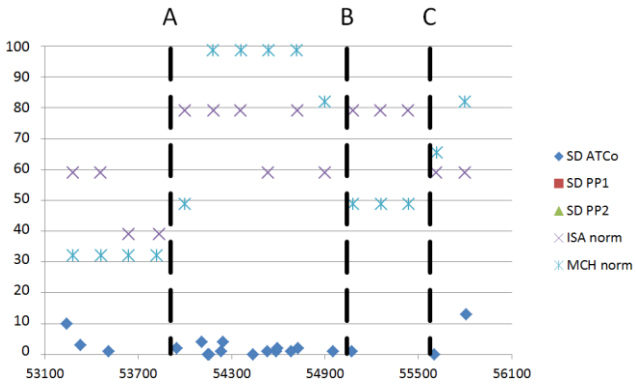

FIGURE 6. Results of exercise run no. 14

In this exercise run, the stress detection of pseudo pilot 1 (PP1) and pseudo pilot 2 (PP2) always delivered a score of zero.

It can be seen that the ISA graph and the MCH graph shows the same behavior between timestamp 53600 and timestamp 55000 while the behavior is completely different between timestamp 53100 and 53600 as well as after timestamp 55000. The correlation between ISA and MCH was rated as 'maybe'.

One very interesting fact is that all three measurements indicate an increasing workload at timestamp 53900 (marked with line 'A'). At line 'B' the MCH graph as well as the stress detection graph indicates a decreasing level of workload. Both graphs show again an increasing level of workload at line 'C'. The correlation between MCH and stress detection was rated as 'clear'.

FIGURE 7 shows the measured graphs of exercise run no. 13 (L36, scenario PF2). Again, the x-axis shows the simulation timestamp (seconds elapsed since midnight); the y-axis either shows the level of workload (ISA and MCH) or the measured stress detection score.

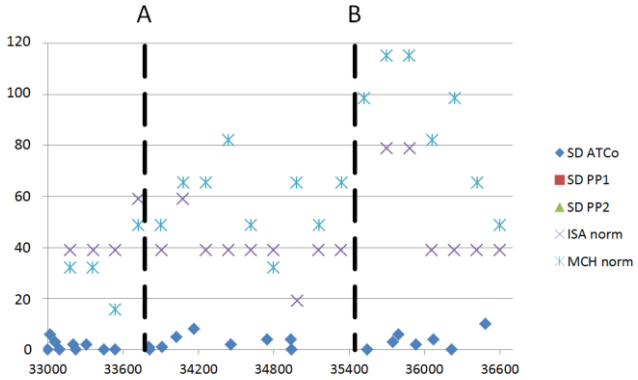

FIGURE 7. Results of exercise run no. 13

Like in run no. 14, the vocal stress measurement always delivered a score of zero for both pseudo pilots.

It can be seen that the MCH graph fluctuates more than the ISA graph. The correlation between both graphs seems to be less developed compared to exercise run no. 14. However, both graphs indicate an increased workload at line 'A' and line 'B'. The correlation between ISA and MCH was rated as 'maybe'.

Again, also the vocal stress measurement roughly indicates an increasing workload at line 'A' and 'B'.

6.4. Stress Detection Results of the Pseudo Pilots

One interesting outcome was that sometimes also the stress detection scores of the pseudo pilots seem to be correlated with the workload of the controller.

FIGURE 8 shows the graphs of exercise no. 3 (L34, scenario PF2). Again, the x-axis shows the simulation timestamp (seconds elapsed since midnight); the y-axis either shows the level of workload (ISA and MCH) or the measured stress detection score.

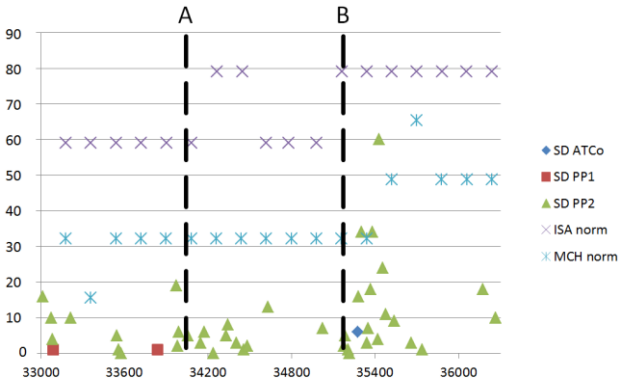

FIGURE 8. Results of exercise run no. 3

In this exercise run, the vocal stress measurement produced output for all participants (the air traffic controller and both pseudo pilots) while by far the most output is produced for pseudo pilot two.

It can be seen that the stress detection output of pseudo pilot 2 clearly correlates with an increased ISA rating at line 'A' as well as line 'B'. During some minutes behind line 'B' there could also be a correlation with the MCH graph. In addition to that, there was a stress indication of the ATCo also falling together with line 'B'.

More of these kinds of correlations between the workload of the air traffic controller and the stress score of the pseudo pilots were suggestively found in exercise runs no. 1, 2, 6 and 11.

In an ATC Simulation with a constant number of pseudo pilots (in this study there were always 2 pseudo pilots involved) the workload and stress of the Air Traffic Controller might in fact be in correlation with the workload of the pseudo pilots because:

- An increased number of aircraft involved in the traffic situation directly means also an increased number of aircraft to be handled by a pseudo pilot,
- Mistakes of the controller (e.g. information mix-up) immediately lead to inquiries, wrong inputs or necessary corrections for the pseudo pilots,
- A frantic way of communicating with the pseudo pilots might directly lead to an increased overall tension in the whole simulation.

7. CONCLUSION

The work on hand brought very interesting new insights and new experiences with stress measurement by voice in an ATC simulation. Although the stress detection did often not produce stress scores >0 there were some results showing that using this technique for objective stress measurement in real time is basically possible.

One problem which was identified is the correct setting of the sensitivity of the vocal stress measurement. As it often produced a stress score of zero the sensitivity setting obviously was too low. One of the next steps therefore can be to repeat this campaign using a higher sensitivity.

Another problem was to refine the gathered data in order to achieve normalization on one hand and comparability between different persons under assessment on the other hand. The vocal stress measurement was found to be very dependent on the person and on his or her resistance against stress. Further dependencies such as on the audio quality of the voice communication channels are obvious.

Although the vocal stress measurement does not directly measure the level of workload it can be expected that with further investigation, research and development (especially to solve the mentioned problems), this technique has the potential to close up with well-established workload measurements, to overtake them or to supplement them.

As a final remark, this work inspired the idea that workload measurement may also be possible by considering the workload and the performance of the pseudo pilots in an ATC simulation.

8. ACKNOWLEDGEMENT

The authors would like to thank all GAMMA consortium members that contributed to this paper through stimulating discussions around the contents presented.

9. DISCLAIMER

The views expressed herein are the authors' own and do not reflect a GAMMA consortium and/or their employers' position or policy.

10. REFERENCES

[1] International Civil Aviation Organization (ICAO), Doc 4444, "Procedures for Air Navigation Services – Air Traffic Management", 15th Edition, 2007
[2] G. Tobaruela, A. Majumdar, W. Y. Ochieng, "Identifying Airspace Capacity Factors in the Air Traffic Management System", ATACCS, 2012
[3] R. Ehrmanntraut, S. McMillan, "Airspace design process for dynamic sectorization", 2007
[4] M. Peters, A. Papenfuss, "Analysis of Critical Situations at Remote Tower Operated Airports", 2012
[5] H. Helmke, O. Ohneiser, J. Buxbaum, C. Kern, „Increasing ATM Efficiency with Assistant Based Speech Recognition", 2017
[6] H. Zeier, "Workload and psychophysiological stress reactions in air traffic controllers", 2007
[7] T.C. Hankins, G.F. Wilson, "A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight", 1998
[8] N.A. Stanton, P.M. Salmon, L.A. Rafferty, G.H. Walker, C. Baber, D.P. Jenkins, "Human Factors Methods: A Practical Guide for Engineering and Design", 2005
[9] Information Society Technologies Programme, Deliverable D2.2.6, "Subjective Assessment Methods for Workload", January 2006
[10] http://www.think.aero/isa/, 2017
[11] https://ext.eurocontrol.int/ehp/?q=node/1585, 2017
[12] G. E. Cooper, R. P. Harper Jr., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities", National Aeronautics and Space Administration (NASA), April 1969
[13] NASA Ames Research Center Human Performance Research Group, "NASA Task Load Index (TLX)", 1986
[14] https://www.baua.de/DE/Themen/Arbeit-und-Gesundheit/Psychische-Gesundheit/Mentale-Gesundheit-und-kognitive-Leistungsfaehigkeit/Mentale-Beanspruchung.html, 2017
[15] C. Dussault, J.-C. Jouanin, M. Philippe, C.-Y. Guezennec, "EEG and ECG Changes During Simulator Operation Reflect Mental Workload and Vigilance", 2005
[16] R.S. Lazarus, J.C. Speisman, A. Mordkoff, „The Relationship Between Autonomic Indicators of Psychological Stress: Heart Rate and Skin Conductance", 1963
[17] M. Rusko, M. Finke, "Using Speech Analysis in Voice Communication", 2016