Fifteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)

Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload

Hartmut Helmke, Matthias Kleinert, Nils Ahrenhold, Heiko Ehr, Thorsten Mühlhausen, Oliver Ohneiser
German Aerospace Center (DLR), Braunschweig, Germany
First.Lastname@dlr.de, Thorsten.Muehlhausen@dlr.de

Lucas Klamert
Austro Control, Vienna, Austria
Lucas.Klamert@austrocontrol.at

Petr Motlicek[†], Amrutha Prasad[†,§], Juan Zuluaga-Gomez[†,¶]
[†] Idiap Research Institute, Martigny, Switzerland
[§] Brno University of Technology, Brno, Czech Republic
[¶] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
First.Lastname@idiap.ch, juan-pablo.zuluaga@idiap.ch

Jelena Dokic, Ella Pinska Chauvin
Integra Consult A/S, Stubbeled 2, 2950 Vedbaek, Denmark
JDJ@integra.dk, EPC@integra.dk

***Abstract*—Air traffic controllers (ATCos) from Austro Control together with DLR quantified the benefits of automatic speech recognition and understanding (ASRU) on workload and flight safety. As the baseline procedure, ATCos enter all clearances manually (by mouse) into the aircraft radar labels. As part of our proposed solution, the ATCos are supported by ASRU, which is capable of delivering the required inputs automatically. The ATCos are only prompted to make corrections, when ASRU provided incorrect output. Overall amount of time required for manually inserting clearances, i.e., by clicking and selecting the correct input on the screen, reduced from 12,800 seconds during 14 hours of simulations time down to 405 seconds, when ATCos were supported by ASRU. A reduction of radar label maintenance time through ASRU might not be surprising given earlier experiments. However, a factor greater than 30 outperforms earlier findings. In addition, this paper also considers safety aspects, i.e., how often ATCos support provided an incorrect input into the aircraft radar labels with and without ASRU. This paper shows that ASRU systems based on artificial intelligence are reliable enough for their integration into air traffic control operations rooms.**

Keywords—automatic speech recognition, automatic speech understanding, situation awareness, saftety, artificial intelligence, human factors, air traffic controller's workload

I. INTRODUCTION

Automatic speech recognition (ASR) systems are widely used in everyday life (e.g., Siri®, Alexa®). ASR has been used in air traffic control (ATC) training simulators to replace expensive simulation pilots since the late 1980s. The German air navigation service provider DFS has been focusing on deployment of ASR for more than 15 years for their trainees in the simulator environment [1]. Moreover, the European Commission is funding the integration of ASR into air traffic management applications at least since 2016, starting with the project "Machine Learning of Speech Recognition Models for Controller Assistance," MALORCA [2]. However, even though many prototypes were already implemented in laboratory environments over the past, ASR has not yet been productively used in an operational environment, i.e., in an operations room (ops room).

A. Speech Recognition is not Speech Understanding

One of the reasons for not deploying ASR systems in ops rooms until now is that speech recognition is not the *end of the story,* because ASR is just the process of translating the voice signal to a sequence of recognized words. To make actual use of the information carried in such words, speech understanding is also required. If the air traffic controller (ATCo) e.g. says "*speed bird two thousand one eight zero knots until four to tower eighteen seven bye*" and even having a perfect ASR system, which correctly recognizes each of the 15 words, any digital assistant still does not understand the semantics of this utterance, i.e., callsign and command elements such as values and units. Therefore, the term automatic speech recognition and understanding (**ASRU**) is used in the remaining part of the paper. ASRU comprises speech-to-text (automatic transcripts) as well as text-to-concepts (automatic extraction of meanings; annotation). Understanding becomes even more important in case not all of the spoken words are correctly recognized by ASR (which is obviously expected). Humans are also prone to errors as tasks become more and more cognitively demanding, but are still able to grasp the meaning of the communication most of the time. ATCo-pilot communication is of course standardized by ICAO [3], but there is a big difference between what is written in ICAO documents and what happens in real life communication between ATCos and pilots, which makes understanding even more challenging.

B. Research Question

Recently, the STARFiSH project [4] demonstrated that acceptable command recognition rates are possible for Frankfurt apron simulation environment. Command recognition rates of 90% were achieved. Are 90% enough? A simple answer would be "yes": The ATCo only needs to manually input/correct one of ten commands. Even if this correction requires doubled effort compared to the input without ASRU support, the workload for manual input is still reduced by a factor of four. However, the main question is, what happens, if the ATCo does not detect that there is an error in the ASRU output, i.e., the content in the radar label is wrong or incomplete. Derived from this issue, one also

Supported by SESAR Joint Undertaking, grant agreement No 874464.

needs to ask if a very good ASRU performance can even result in an over-trust into the system.

C. Derived Research Questions

- The ASRU availability serves as independent variable for our human-in-the-loop real-time simulation study. Another independent variable is whether an ATCo starts with or without ASRU support. Normally, as shown in [5], ATCos perform better when they are already familiar with tools and scenarios. How can we compensate unintended sequence effects, i.e., the second independent variable?
- If ASRU errors are not detected and corrected by the ATCo, how can we reduce the risk that wrong or missing inputs remain undetected and uncorrected?
- First, however, we need to calculate the metrics. How can we effectively and even more important efficiently analyze, which clearances are given, which ones are correct in the radar labels, and which ones are missing or have wrong values?

D. Paper Structure

After presenting related work in section II, section III describes the validations, being performed with 12 ATCos from Austro Control (ACG) in DLR's lab environment from September to November 2022. As part of the baseline runs, ATCos enter all clearances manually. In solution runs, ATCos are supported by ASRU. The ASRU architecture is described together with its performance in section IV. Section V contains objective validation results on ATCo performance and safety questions, whereas section VI presents subjective feedback from ATCos from questionnaires and together with section V gives answers to the derived research questions. Section VII concludes. The appendix compares our sequence compensation approach to a multi-factor ANOVA approach.

II. Background

After focusing on replacing ATC simulation pilots by ASR [6] in the last decades, the evaluation of ATC workload using ASR data [7], [8] was a subsequent step, during which the limitations of grammar-based ASR approaches became apparent. In 2018 Airbus launched a speech recognition challenge to encourage the development of ASR for ATC scenarios [9], which provided academia and industry with access to real life, manually transcribed ATC utterances. Although many applications achieved acceptable recognition performance in terms of word error rate, it became clear that the lack of context information resulted in inadequate results especially at the conceptual level. One promising approach to improve ASR performance is the use of context knowledge related to expected utterances, with early attempts dating back to the 1980s [10], [11]. This context knowledge can significantly reduce the search space and lead to fewer misrecognitions [12]. In this course, the benefits of using context information for pre-processing versus using context for post-recognition have been analyzed [13].

Later, the context was extended by generating it also with an assistance system; here an arrival manager [14]. The extension started with a study in 2011 [15]. In a pilot study with a limited set of callsigns and commands, [16] reported command (recognition) error rates below 5%. They used an acoustic model derived from the Wall Street Journal recognition corpus. In 2016, it was shown that using ASRU with using context from an assistant system can significantly reduce ATCo workload, which translates into fuel burn reduction and an increased runway throughput. These results were quantified in [5] and [17], respectively. MALORCA project aimed at automatically adapting the speech recognition building blocks to different approach areas. The learning mechanism of command prediction, i.e., the relevant part of the assistant system, was described in [18]. Automatic adaptation results for Vienna and Prague approach areas from ATCo-pilot speech recordings and the corresponding radar tracks were presented in [19].

ASRU was also used to automatically detect readback discrepancies in the US tower control [20] and Icelandic enroute airspace [21]. Reference [22] presented a safety monitoring framework that applied ASR to flight conformance monitoring and conflict detection. Finally, the accuracy and robustness achieved by mature in-domain ASR has enabled mining of large-scale ATC communication recordings for post-operational analyses [23], [24], [25]. The approach procedure deployment across the U.S. National Airspace System using automatically transcribed radio communications in post analyses was investigated in [26]. Similarly, the quantity of pilot weather reports delivered over the radio against the quantity of pilot reports manually filed during the same time frame was compared in [27]. Another work focused on detecting who is speaking (ATCo or pilot) purely based on transcripts generated by an in-domain ASR system was recently proposed on [28]. Likewise, callsign recognition based on ASR transcripts has also been explored in [29].

The AcListant®-Strips project has demonstrated that ASRU supported radar label maintenance can reduce fuel burn by 50 to 60 liters of kerosene per flight [5]. The next step was to integrate an industrial ASR prototype into the TopSky controller working position (CWP) in SESAR 2020 solution PJ.16-04. TopSky is an operational radar screen, developed by Thales LAS. Austro Control uses the Thales TopSky System [30] since 2014 to manually input given clearances into the radar label. The feedback of the participating ATCos related to a usability and integration of ASRU into the human machine interface (HMI) was very positive [30]. Later, it was decided to repeat the experiment with an advanced ASRU system integrated into an HMI being adapted to Vienna approach area. Open questions were still posed: e.g., "*what happens when the ATCo does not recognize when ASRU fails and a misrecognition is not corrected in the radar labels?*". All these questions are addressed by a validation exercise in the SESAR funded project "PJ.10-96-W2 HMI Interaction modes for ATC center" presented in the next section.

III. Validation Exercise: Setup and Execution

The main purpose of the exercise was to quantify the benefits of ASRU with respect to safety and ATCo workload. The benefit should arise from supporting approach ATCos with aircraft radar label maintenance. Therefore, in the baseline runs the modalities of a "typical" manual *mouse-only* input for ATCo commands into the human machine interface were compared to a setup with *ASRU + mouse* input in the solution runs. First, the flow of ATCo commands into the HMI is detailed. Second and

third, the experiment setup as well as the scenarios and configurations for our experiments are explained. Fourth, the study participants are described.

A. ATCo Command HMI Input

If the ATCo clicks on one of the nine underlined radar label cells shown in Figure 1, a drop-down menu opens to enter the given clearance values, e.g., for altitude, speed, heading, waypoint etc.

Figure 1. Left: Interactive radar label cells (red underlined); Right: Drop-down menu to enter given transition name, which opens with a click on the *transition* field (in this example the value *MABOD*). Yellow cross for rejecting all and green checkmark to accept all recognized values in the label.

In solution runs, the ATCo command values are extracted from the radio telephony utterance and automatically appear in the label cells in purple, not shown in Figure 1, but later in Figure 7. Thus, the ATCo only needs to check and confirm with a mouse click or corrects values in seldom cases of misrecognition. Accepted cell values will turn into light green as soon as the ATCo accepts them with a click on the green checkmark in the first label line. The command values are also automatically accepted after 10 seconds if the ATCo does not reject or correct them, which was a result of the AcListant®-Strips project [5] and also of solution PJ.16-04-ASR "CWP HMI" [30]. Nevertheless, the ATCo can manipulate cell values at any time.

B. Setup of Experiments

The exercise included iterative validation trials. Three pre-validation rounds took place in October 2021, December 2021, and in March 2022. Figure 2 shows the basic validation setup for the CWP in DLR's simulation facility, ATMOS (Air Traffic Management and Operations Simulator). The ATCos communicated with simulation pilots in another room via voice-over-IP (VoIP). The headset microphone signal of the ATCo served as the input signal for the speech recognition engine.

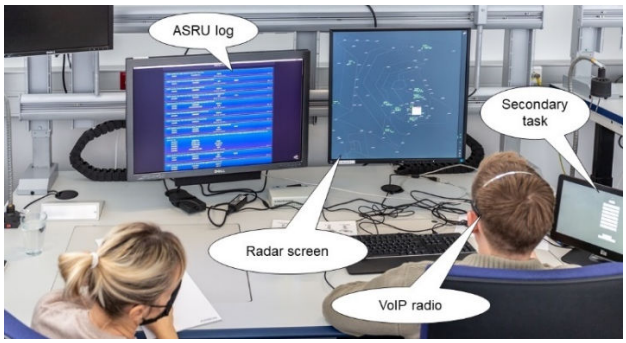

Figure 2. Basic validation setup during final trials.

The main radar screen (right square monitor in Figure 2) displayed the simulated Vienna (LOWW) airspace. The display consisted of the Vienna terminal maneuvering area (TMA) structure with the area navigation (RNAV) holding points BALAD, PESAT, MABOD, and NERDU, associated waypoints and RNAV routes, and radar data of inbound aircraft to runway 34 (see also Figure 4). The left monitor (ASRU log) shows the recognized ATCo utterances with the extracted concepts: callsign, command type, value etc. in the solution runs with ASRU support. The small touch screen monitor to the right provides the interface for a secondary task.

The ISA (Instantaneous self-assessment of workload) interface is integrated into the radar screen and requires ATCo feedback regarding individual self-assessed workload for the last five minutes [31]. Subjective rating measures on amongst other aspects workload and situation awareness were captured after each simulation run using NASA-TLX (National Aeronautics and Space Administration Task Load Index) [32], Bedford Workload Scale [33], SUS (System Usability Scale) [34], CARS (Controller Acceptance Rating Scale) [35], and the three SHAPE questionnaires (Solutions for Human Automation Partnerships in European ATM) [36] – SASHA (Situation Awareness for SHAPE) ATCo, SATI (SHAPE Automation Trust Index), and AIM-s (Assessing the Impact on Mental Workload). A secondary task was evaluated to gather an objective measure for ATCo workload during the trials. The task is based on the Stroop test [37] implemented in the STARFiSH project [4]. A higher number of correct tests indicates more mental capacity available for the secondary task. Thus, less workload capacity is consumed by the primary task [38]. After ten minutes of a simulation run, the ATCos had to perform the secondary task for ten minutes in addition to controlling the traffic.

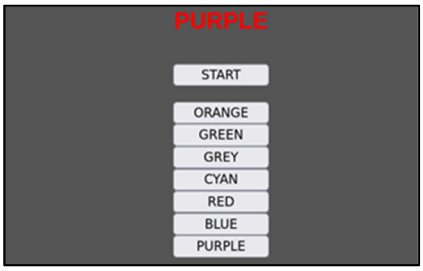

Figure 3. Stroop test screenshot.

When the user presses the "START" button (Figure 3), the app shows a word for a color printed in a different color. The user has to select the color of the print from a list of buttons labelled with the names of colors. The user has to select the word *RED* in Figure 3.

C. Scenarios and Configurations of Experiments

Two different scenarios were created: a medium density traffic scenario with 30 arrivals per hour and a heavy density traffic scenario with 42 arrivals per hour. The scenarios did not contain departures, overflights or other types of traffic TMA ATCos frequently deal with in reality. In the heavy density traffic scenario, the voice frequency was occupied by the ATCo 35.2% of the time (medium density scenario: 30.9%, average of all scenarios: 33.1%), plus roughly 45% of the time by the simulation pilots. The ATCo area of responsibility (see Figure 4) is comparable to a combined pickup/feeder sector in Europe, which corresponds to a combined feeder/final sector in the US. All scenarios and input modalities were trained in additional training runs, before

the solution and baseline runs started. Each ATCo started with the medium traffic scenario. Then a heavy traffic scenario followed, afterwards again the medium traffic scenario, and the last scenario was always with heavy traffic. 55% of the ATCos started with solution runs and 45% started with baseline runs. If an ATCo started with a solution run, s/he ended also with a solution run, i.e., in between were two baseline runs. The same applied the other way around, when the ATCos started and ended with the baseline runs. Each scenario lasted for 35 minutes. The validation team has heavily discussed whether also to counterbalance the order of traffic load. We decided to always start with the medium scenario to give the ATCos the opportunity for an additional training opportunity without calling this a training run. The main argument is presented in subsection V.A, because this approach enables us to compensate sequence effects.

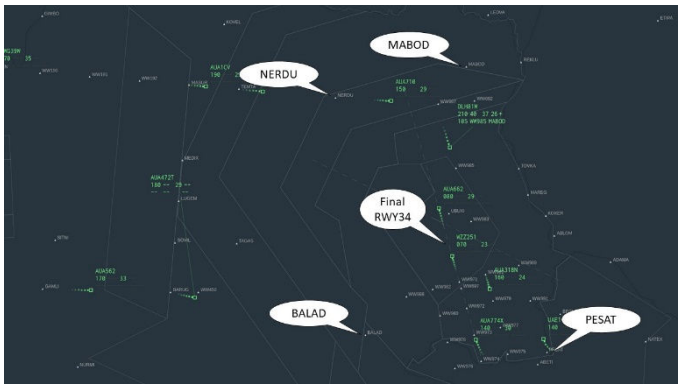

Figure 4. ATCo's area of responsibility with main holding points and final.

D. Study Participants

The validation exercises were performed between September, 14th and November, 3rd 2022 by 12 ATCos from Austro Control. One of them was female. The average age of participating ATCos was 32 years (*Standard Deviation SD=Sigma* = 7.3; age interval between 25 and 44 years). Their professional work experience was 8 years on average (*SD* = 6.8; experience interval between 1 and 20 years).

On each of the six validation days two ATCos were available. They started at 08:30 a.m. and ended around 04:30 p.m. While one of them was doing one of the four exercises (medium baseline, heavy solution, medium solution, heavy baseline) the other was doing the questionnaires and had free time to rest, respectively. Thus, ATCos were not working in parallel.

IV. Automatic Speech Recognition and Understanding

The exercise implements the ASRU system as defined by the project HAAWAII. The ASRU core mainly relies on four modules, performing voice activity detection (VAD), speech-to-text transformation (S2T), prediction of relevant context (Callsign Prediction), and extraction of semantic meaning (Concept Recognition). Figure 5 gives an overview about the integration of the ASRU components (light blue) in the context of the project.

Voice Activity Detection (VAD): The process is relatively straightforward, as the push-to-talk signal is readily available. However, there may be instances, where the last one or two words of the simulation pilots' readback are overlapping with the next utterance from the ATCo. Despite this, it does not have a significant impact on the performance of *Concept Recognition.*

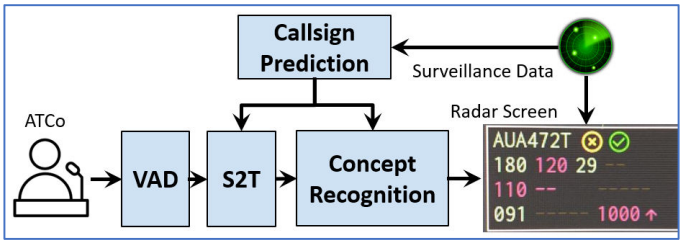

Figure 5. ASRU components in validation setup.

Speech-to-text (S2T): Whenever the VAD detects a transmission, the signal is forwarded to S2T and the recognition process starts in real time. The idea is to transform only the portion of the signal that contains audio, into word sequences. This means the S2T delivers intermediate recognitions as soon as an ATCo starts speaking and updates the recognized words continuously until the end of the transmission (end point). The S2T utilizes a combination of cutting-edge technologies to achieve optimal performance. The architecture is a hybrid deep neural network hidden Markov model (DNN-HMM). It is based on an HMM model combined with a convolutional neural network factorized time delayed neural network (CNN-TDNNF). The entire system is trained using the lattice-free maximum mutual information objective function. The implementation follows the widely-known Kaldi toolkit, which uses Mel frequency cepstral coefficients (MFCC) and i-vector as input features. It also incorporates techniques such as 3-fold speed perturbation and one third frame sub-sampling. Additionally, a 3-gram language model (LM) was trained and adapted using in-domain data to further enhance the accuracy of the system. Table I shows the performance on word level, which is based on the **word error rate** (WER), i.e., the percentage of words not correctly recognized, which is based on the Levenshtein distance [39]. It also lists the number of uttered words, the number of substitutions (Subst), deletions (Del), and insertions (Ins) representing the difference between recognized words and the actual uttered words. The best performance, i.e., lowest WER, for a single ATCo on all his/her four runs was 0.7%, the worst one was 8.2%.

TABLE I. Word Error Rate, i.e., Performance at Word Level

	# Words	Levenshtein Distance	# Subst	# Del	# Ins	WER
Total	118816	3712	1853	1324	535	3.12%
Heavy	64441	2148	1066	729	353	3.33%
Medium	54375	1564	787	595	182	2.88%
Solution	59180	1805	881	686	238	3.05%
Baseline	59636	1907	972	638	297	3.20%

Manual, i.e. gold, transcriptions were generated for all ATCo utterances including baseline runs.

Callsign Prediction: This module considers surveillance data to determine if any recognized callsign could reasonably be part of an ATCo voice transmission. The output is used by S2T and Concept Recognition to enhance the recognition quality of both.

Concept Recognition: Every time a word sequence is forwarded, it is analyzed by the concept recognition module. The analysis result is then transformed into relevant ATC commands as defined by SESAR project PJ.16-04 CWP HMI [40] and extended by the HAAWAII project [41]. The Concept Recognition provides early recognitions of callsigns (Csgn) and also of commands (Cmd), i.e., they are already in the intermediate output even before the utterance completely ended. The implementation relies on a rule-based algorithm, which determines the relevant parts in a step-by-step manner [41]. Plausibility values are assigned to each extracted command. If the plausibility is too low (< 51%), it is not forwarded to the radar screen.

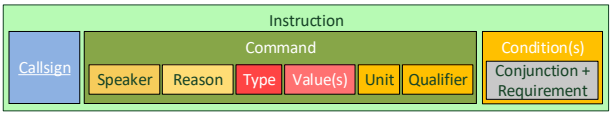

Figure 6. Elements of an instruction consisting of callsign, command etc.

A command at semantic level consists of callsign, speaker (ATCo or pilot), reason (report, request, readback, command), see Figure 6, a type, values, unit, qualifier, and can have optional conditions. Therefore, all these parts must be correctly extracted at semantic level to be counted as a correct recognition. Otherwise it is counted as an error or a rejection. We concentrated on ATCo utterances in this study. Therefore, speaker and reason are not used. Table II summarizes the performance of the Concept Recognition block of Figure 5.

TABLE II. PERFORMANCE AT COMMAND / SEMANTIC LEVEL

	WER	Cmd-Recog-Rate	Cmd-Error-Rate	Csgn-Recog-Rate	Csgn-Error-Rate
Full Command	3.1%	92.1%	2.8%	97.8%	0.6%
Only Label		92.5%	2.4%	97.8%	0.6%
Label Without Csgn		94.3%	1.9%	---	---
Only Label, offline		93.4%	1.7%	97.9%	0.5%
Only Label, gold	0.0%	99.3%	0.3%	99.9%	0.1%

The columns "Cmd-Recog-Rate" and "Cmd-Error-Rate" show the percentage of correct command recognitions and errors, respectively. The difference of the sum of these two columns to 100% correspond to rejected commands. The last two columns show the same metrics for the callsign only. Details with respect to the used metrics can be found in [42].

The row "Full Command" visualizes the quality on all instruction elements, even if they are never shown in the radar label of this application. As for our application only callsign, type, and value are important, the row "Only Label" shows the rates, when we ignore unit, qualifier etc. If we also ignore the callsign ("Label Without Csgn"), the command recognition rates increase to 94.3%, i.e., 2.2% of the wrong or missing label entries result from wrong or unrecognized callsign. After the exercise the rates were recalculated again offline by also eliminating some obvious software bugs. The recalculated rates of row "Only Labels" on the same word sequence inputs are shown in row "Only Label, offline". All these reported rates of those four rows receive the same word sequences with an average WER of 3.1% as input. If we would assume a perfect S2T block, i.e., a word error rate of 0%, we measure a command recognition rate of 99.3%, which shows that the used phraseology is very well modelled for the Concept Recognition module.

V. OBJECTIVE VALIDATION RESULTS

The following table III shows the result of the Stroop test already described in Figure 3 for medium traffic scenarios. Columns "ATCo-Id" show the identifier of the participant. "Sol" and "Base" show the Stroop test results (the higher the number of correct tests, the better). The number "1/2" indicates whether the participant started with a baseline or solution run ("1") and ended with a baseline or solution run ("2").

TABLE III. NUMBER OF SUCCESSFUL STROOP TESTS; RESULTS FOR MEDIUM TRAFFIC SCENARIO

ATCo Id	Sol 1	Bas 2	*ATCo-Id*	Bas 1	Sol 2
1	66	31	***3***	30	34
2	106	66	***5***	17	65
4	28	20	***7***	28	50
6	10	39	***9***	3	42
8	30	53	***11***	18	51
10	44	78			
12	34	41			
Average	45.4	46.9	***Average***	19.2	48.4
Average Run 1		34.5	***Average Run 2***		47.5

A. Compensating Sequence Effects

Due to sequence effects, the results in the second run were mostly better than in the first run of the ATCo and from second to third they also slightly improved etc. This averages out, when 50% started with baseline and 50% with solution run, but being able to compensate will increase statistical significance. Furthermore, a run with ATCo-Id 2 was repeated, so that seven ATCos started the medium scenario with solution runs and only five with baseline runs. The following approach adapted from [5], was used to compensate for these sequence effects:

TABLE IV. NUMBER OF SUCCESSFUL STROOP TEST RESULTS FOR MEDIUM TRAFFIC SCENARIO AFTER CLEANING SEQUENCE EFFECTS

ATCo-Id	Sol1	Bas2	Diff	ATCo-Id	Bas1	Sol2	Diff
1	72.5	24.5	48.0	***3***	36.5	27.5	-9
2	112.5	59.5	53.0	***5***	23.5	58.5	35
4	34.5	13.5	21.0	***7***	34.5	43.5	9
6	16.5	32.5	-16.0	***9***	9.5	35.5	26
8	36.5	46.5	-10.0	***11***	24.5	44.5	20
10	50.5	71.5	-21.0				
12	40.5	34.5	6.0				
Average	51.9	40.4	11.6	***Average***	25.7	41.9	16.2
Average Run 1		41.0		***Average Run 2***		41.0	

The average values of all 12 ATCos for the first run and the second run were calculated, see second last and last rows of table III. The averages of the last row were used to correct the Stroop test results. During the first runs the ATCo all together achieved an average number of 34.5 successful Stroop tests (average of all values in column "Sol 1" and column "Base 1", marked in light yellow). In the second run with medium traffic the ATCos successfully completed on average 47.5 Stroop tests (columns "Base 2" and "Sol 2", marked in light green). Therefore, we corrected all entries of table III by 50% of the difference between

34.5 and 47.5. As shown in table IV, first runs were corrected by adding 6.5 and second runs were corrected by subtracting 6.5.

B. Paired t-Tests to Evaluate Statistical Significance

The differences between runs of the same ATCo in baseline and in solution runs get smaller now, i.e., sigma decreases. This is also shown by the performed paired t-test. The null hypothesis H_0 is "No ASRU support increases the number of successfully performed Stoop tests compared to solution mode". The test value is defined by:

$$T = (D - \mu_0)\,\frac{\sqrt{n}}{SD}, \qquad (1)$$

The differences of the successful Stroop tests (solution minus baseline runs) for each run of table IV is calculated, e.g., 72.5 minus 24.5 for ATCo-Id 1. The number of differences (ATCos) is n (12 in our case). D is the mean value of the performed Stroop test differences "Diff", i.e. 13.5. SD is the standard deviation of the differences, i.e., 23.6. We are just interested in checking, whether ASRU input enables more correct Stroop tests than mouse input. Therefore, μ_0 is set to 0. The value T of 1.98 is calculated.

As T obeys a t-distribution with n-1 degrees of freedom we can reject our null hypothesis H_0 with probability of α (p-value), if the calculated value for T is bigger than the value of the inverse t-distribution at position $t_{n-1,1-\alpha}$ with n-1 degrees of freedom (in our case 1.80 for α=0.05). Therefore, the hypothesis H_0 is rejected, because T=1.98 > 1.80 holds. We could even calculate the minimal α so that $T > t_{n-1,\,1-\alpha}$ is still valid. This is in our case α=0.036. Our results falsify the negatively formulated null hypothesis. Hence, more correct Stroop tests were done, when ATCos had ASRU support.

The probability to reject the null hypotheses for the heavy traffic scenario and for both scenarios together is calculated, too. Table V shows the minimal α values, so that $T > t_{n-1,\,1-\alpha}$ holds. We will mark in the following in green, if α is less than 5%. With considering sequence effects all null-hypotheses can be rejected with α less than 5%.

TABLE V. MIN ALPHA FOR FOR SUCCESSFUL STROOP TESTS

Hypotheses	Medium	Heavy	Both
With considering sequence effects	3.6%	2.3%	0.3%
Without considering sequence effects	9.3%	3.6%	1.3%

Minimal α values, shaded in green for 0%≤α<5%, in light green for 5%≤α<10%

Our approach to compensate the sequence effects by subtracting or adding the average values and then performing a paired t-test on the transformed values is not the only approach. The two-way analysis of variance (ANOVA) might provide similar results [43], but we did not further investigate. More details to increase trust in our results can be found in the appendix.

C. Workload Resulting from Radar Label Maintenance

We wanted to know how much time the ATCo could save for maintaining the command values in the radar label when being supported by ASRU. If the ATCo gives the command "AUA472T DESCEND 120 FL, AUA472T HEADING 110, AUA472T RATE_OF_DESCENT 1000 ft_min OR_GREATER", the ATCo has to click into the radar label (see Figure 1), i.e., into the active altitude field, scroll in the menu, and select the value 120, then click on the active heading field and scroll down or up to the heading value 110, click in the vertical rate field, select the 1000, and click into the field for "OR_GREATER". The ATCo can update the cell at any time, e.g., 5 seconds before the command is spoken or 10 seconds later or during readback of the pilot. In solution runs, the ATCo just checks the three purple values and the purple arrow indication "OR GREATER" in the radar label of the callsign, see Figure 7.

Figure 7. Radar label showing recognized flight level (120), heading (110 and "—" for waypoint (WP) cell), and descent rate (1000 or greater).

If the values are correct, nothing needs to be done. If one value is wrong or missing, it needs to be manually overwritten. If more values are wrong, the ATCo's workload would of course be higher compared to entering the values manually from the beginning. The worst case would be, if even the callsign is wrongly recognized. Then, the ATCo needs to reject the values for the wrong callsign and to manually input all of them for the correct callsign. As the callsign recognition error rate is very low, see Table II in section IV, this last case almost never happens.

Table VI shows the time (column 2) needed for entering the values into the radar label cells. Both baseline and solution scenarios for 12 ATCos each sum up to 50,400 seconds duration. In the baseline scenarios the ATCos spent 25.3% of their time, i.e., 12,763 of 50,400 seconds, just for clicking, whereas in the solution runs only 0.8% of the time was needed, i.e., a factor of more than 31. It needs to be mentioned that ATCos are used to doing some tasks in parallel. During those 12,763 seconds they are, e.g., partly speaking to the pilot or checking pilots` readbacks. The ASRU system, however, is not perfect. The ATCo is still responsible for the contents in the radar labels cells. Especially during solution runs, the ATCos need to check contents of the cells and compare it to the pilot readback. This requires mental workload, which also needs to be considered.

TABLE VI. RADAR LABEL MAINTENANCE EFFORT: DURATION AND CLICKS FOR CERTAIN INPUT TYPES

		Number of Mouse Clicks for Radar Label Input						
Scenario	**Time [s]**	**Total**	**Altitude**	**Head**	**Speed**	**Trans**	**Rates**	**WP**
Baseline	12,763	58,77	1,906	*572*	1,074	216	74	589
Solution	405	154	28	*7*	34	7	11	20

The call values do not sum up to "Total", because the numbers of for elements of Figure 1 are not shown. Especially the field with ILS and Handover contains 936 clicks.

As done in [17], we used the Keystroke-Level-Model (KLM) [44] for the above analysis. This model defines execution times for different types of human-computer interaction, e.g., press or release a button, move the mouse to a specific position on the screen, the mental process of thinking what to do next. Since the calculation of input commands via mouse starts with the first click in the respective label, we ignore the time the ATCo needs

to move the mouse to the label. We estimate the additional time compared to the baseline ("mouse only") scenario with 1200 milliseconds for every command that was accepted without any correction. This time correlates with the duration needed for a single mental process thinking of what to do next. In all solution runs 6,480 commands were given, which were relevant for the radar label cells, i.e., 6,480 * 1.2 seconds need to be added to the 405 seconds for correcting the wrong and missing recognitions. This sum of 8,200 seconds is still less than the 12,763 seconds needed for clicking in the baseline runs.

D. Missing Information in Radar Label Cells

Knowing now that ASRU support reduces ATCo workload, we need to know if all given commands show the current situation in digital form, i.e., are the contents of the radar label cells correct, after the ATCo has corrected the ASRU output. How often, do we have wrong or missing inputs? A person who in theory counts how often the radar label contents are different from the spoken commands is needed. However, this approach is hopeless. Nobody is able to listen to ATCo utterances, understand on word and on semantic level, and check the radar label contents with the needed accuracy. A deviation in the order of 1% is expected. From transcription experiments it is already known that a person's word error rate is in the order of 4% to 11%, especially when a person can only listen once [45]. A computer-based solution is necessary.

During the experiments all mouse clicks changing the radar label cell contents are recorded; table VI resulted from these recordings. The correct contents of each cell for each callsign at any point in time is indirectly given: All voice recordings were transcribed and annotated. These so-called *gold annotations* are replayed and sent to the process, which has generated the contents of the radar label cells, i.e., the clicks are recorded again, i.e., we got the *gold contents* (assumed to be correct) of each cell for each callsign at any point in time. Then the cell contents during the experiments are compared with the correct/gold contents. The comparison of the label cell contents during the experiments to the correct contents can be done automatically, and best of all, the calculation can be redone whenever we find, e.g., an inconsistency in the gold annotations.

Table VII shows the result for the baseline and the solution runs. The first column shows the number of clearances given for each cell. We did not count commands which cleared a value in a field, e.g., "own navigation" or "no speed restrictions", but we considered them when a calculation was missing or in case of wrong cell entries. "Gold" contains the number of commands of this type, resulting from the replay of the manual annotations. "Clicks" counts the number of clicks into this cell, which changed the value in this cell, ignoring clicks clearing the value.

"Miss" counts the number of cell values, which were missing and "Add" the number of cell values which were in the cells, but not said at that time. If the 250 knots were intended/said and the value 240 was accidently entered or wrongly recognized and, therefore, not correct, we counted this twice as missing 250 and as additional 240 in the "Spd" row. "RR" is the command recognition rate for that type. The entries in the cells "Miss" and "Add" are corrected by sequence effects as described in subsection V.A. The compensation effects, however, were much smaller than for the Stroop test. The biggest change is by 1.3 in absolute numbers. Some cells of Table VII are marked in orange, which require a deeper analysis or some more explanations.

TABLE VII. NUMBER OF ERRORS IN RADAR LABEL CELLS AFTER CLEANING BY SEQUENCE EFFECTS FOR HEAVY AND MEDIUM SCENARIOS

	Baseline					Solution				
Type	**Gold**	**Clicks**	**Miss**	**Add**	**RR**	**Gold**	**Clicks**	**Miss**	**Add**	**RR**
Alti	1,950	1,906	62	20	95%	1,978	28	19	16	95%
Spd	1,102	1074	70	35	89%	1,183	34	17	3	89%
Head	936	572	351	8	94%	894	7	30	11	94%
WP	598	589	29	14	85%	604	20	18	25	87%
Tran	301	216	89	12	85%	289	7	23	1	88%
Rate	63	74	13	4	67%	64	11	6	1	74%
Spec	1,367	936	14	15	93%	1,372	19	34	15	92%

a. "Alti" shows the number of commands, which were spoken and would require an input into the altitude cell in the radar label. "Spd" for speed cell, "Head" for the heading cell, "WP" for the waypoint cell, "Tran" for the "Transition/Route", "Rate" for the descent rate and "Spec" for the ILS/approach clearance and the change frequency command type.

- 62 altitude commands were not or wrongly entered into the radar cells during baseline runs. Reasons are that 39 times the clearance "*descend three thousand feet, cleared ILS runway three four*" was only entered as ILS clearance. 12 times a clearance to 2600 feet was given. The input of an altitude clearance, which is not a multiple of 1000 was not supported by the used HMI. Nevertheless, these values were in the altitude cells with ASRU support.
- 351 heading values were not entered into the label cells in baseline runs. This happens 218 times, when the ATCo gives an ILS clearance ("heading three one zero cleared ILS three four"). ATCos from ACG explained that they are never inputting heading values in their TopSky system at home, when giving an ILS clearance. Nevertheless, this is an advantage for ATCos, when supported by ASRU.
- 19 altitude values were not corrected in the solution runs. This mostly happened, when recognizing a wrong callsign.
- 25 misrecognitions of waypoint values were not corrected in the solution runs. 15 times a waypoint name was wrongly extracted instead of another command type (e.g., "*reduce two two zero two*" extracted as abbreviation of waypoint *WW202*). Six out of these 15 cases happened together with the handover to tower. A reason for forgetting the correction could be that the aircraft are not in the focus of the ATCo anymore. Furthermore, seven wrong waypoint values appeared, when the ATCos advised a holding at a waypoint or directed the aircraft to the starting waypoint of a transition, which was misrecognized as DIRECT_TO. From the point of evaluation this is a wrong, not corrected entry, but operationally it does not really matter, in which cell the waypoint is shown. Twice a wrong callsign was extracted, which had already been handed over to tower. The ATCos did not pay attention anymore to those cell entries.
- The 34 values in the special field in the fourth label line mostly (30 times) relate to a misrecognition of a CLEARED ILS RW34 command, which was recognized as EXPECT ILS RW34. The ATCo did not realize this misrecognition, because it is shown only when the mouse is hovered over the label. The hidden forth label line, shown in left part of

Figure 8, was implemented by intention to find out, if this has effects on situation awareness and safety. It would have and, therefore, the fourth label line must always be shown when it changes and not only when the mouse hovers over.

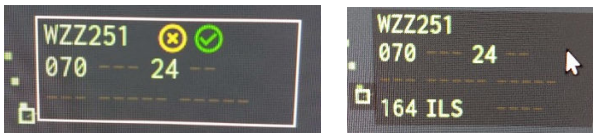

Figure 8. Radar label with hidden (left) and shown (right) fourth label line.

The initial question was how it can be verified how many ATCo commands are missing in the label cells with and without ASRU support. This was done by replaying the annotated utterances. The next question was if all missing commands were corrected by the ATCo. The numbers in Table VII clearly show that this is not the case. However, it is also not the case when the ATCo manually inputs all commands. We performed paired t-tests as described in subsection V.A to validate if the differences are statistically significant. We used the data after cleaning the sequence effects.

TABLE VIII. MIN ALPHA FOR HYPOTHESES THAT ASRU IMPROVES CORRECTNESS OF RADAR LABEL CELL CONTENTS

Hypotheses	Medium	Heavy	Both
Total Missing	1.6E-04	4.9E-05	6.1E-08
Altitude cell	1.7E-04	7.2%	2.1E-04
Speed cell	1.9%	0.5%	4.9E-04
Heading cell	3.4E-04	3.0E-04	1.1E-06
Waypoint cell	16.9%	9.8%	5.2%
Transition cell	2.8%	5.0E-04	1.0E-04
Vertical Rate cell	-18.6%	2.9%	12.7%
Special cell	-14.8%	-9.5%	-4.8%

Minimal α values, shaded in green for 0%≤α<5%, in light green for 5%≤α<10%, in orange if we have evidence that results are worse with ASRU support, and in yellow for the rest (|α| ≥ 10%).

Table VIII shows the minimal α values. We have (very strong) statistical significance $\alpha < 10^{-7}$ % (column *Both*, row *Total Missing*), that the contents of cell values are better, when the ATCo is supported by ASRU for the majority of command types. We already discussed that, when explaining Table VII.

VI. RESULTS FROM SUBJECTIVE ATCO FEEDBACK

The ISA test aims at a retrospective self-assessment of the workload during the last 5 minutes. Every five minutes a five-points rating scale was shown on the radar screen. The ATCo has to select one of five numbers (1) Under-utilised, (2) Relaxed, (3) Comfortable, (4) High, (5) Excessive. In total, we got 6 to 7 values from the 48 simulation runs resulting in 327 feedbacks.

TABLE IX. MIN ALPHA AND DIFFERENCE OF AVERAGE ISA VALUES

		Medium	Heavy	Both
delta ISA	no consideration of sequence effects	-0.03	-0.39	-0.21
min α		42.6%	1.1%	3.1%
delta ISA	consideration of sequence effects	-0.09	-0.39	-0.25
min α		10.6%	0.5%	0.3%
Average ISA	Without ASRU	2.48	3.26	2.87
	With ASRU	2.39	2.87	2.63

Minimal α values, marked green (0%≤α<5%), light green (5%≤α<10%), and yellow (α ≥ 10%).

Table IX shows the results of the paired t-test. Negative values in column "*delta ISA*" indicate that the average ISA score was 0.39 units less for the heavy traffic scenario, i.e., better in the solution runs. The last two rows show the absolute ISA scores without sequence effects. The consideration of the sequence effects is important for the medium traffic scenarios. The compensation of sequence effects does not influence the average value of the ISA scores, but reduces the sigma and therefore improves statistical significance, e.g. for the heavy traffic scenario from α=1.1% down to 0.5%.

Six questions are used for the NASA-TLX with the ten answer possibilities (1) *Low* to (10) *High*. The Bedford Workload scale has two different questions with also ten alternatives from (1) *Workload Insignificant,* (2) *Workload Very Low,* … (9) *Workload Very High,* (10) *Workload Extremely High*. SASHA offers six questions with answers from (1) *Never*, (2) *Seldom*, … (6) *Very Often*, (7) *Always*. AIM-s has 15 questions with seven answer alternatives: (1) *None*, (2) *Little* … to (7) *Extreme*. The System Usability Scale (SUS) uses ten questions with five alternatives from (0) *fully disagree* … to (4) *fully agree*. CARS offers ten alternatives for one question: (1) *Improvement mandatory. Safe operation could not be maintained* … (5) *Very Objectionable Deficiencies. Maintaining adequate performance requires extensive controller compensation* to (10) *Deficiencies are rare. System is acceptable and controller doesn't have to compensate to achieve desired performance*.

We show the answers to the following 10 safety related questions; in brackets the row name in table X and the test category:

1. How insecure, discouraged, irritated, stressed, and annoyed were you? (Stress annoyed, NASA-TLX)
2. What was your peak workload? (Peak workload, Bedford)
3. In the previous run I … started to focus on a single problem or a specific aircraft. (Single aircraft, SASHA)
4. In the previous run there … was a risk of forgetting something important (such as inputting the spoken command values into the labels). (Risk to Forget, SASHA)
5. In the previous run, how much effort did it take to evaluate conflict resolution options against the traffic situation and conditions? (Conflict resolution, AIM-s)
6. In the previous run, how much effort did it take to evaluate the consequences of a plan? (Consequences, AIM-s)
7. In the previous working period, I felt that … the system was reliable. (Reliable, SATI)
8. In the previous working period, I felt that … I was confident when working with the system. (Confidence, SATI)
9. I … found the system unnecessarily complex. (Complexity, SUS)
10. Please read the descriptors and score your overall level of user acceptance experienced during the run. Please check the appropriate number. (User Acceptance, CARS)

The results of each question with elimination of sequence effects are shown in table X. After elimination of sequence effects, all values are normalized to the interval [1..10], so that value 1 corresponds to safe and 10 to unsafe. With this normalization, we could calculate a total safety value by calculating the average of all 10 questions. We see the results for Medium, Heavy, and Both traffic scenarios. "Diff" shows the average differences of solution run minus baseline run. A negative value means that the

ATCo thinks the solution is safer. Column "α" shows the p/α-value of the performed paired t-tests. We are now interpreting the green and orange α-values:

TABLE X. MIN ALPHA OF PAIRED T-TESTS AND DIFFERENCE OF AVERAGE FOR SAFETY RELEVANT SUBJECTIVE FEEDACK OF ATCOS

	Medium		Heavy		Both	
Question	***Diff***	***α***	***Diff***	***α***	***Diff***	***α***
Stress, annoyed	-0.50	18%	0.18	-38%	-0.16	34%
Peak workload	-0.44	4.6%	-0.19	33%	-0.32	9.9%
Single aircraft	0.24	-23%	-0.13	32%	0.04	-41%
Risk to forget	-0.83	0.2%	-0.46	14%	-0.64	0.7%
Conflict resolution	0.22	-35%	-0.74	5.6%	-0.26	24%
Consequences	0.77	-10%	-0.17	34%	0.30	-21%
Reliable	0.19	-40%	-0.63	13%	-0.24	30%
Confidence	-1.75	3.7%	-1.44	7.8%	-1.59	1.1%
Complexity	-2.36	0.4%	-1.60	1.5%	-1.98	2.E-04
User Acceptance	-1.72	3.7%	-0.30	37%	-1.01	6.3%
Total	-0.49	4.8%	-0.62	2.2%	-0.56	0.4%

Minimal α values, shaded in green for 0%≤α<5%, in light green for 5%≤α<10%, in orange if we have evidence that results are worse with ASRU support, and in yellow for the rest (|α| ≥ 10%).

- The feedback to "Stress, annoyed", "Single aircraft", "Conflict resolution", "Consequences", and "Reliability" show trends, but no statistical significance.
- Peak workload seems to be less with ASRU support especially in the medium traffic scenarios. The "Risk to forget" something is significantly higher (α=0.7%, both), when no ASRU support is available.
- The feedback to "Confidence" and "Complexity" of use was better in solution runs in all three cases, i.e., *Medium*, *Heavy*, and *Both*. The α-value of 0.0002 for system complexity is very clear, i.e., statistically significant. The same applies for the absolute difference of 1.98 for both on the 10-digit scale and even 2.36 for the medium scenario.
- The "User Acceptance" with respect to safety is clear with preference to ASRU support, which is not statistically significant for the heavy scenarios, which requires additional work and further analysis.
- In "Total" feedback to the safety related topics of all the questionnaires strongly prefers ASRU support with a statistical significance of 0.4%.

VII. CONCLUSIONS

The paper compared ATCos' workload and safety effect between solution runs with automatic speech recognition and understanding (ASRU) support and baseline runs without ASRU support. The evaluated application was inputting spoken commands into the aircraft radar labels on the radar screen. Two main variables affected the results: (1) usage of ASRU or not and (2), sequence effects, i.e., whether ATCos started with ASRU support or not. An approach was described and successfully applied to compensate these sequence effects. This increased statistical significance of the results; in many cases we got the same results compared to doubling the number of study participants.

With ASRU support, the time ATCos need to insert given commands into the radar labels, could be reduced from more than 12,000 to 400 seconds. 4% of label values were wrong or missing, when ASRU support was used even after manual correction with mouse and keyboard. This could have effects on safety, because also altitude values were missing or wrong. It was, however, also evaluated that without ATCo support 11% of given commands were wrongly or not inputted into the radar labels, i.e., it was shown that label error with ASRU support is smaller than without ASRU support. Evaluation of subjective ATCo feedback from questionnaires provided the same results with high statistical significance (α=0.3%). A **new replay technique** was implemented, which **enables a very accurate calculation of missing and wrong label entries**, based on manual transcription and annotation of all spoken ATCo transmissions to pilots.

The ATCos that participated in the trials recognized the **significant potential of ASRU to increase** their available **mental capacity** for effective traffic management. This increase in mental capacity is particularly relevant in high-density traffic and/or when working in a control sector, which inherently requires a steady flow of instructions to aircraft to improve final approach path spacing. The more intuitive and well-implemented the integration of ASRU is in the radar HMI, the greater the potential benefits will be. This can be achieved by providing easy methods to check and, if necessary, fix any incorrect or missing ASRU output. Additionally, incorporating a mechanism verifying the contents of radar labels against radar data and Mode-S downlinked target values can further increase safety and is already available in existing assistant systems.

It is important to note that during the trial, **no major safety concerns** were reported by the participating ATCos. This is a crucial requirement for the real-world implementation of ASRU systems. In fact, it is possible that the use of ASRU may lead to a reduction of risk caused by errors in inputting data in the radar labels. The main conclusion of the study is that **Artificial Intelligence-based applications are ready for integration in the operations room --** at least for ASRU. While they are not yet perfect and achieving recognition rates of 99.9% is still a goal to strive for, it is also true that human operators make wrong undetected inputs. By **considering ASRU and ATCos as a team**, the safety of air traffic management can be enhanced and workload can be reduced.

ACKNOWLEDGMENT

The authors want to thank all the ATCos from Austro Control who supported the validations with their feedback and their voice utterances. Also, many thanks to the employees of DLR doing the tedious transcription and annotation work.

REFERENCES

[1] S. Ciupka, "Siris big sister captures DFS" original German title: "Siris große Schwester erobert die DFS," transmission, Vol. 1, 2012.

[2] M. Kleinert, H. Helmke, G. Siol, H. Ehr, D. Klakow, M. Singh, P. Motlicek, Chr. Kern, A. Cerna, and P. Hlousek, "Adaptation of Assistant Based Speech Recognition to New Domains and its Acceptance by Air Traffic Controllers," in Proc. of the 2nd International Conference on Intelligent Human Systems Integration (IHSI 2019): Integrating People and Intelligent Systems, 2019, San Diego, CA, USA.

[3] ICAO, “Doc 4444, Procedures for Air Navigation Services, Air Traffic Management,” ICAO, Montréal, Canada, 2016.

[4] M. Kleinert, S. Shetty, H. Helmke, O. Ohneiser, H. Wiese, M. Maier, S. Schacht, and I. Nigmatulina, “Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System,” 12th SESAR Innovation Days, Budapest, Hungary, 2022.

[5] H. Helmke, O. Ohneiser, J. Buxbaum, and Chr. Kern, “Increasing ATM efficiency with assistant-based speech recognition”, 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 2017.

[6] R. Tarakan, K. Baldwin, and R. Rozen, ”An Automated Simulation Pilot Capability to Support Advanced Air Traffic Controller Training,” in 26th Congress of the International Council of the Aeronautical Sciences, Anchorage, AK, USA, 2008.

[7] J.M. Cordero, M. Dorado, and J.M. de Pablo, “Automated speech recognition in ATC environment,” in 2nd International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS‘12), IRIT Press, Toulouse, France, 2012, pp. 46–53.

[8] J.M. Cordero, N. Rodríguez, J.M. de Pablo, and M. Dorado, “Automated Speech Recognition in Controller Communications Applied to Workload Measurement,” 3rd SESAR Innovation Days, Stockholm, Sweden, 2013.

[9] 2018 AIRBUS Air Traffic Control Challenge Workshop: https://www.irit.fr/recherches/ SAMOVA/pagechallenge-airbus-atc-workshop.html

[10] S.R. Young, W.H. Ward, and A.G. Hauptmann, “Layering predictions: Flexible use of dialog expectation in speech recognition,” in Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI89), Morgan Kaufmann, 1989, pp. 1543–1549.

[11] S.R. Young, A.G. Hauptmann, W.H. Ward, E.T. Smith, and P. Werner, “High level knowledge sources in usable speech recognition systems,” in Commun. ACM, vol. 32, no. 2, Feb. 1989, pp. 183–194.

[12] D. Schäfer, “Context-sensitive speech recognition in the air traffic control simulation,” Eurocontrol EEC Note No. 02/2001 and PhD Thesis of the University of Armed Forces, Munich, Germany, 2001.

[13] Y. Oualil, M. Schulder, H. Helmke, A. Schmidt, and D. Klakow, “Real-Time Integration of Dynamic Context Information for Improving Automatic Speech Recognition,” Interspeech, Dresden, Germany, 2015.

[14] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, and M. Schulder, “Assistant-Based Speech Recognition for ATM Applications,” in 11th USA/ Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[15] T. Shore, “Knowledge-based word lattice re-scoring in a dynamic context,” Master Thesis, Saarland University (UdS), Germany, 2011.

[16] T. Shore, F. Faubel, H. Helmke, and D. Klakow, “Knowledge-Based Word Lattice Rescoring in a Dynamic Context,“ Interspeech 2012, Sep. 2012, Portland, OR, USA.

[17] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, “Reducing Controller Workload with Automatic Speech Recognition,” in IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016.

[18] M. Kleinert, H. Helmke, G. Siol, H. Ehr, M. Finke, A. Srinivasamurthy, and Y. Oualil, “Machine learning of controller command prediction models from recorded radar data and controller speech utterances,” 7th SESAR Innovation Days, Belgrade, Serbia, 2017.

[19] M. Kleinert, H. Helmke, G. Siol, H. Ehr, A. Cerna, C. Kern, D. Klakow, P. Motlicek et al., ”Semi-supervised Adaptation of Assistant Based Speech Recognition Models for different Approach Areas,” in IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018.

[20] S. Chen, H. Kopald, R. S. Chong, Y.-J. Wei, and Z. Levonian, “Readback error detection using automatic speech recognition,” in 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 2017.

[21] H. Helmke, M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr, H. Ariliusson, T. S. Simiganoschi, A. Prasad, P. Motlicek, K. Vesely, K. Ondrej, P. Smrz, J. Harfmann, and C. Windisch, “Readback error detection by automatic speech recognition to increase ATM safety,” 14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021), Virtual Conference, 2021.

[22] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang. and B. Yang, “A real-time ATC safety monitoring framework using a deep learning approach,” IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 11, 2020, pp. 4572–4581.

[23] J. Zuluaga-Gomez, P. Motlicek et al., “Automatic Speech Recognition Benchmark for Air-Traffic Communications,” Interspeech, 2297–2301. https://doi.org/10.21437/Interspeech.2020-2173.

[24] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina et al., “How does pre-trained wav2vec 2.0 perform on domain shifted ASR? An extensive benchmark on air traffic control communications,” 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023.

[25] J. Zuluaga-Gomez, K. Vesely, I. Szoke, P. Motlicek et al., “ATCO2 corpus: A largescale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications,” arXiv preprint arXiv:2211.04054.

[26] S. Chen, H. Kopald, R. Tarakan, G. Anand, and K. Meyer, “Characterizing national airspace system operations using automated voice data processing,” 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria, 2019.

[27] S. Chen, H. Kopald, B. Avjian, and M. Fronzak, ”Automatic pilot report extraction from radio communications,” 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 2022.

[28] J. Zuluaga-Gomez, S.S. Sarfjoo et al., “Bertraffic: Bert-based joint speaker role and speaker change detection for air traffic control communications,” 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023.

[29] J. Zuluaga-Gomez, K. Vesely et al., “Automatic call sign detection: Matching air surveillance data with air traffic spoken communications,” Multidisciplinary Digital Publishing Institute Proceedings, 59 (1), 14, 2020

[30] M. Kleinert, H. Helmke, S. Moos, P. Hlousek, C. Windisch, O. Ohneiser, H. Ehr, and A. Labreuil, “Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance,” 9th SESAR Innovation Days, Athens, Greece, 2019.

[31] C.S. Jordan and S.D. Brennen, “Instantaneous self-assessment of workload technique (ISA),” Defence Research Agency, Portsmouth, 1992.

[32] S.G. Hart and L.E. Staveland, “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,” in Human mental workload, P.A. Hancock and N. Meshkati, Eds., Amsterdam, North-Holland, 1988, pp. 139–183.

[33] A.H. Roscoe, “Assessing pilot workload in flight,“ Proceedings of the AGARD Conference Proceedings Flight Test Techniques, Lisbon, Portugal, 1984.

[34] J. Brooke, “SUS-A Quick and Dirty Usability Scale,” Usability Evaluation in Industry, Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester, B.A., Eds.; Taylor and Francis: London, UK, 1996, pp. 189–194.

[35] K.K. Lee, K. Kerns, R. Bone, and M. Nickelson, “The Development and Validation of the Controller Acceptance Rating Scale (CARS): Results of Empirical Research,” 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 2001.

[36] D.M. Dehn, “Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires,“ Air Traffic Control Quarterly, 16(2), 2008, pp. 127–146.

[37] J.R. Stroop, “Studies of interference in serial verbal reactions,” Journal of experimental psychology 18.6, 1935, pp. 643–662.

[38] S.M. Casner and B.F. Gore, “Measuring and evaluating workload: A primer,” NASA Technical Memorandum, 216395, 2010.

[39] V.I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in Soviet Physics -- Doklady 10.8, Feb. 1966.

[40] H. Helmke, M. Slotty, M. Poiger, D.F. Herrer, O. Ohneiser et al., “Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04,” IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018.

[41] M. Kleinert, H. Helmke, S. Shetty, O. Ohneiser, H. Ehr, A. Prasad, P. Motlicek, and J. Harfmann, "Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning," IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2021.

[42] H. Helmke, S. Shetty, M. Kleinert, O. Ohneiser, H. Ehr, A. Prasad, P. Motlicek, A. Cerna, and C. Windisch, "Measuring Speech Recognition And Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates," 11th SESAR Innovation Days, Virtual Conference, 2021.

[43] R.A. Fisher, "Statistical Methods for Research Workers," In: S. Kotz and N.L. Johnson (eds) "Breakthroughs in Statistics," Springer, New York, 1992.

[44] D. Kieras, "Using the Keystroke-Level Model to Estimate Execution Times," http://www-personal.umich.edu/~itm/688/KierasKLMTutorial 2001.pdf, 2001.

[45] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," in Proc. Interspeech, pp. 137– 141, Stockholm, Sweden, 2017.

APPENDIX

We have treated the statistical analysis as many problems with one independent variable, in our case ASRU support, and one dependent variable. In each problem we consider a different dependent variable, e.g. successfully performed Stroop- tests, missing label value entries, answers to questionnaires etc. We, however, have three independent variables, i.e. with ASRU support, traffic complexity, i.e. heavy or medium, and when the scenario was performed, i.e. solution run first or baseline first.

This could also be treated by as a multi-factor variation analysis problem without repetition, i.e. a multi-factor ANOVA already introduced by the work of Fisher [43]. We treat the problem without repetition, because we only measured once. It would have been with repetition, if the measurements would have been e.g. after ten minutes, 20 minutes and at the end. The second independent variable would have been "when taken".

For comparison our approach of elimination of sequences effects presented in subsection V.A, we use the Stroop-Test results of the heavy scenarios, shown again in table XI.

TABLE XI. NUMBER OF SUCCESSFUL STROOP-TESTS WITH SEQUENCE EFFECTS

ATCO	Heavy				Medium			
	Solution		Baseline		Solution		Baseline	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
1		18	9		66			31
2	40			0	106			66
3	33			19		34	30	
4		34	7		28			20
5	33			10		65	17	
6		23	16		10			39
7	16			34		50	28	
8		71	38		30			53
9	12			24		42	3	
10		48	31		44			78
11	14			45		51	18	
12		68	6		34			41

We can perform a two-factor ANOVA with solution/baseline and first/second run as independent variable. It enables now to test multiple null-hypotheses at the same time.

1. There is no statistically significant difference between the means of successful Stroop-tests in baseline and solution run, respectively (H1).
2. There is no statistically significant difference between the means of successful Stroop-tests between first and second run, respectively (H2).
3. The two factors do not influence each other with respect to successful Stroop-tests, i.e. we observe the same effects of ASRU support on successful Stroop-Test independent whether ASRU runs are performed first or second (H3).

Considering only the heavy traffic scenario, we get for hypothesis H1 with an f-Test a p-value of 4.8%, for H2 a p-value of 10.2% and for H3 a p-value of 28.6%, which means that the ASRU support has a big effect (4.8%), sequence of ASRU support might have an effect (10.2%) and the two variables do not interact on the result (28.6%).

We can also test with two different unpaired t-test, which, however, do not consider H3. We get a p-value of 5.7% for H1 and 12.7% for H2. Our p-value provided in table V is with 3.6% smaller, because the samples are dependent, a paired t-tests is performed, considering only one side, i.e. our null hypothesis is "*no ASRU support increases the number of successfully performed Stoop tests compared to solution mode*". The null hypothesis of the two-sided test would be "*With and without ASRU results in the same number of successfully performed Stoop tests*". Furthermore, we are not interested in the p-values of H2 and H3. We know/assume that we have the sequence effects and we are looking for a way to compensate them. The "easiest" way would be to increase the number of participants by a factor of four, so that each participant only performs one experiment. Constraints on budget for the simulator and also for participating ATCos, pseudo-pilots and validations specialists did of course not allow this.