aerospace

MDPI

Article

Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment

Oliver Ohneiser [1,*], Hartmut Helmke [1], Shruthi Shetty [1], Matthias Kleinert [1], Heiko Ehr [1], Sebastian Schier-Morgenthal [1], Saeed Sarfjoo [2], Petr Motlicek [2], Šarūnas Murauskas [3], Tomas Pagirys [3], Haris Usanovic [4], Mirta Meštrović [5] and Aneta Černá [6]

[1] German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; hartmut.helmke@dlr.de (H.H.); shruthi.shetty@dlr.de (S.S.); matthias.kleinert@dlr.de (M.K.); heiko.ehr@dlr.de (H.E.); sebastian.schier@dlr.de (S.S.-M.)

[2] Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland; saeed.sarfjoo@gmail.com (S.S.); petr.motlicek@idiap.ch (P.M.)

[3] AB "Oro Navigacija" (ON), Air Navigation Service Provider of Lithuania, Balio Karvelio St. 25, 02184 Vilnius, Lithuania; murauskas.s@ans.lt (Š.M.)

[4] Austro Control (ACG), Österreichische Gesellschaft für Zivilluftfahrt mbH, Air Navigation Service Provider of Austria, Schnirchgasse 17, 1030 Vienna, Austria

[5] Croatia Control (CroControl), Air Navigation Service Provider of Croatia, Rudolfa Fizira 2, 10410 Velika Gorica, Croatia

[6] Air Navigation Services of the Czech Republic (ANS CR), Navigační 787, 25261 Jeneč u Prahy, Czech Republic

[*] Correspondence: oliver.ohneiser@dlr.de; Tel.: +49-531-295-2566

Abstract: Assistant Based Speech Recognition (ABSR) systems for air traffic control radiotelephony communication have shown their potential to reduce air traffic controllers' (ATCos) workload. Related research activities mainly focused on utterances for approach and en-route traffic. This is one of the first investigations of how ABSR could support ATCos in a tower environment. Ten ATCos from Lithuania and Austria participated in a human-in-the-loop simulation to validate ABSR support within a prototypic multiple remote tower controller working position. The ABSR supports ATCos by (1) highlighting recognized callsigns, (2) inputting recognized commands from ATCo utterances in electronic flight strips, (3) offering correction of ABSR output, (4) automatically accepting ABSR output, and (5) feeding the digital air traffic control system. This paper assesses human factors such as workload, situation awareness, and usability when ATCos are supported by ABSR. Those assessments result from a system with a relevant command recognition rate of 82.9% and a callsign recognition rate of 94.2%. Workload reductions and usability improvement with p-values below 0.25 are obtained for the case when the ABSR system is compared to the baseline situation without ABSR support. This motivates the technology to be brought to a higher technology readiness level, which is also confirmed by subjective feedback from questionnaires and objective measurement of workload reduction based on a performed secondary task.

Keywords: air traffic controller; multiple remote tower; assistant-based speech recognition; automatic speech recognition and understanding; electronic flight strips

check for updates

Citation: Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Schier-Morgenthal, S.; Sarfjoo, S.; Motlicek, P.; Murauskas, Š.; Pagirys, T.; et al. Assistant Based Speech Recognition Support for Air Traffic Controllers in a Multiple Remote Tower Environment. *Aerospace* **2023**, *10*, 560. https://doi.org/10.3390/aerospace10060560

Academic Editor: Judith Rosenow

Received: 5 April 2023
Revised: 27 May 2023
Accepted: 7 June 2023
Published: 14 June 2023

1. Introduction

Speech recognition and speech understanding have found their way into use in daily life. While speech recognition has become quite robust with growing amounts of data, speech understanding remains a challenge given the complexity of verbal utterances' semantics. However, high accuracy in speech understanding is needed for human operators that supervise safety-critical processes, such as in aviation. Only then, users of speech recognition and understanding systems such as controllers will accept them and can benefit from their support, e.g., through workload reduction. Nowadays, tower controllers are burdened with manually maintaining flight strips, even if the content that needs to be entered in such flight strips is also communicated verbally in air traffic control radio telephony. This article presents one of the first prototypes of a speech recognition and understanding system to support ATCos in the tower environment in maintaining digital flight strips—in our case, even in a simulated multiple remote tower environment.

Our conducted validation study with ten air traffic controllers (1) quantifies any productivity enhancements in terms of mental workload, situation awareness, satisfaction, acceptance, trust, and usability through the advanced support functionalities in the digital system with automatic flight strip maintenance and highlighting features (independent variable); (2) quantifies the quality of speech-to-text and text-to-concept functionality; and (3) gathers feedback on the prototypes' functionality and visualization.

1.1. Related Work

1.1.1. Automatic Speech Recognition and Understanding in Air Traffic Management

During the last decades, a row of prototypes for speech recognition and understanding [1] in the air traffic management (ATM) domain has been developed. Early prototypes intended to support air traffic control (ATC) training and to reduce the number of required simulation pilots [2,3]. ATC events have been recognized from utterances to estimate controller workload [4,5]. The integration of contextual knowledge from an electronic assistant system for the speech recognition and understanding process [6] reduced recognition error rates [7]. These so-called assistant-based speech recognition (ABSR) systems initially focused on the approach environment [8]. For interoperability and comparability, rules for transcription (speech-to-text) and annotation (text-to-concepts)—so-called ontologies—have been defined and agreed upon between the major European ATM stakeholders [9]. Due to these rules, ATC utterances always comprise a callsign and at least one command that can consist of a type, unit, qualifier, and conditions. Later, ABSR systems were enhanced and enrolled on the en-route [10], apron [11,12], and tower environment [13]. This included the prediction and extraction of ATC commands [14]. Further research prototypes enhanced the ontologies, worked on speech recordings and radar data from real operations rooms, especially, but not limited to, recognizing callsigns [15–17], pre-filled aircraft radar labels that reduced the workload of ATCos [18,19], and implemented automatic readback error detection [10,20]. However, there was no validation of a sophisticated ABSR system's support for tower controllers, especially in a multiple remote tower setup using such a system in a high-fidelity laboratory environment.

1.1.2. Multiple Remote Air Traffic Control Tower and Human Operator Performance

The history of laboratory remote tower working positions started over two decades ago [21]. Recent research focused on human performance in multiple remote tower environments, i.e., where an ATCo is responsible for more than one remote airport at the same time. This started with analyzing eye-tracking data to characterize tower controllers' visual attention [22]. The research went on to investigate the changes in monitoring tasks and drafting multimodal interaction to support human operators at the controller working position (CWP) [23]. The latest research concentrated on workload assessment [24], operational feasibility and safety [25], as well as a supervisor position [26]. With fostering the technology maturity, questions regarding standardization with the European Organization for Civil Aviation Equipment (EUROCAE) and the European Union Aviation Safety Agency (EASA) guidelines have been developed [21]. Furthermore, the certification process for multiple remote tower operations has been sketched [27].

In the multiple remote tower environment, the human ATCo remains a central mean for the overall performance, with or without ABSR support. Related work on human performance assessment with standardized questionnaires is explained together with their results in the subsections of the result Section 3.

1.2. Structure of the Article

Section 2 describes the setup for the validation of ABSR support for ATCos and the conduction of this study. Section 3 presents the study results for the two aspects "Application of ABSR" and "ABSR in an ATM environment", i.e., results on speech recognition performance (Section 3.1) and speech understanding performance (Section 3.2) as well as on human factors such as mental workload, situation awareness, satisfaction, acceptance, trust, and usability (Sections 3.3–3.10), and ends with general feedback from ATCos (Section 3.11). Section 4 discusses the major study results for the fast readers who just quickly scanned Sections 2 and 3. For the very fast overview reader, Section 5 concludes and gives an outlook on future work. A list of abbreviations is provided before the Appendix. For more details and to follow some of the calculations, Appendix A lists results on speech-to-text performance, Appendix B lists results on text-to-concept performance, Appendix C lists the questionnaire statements of this study, and Appendix D details some validation setup views.

2. Materials and Methods

This section describes the hardware and software setup, as well as the methodology for the conduction of a human-in-the-loop simulation study to validate the benefits of an implemented ABSR prototype that was integrated with a prototypic electronic flight strip system for ATCos working within a simulated multiple remote tower environment. The technological validation exercise "006" was part of SESAR2020's wave 2 project PJ.05, "Digital Tower Technologies (DTT)" that received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation program under grant agreement No 874470. More specifically, the exercise was conducted within solution 97, "HMI Interaction modes for Airport Tower," with its "Automatic Speech Recognition (ASR)" activity for "Improving controller productivity by ASR at the TWR CWP".

2.1. Hardware Setup of the Validation Study

Figure 1 shows the hardware setup of a prototypic CWP for a multiple remote tower environment in DLR's TowerLab [28]. Three horizontal rows of monitors (top of Figure 1) visualize the artificial outside view for the three configured airports. The airport layout is generic, but the three airports are named Vilnius, Kaunas, and Palanga.

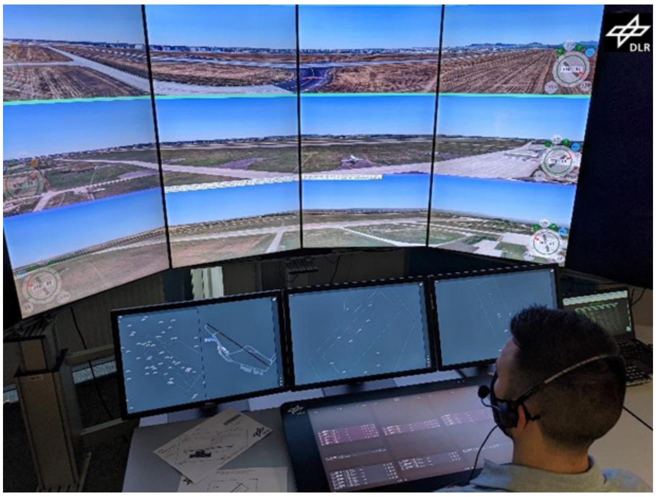


Figure 1. Multiple remote tower environments with a row of monitors per each of the three airports under ATCo control, three radar screens, and the electronic flight strip system that is supported by the output of an assistant-based speech recognition system. The position for Vilnius is always top/left, Kaunas is middle, and Palanga is bottom/right.

The three monitors below on the desk (see Figure 1) depict the air traffic in the airport's vicinity. The touch display at the middle of the desk (see Figure 1) presents the electronic flight strips per airport per column. The ATCo wears a headset with speakers and a microphone that is triggered via a push-to-talk button at the headset's cable. The paper sheets on the left of the desk (see Figure 1) contained the airport layout, aircraft callsigns, and a legend for the symbols of the electronic flight strip system.

2.2. Software Setup and Simulation Environment of the Validation Study

All used software and displays are prototypic DLR developments. They consist of the most common elements that the usual controller working positions of European air navigation service providers offer. Thus, a wide range of ATCos from many different countries can use the systems of the validation study even if the details differ compared to their "usual" systems in daily-life operations. The aircraft and ground vehicle movements relevant to the tower and ground control were simultaneously simulated in three remote Lithuanian airports, i.e., Vilnius, Kaunas, and Palanga.

2.2.1. Outside View for Supervision of Movements on Ground and above the Airfield

The artificial outside view, such as out of a physical tower for those three airports, comprises the runway, taxiways, stands, and some environments, such as landscape and buildings, as shown in Figure 1. On the left and right side of each monitor row, there was a compass rose with additional information relevant to aircraft takeoff and landing (more details in Appendix D). If the validation condition "with ABSR support" was active, the ABSR output was also shown in the ATCo outside view.

2.2.2. Radar Displays to Monitor Air Traffic Close to the Airfield

A radar display for each of the three airports (see Figure 1 middle part) visualized the airspace structure with waypoints and the air traffic in the airport's vicinity. Each aircraft had a radar label displaying the aircraft callsign, weight category, current altitude, rate of descent/climb, speed, heading, and aircraft type. The biggest airport (Vilnius) also had a ground radar display showing the runway, taxiways, stands, and aircraft information, i.e., current and latest positions, aircraft callsign, relevant runway or stand, speed, and aircraft type, as well as a color indicating if the flight is an arrival or departure.

2.2.3. Electronic Flight Strip System (EFS)

The electronic flight strip system on the touch display consisted of one column per airport (see Figure 2). The column heads presented the airport's ICAO code, runways, automatic terminal information service (ATIS) letter, and radio frequency. Each of the three columns, in turn, comprised four different bays—air, runway, ground, and stand—in order to enable managing the flight progress in a procedural way.

Each flight strip (see zoomed white box in Figure 2) offered the option for hand written notes (pen symbol in upper left area), and showed aircraft callsign (BRU835), ICAO weight category (M), runway (34), stand (M1), estimated time of arrival/departure (08:39), aircraft type (A320), flight rules ("I" or "V" for instrument/visual flight rules), origin/destination airport (EDDK), standard instrument departure (such as BELED3D for aircraft GAF612 on the lower right blue flight strip), and squawk (3511).

The EFS for the ATCos further had a number of flight status icons on the right side (see Figure 2). The flight status icons depended on the flight intentions, i.e., blue departure flight strips/purple arrival flight strips, and on the progress, i.e., in which bay the flight strips currently are. Each flight status icon could be toggled, i.e., activated when a status change was initiated or deactivated, e.g., in case of activating by accident. The different flight status icons are shown in Figure 3. If they were activated through the tap of an electronic pen, they turned into a light green color in the electronic flight strip.

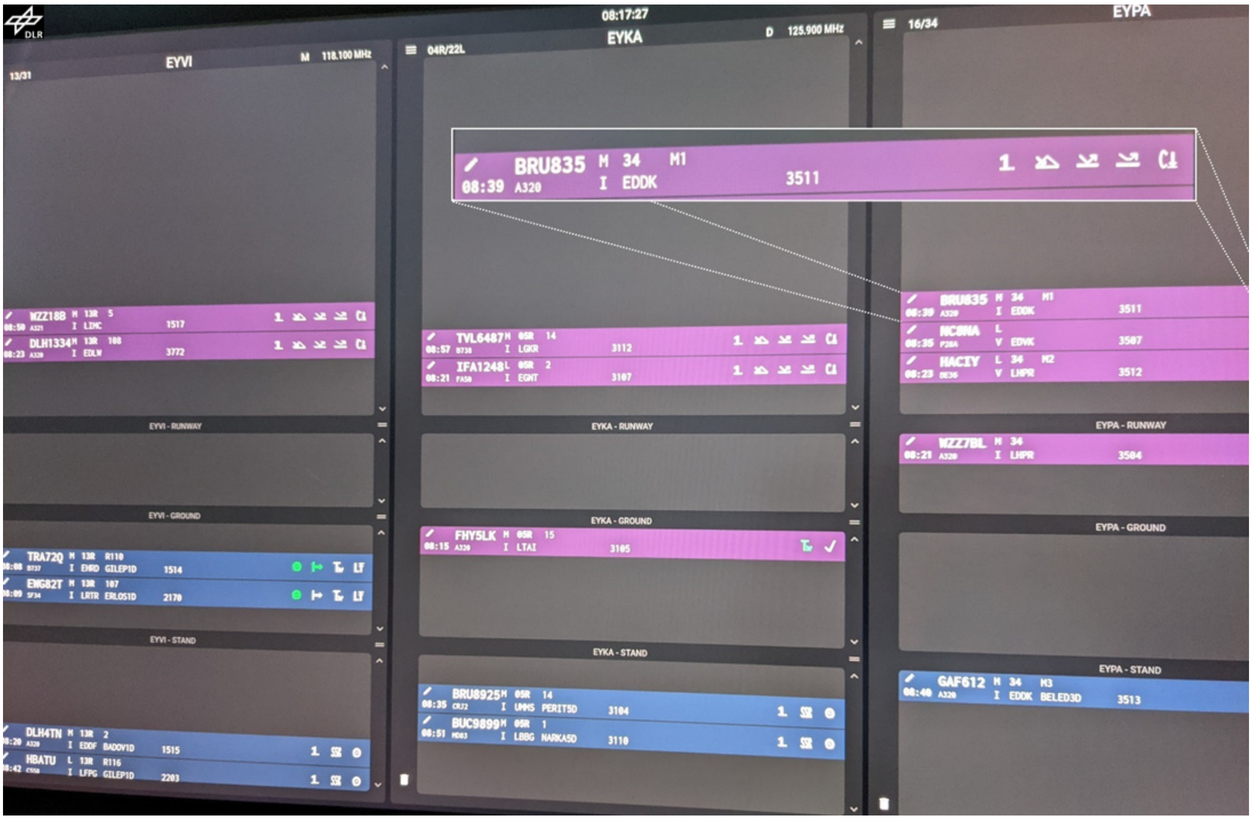
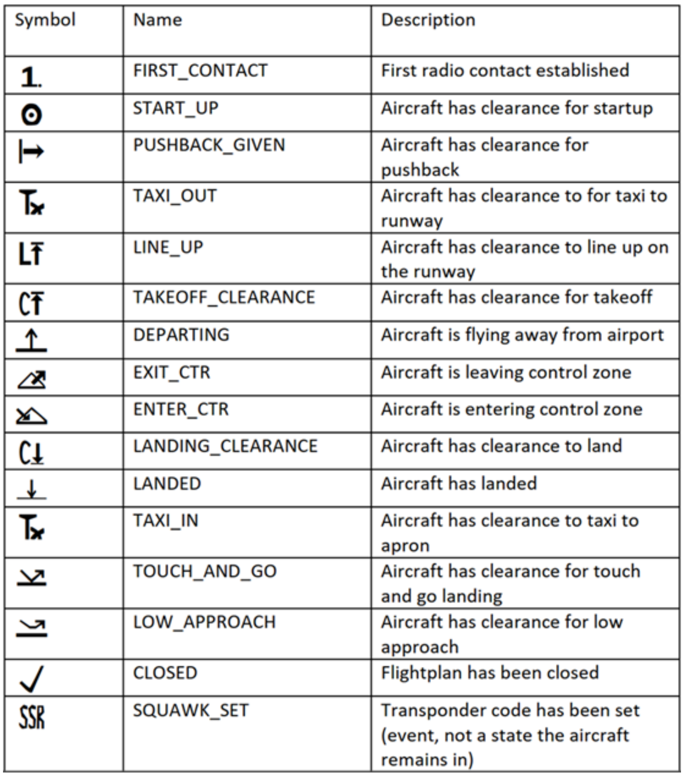


Figure 2. DLR's prototypic electronic flight strip system for aircraft at three remotely controlled airports (from left to right: Vilnius, Kaunas, Palanga).



Symbol	Name	Description
1.	FIRST_CONTACT	First radio contact established
⊙	START_UP	Aircraft has clearance for startup
↦	PUSHBACK_GIVEN	Aircraft has clearance for pushback
Tx	TAXI_OUT	Aircraft has clearance to for taxi to runway
LT	LINE_UP	Aircraft has clearance to line up on the runway
CT	TAKEOFF_CLEARANCE	Aircraft has clearance for takeoff
↥	DEPARTING	Aircraft is flying away from airport
⬈	EXIT_CTR	Aircraft is leaving control zone
⬊	ENTER_CTR	Aircraft is entering control zone
C↓	LANDING_CLEARANCE	Aircraft has clearance to land
↓	LANDED	Aircraft has landed
Tx	TAXI_IN	Aircraft has clearance to taxi to apron
⤻	TOUCH_AND_GO	Aircraft has clearance for touch and go landing
⤼	LOW_APPROACH	Aircraft has clearance for low approach
✓	CLOSED	Flightplan has been closed
SSR	SQUAWK_SET	Transponder code has been set (event, not a state the aircraft remains in)

Figure 3. Flight status icons of electronic flight strips available depending on the current flight status [29].

The electronic flight strips changed their bays with further progress of the flight status when arriving or departing, e.g., after setting the status "LINEUP," the flight strip moved from the ground bay to the runway bay.

2.2.4. Assistant-Based Speech Recognition and Understanding Prototype

The core development for the validation study was a prototypic system for speech recognition and understanding in a multiple remote tower environment. This ABSR system is based on a number of models based on deep neural networks trained by machine learning methods, respectively. The two main steps are (1) speech recognition, i.e., automatic speech-to-text transcription from tower controller audio input, and (2) speech understanding, i.e., automatic semantic text-to-concept annotations from the transcription input (see Figure 4). The speech recognition and understanding models were trained on in-domain and out-of-domain data, specifically 200 h from seven different datasets and 4.5 h (recorded in the later study environment) of manually transcribed speech data, as well as 400 h of untranscribed data from LiveATC (Homepage: https://www.liveatc.net/ (accessed on 4 April 2023)) [30]. Further references on the development of the speech recognition engine with artificial intelligence techniques can be found in [30].

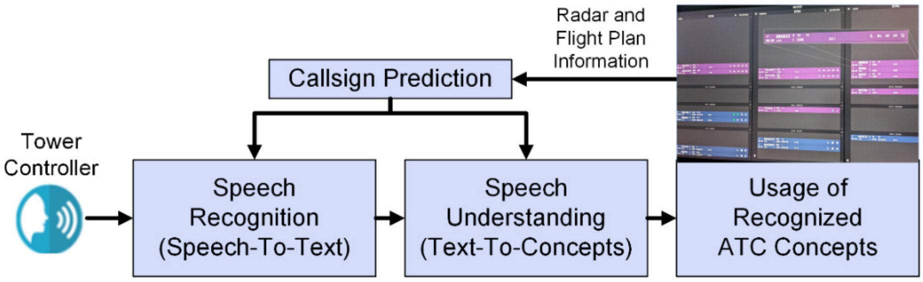


Figure 4. Components of assistant-based speech recognition (ABSR) in the multiple remote tower environment.

Both speech-to-text and text-to-concepts benefit from the use of contextual data, i.e., they consider radar data and flight plan data. The callsign prediction model is used to forecast aircraft callsigns for the next ATCo utterances, i.e., it predicts only those aircraft callsigns which are in the current area of responsibility of the ATCo. Those forecasted callsigns support the speech recognition engine in recognizing the correct word sequences and the speech understanding module in extracting the correct callsigns, especially in cases when not all words of the callsign are correctly recognized.

The command extraction model in the speech understanding module analyses the automatically transcribed ATCo utterances and extracts meaningful content, i.e., ATC concepts such as commands with callsigns, command types, values, units, etc., conform to the defined ontology. Two example transcriptions with their example annotations shall illustrate this:

- *wizz air two echo bravo good morning vilnius tower startup and pushback approved cleared to sofia* via *erlos one delta departure route seven thousand feet squawk two one seven seven QNH one zero one four*

  WZZ2EB GREETING
  WZZ2EB STATION VILNIUS_TOWER
  WZZ2EB STARTUP
  WZZ2EB PUSHBACK
  WZZ2EB CLEARED TO LBSF
  WZZ2EB CLEARED VIA ERLOS_1D
  WZZ2EB ALTITUDE 7000 ft
  WZZ2EB SQUAWK 2177
  WZZ2EB INFORMATION QNH 1014

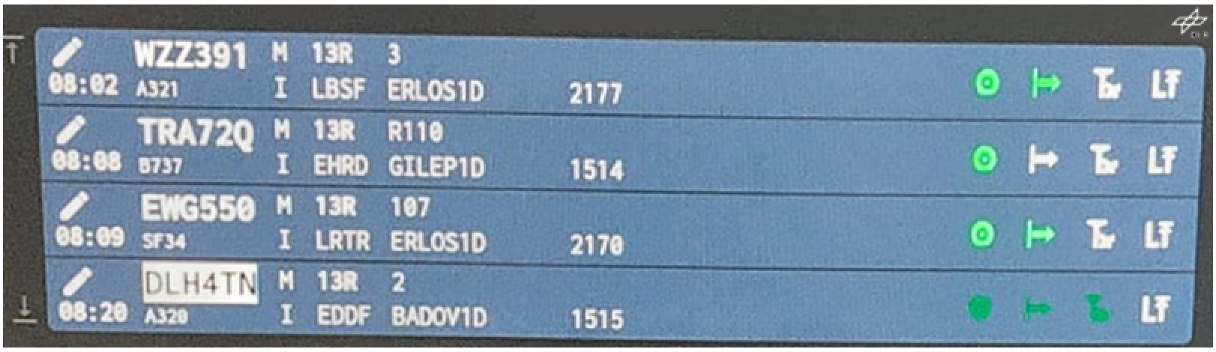
- *hotel tango uniform when you are ready taxi to holding point runway three one correction one three right* via *[hes] golf vilnius*

  HBATU CORRECTION
  HBATU TAXI TO HP_13R WHEN READY
  HBATU TAXI VIA G C WHEN READY

The recognized ATC concepts, i.e., the annotations, are then used for highlighting purposes or supporting manual input in electronic ATC systems.

2.2.5. Visualization of ABSR Output on EFS and Outside View

The ABSR output was visible through different highlighting mechanisms in the electronic flight strips if the validation condition "with ABSR support" was active. If a callsign was recognized [31], the callsign was highlighted by displaying a rectangle in inverted colors for ten seconds at the callsign field of the flight strip (see "DLH4TN" in Figure 5). The callsign was highlighted immediately after being recognized and extracted even before the ATCo finished the utterance by releasing the push-to-talk button.



	WZZ391	M	13R	3		
08:02	A321	I	LBSF	ERLOS1D	2177	LF
	TRA72Q	M	13R	R110		
08:08	B737	I	EHRD	GILEP1D	1514	LF
	EWG550	M	13R	107		
08:09	SF34	I	LRTR	ERLOS1D	2170	LF
	DLH4TN	M	13R	2		
08:20	A320	I	EDDF	BADOV1D	1515	LF

Figure 5. Prototypic electronic flight strips in the ground bay with a highlighted callsign as recognized from an ATCo utterance (DLH4TN), dark green automatically highlighted status icons for DLH4TN (STARTUP, PUSHBACK, TAXI), and five light green highlighted status icons of three other flights after being automatically accepted from the system or manually entered by the ATCo.

If one or more ATC concepts, such as commands and optionally command values, have been recognized, there was a dark green highlighting to support the ATCo in maintaining flight strips. This means the flight status icons on the right side of a flight strip or text values on the left side of a flight strip have been highlighted for ten seconds (see highlighted status icons for STARTUP, PUSHBACK, and TAXI of DLH4TN in Figure 5).

If the flight status icons in dark green mode remained unchanged by the ATCo for ten seconds, they were automatically accepted and turned into light green as with manual activation. In the case of a recognized HOLD_SHORT of runway command, the runway name was highlighted with color inversion for ten seconds as well.

2.3. ATCo Tasks in the Different Validation Conditions

Many of the tasks that ATCos needed to perform during the real-time human-in-the-loop validation study were identical under different validation conditions. Two conditions have been analyzed in the simulated multiple remote tower environment: baseline, i.e., without ABSR support and solution, i.e., with ABSR support. Section 2.3.1 describes the ATCo tasks in the baseline condition; Section 2.3.2 explains the changes induced for the ATCo when working in the solution condition.

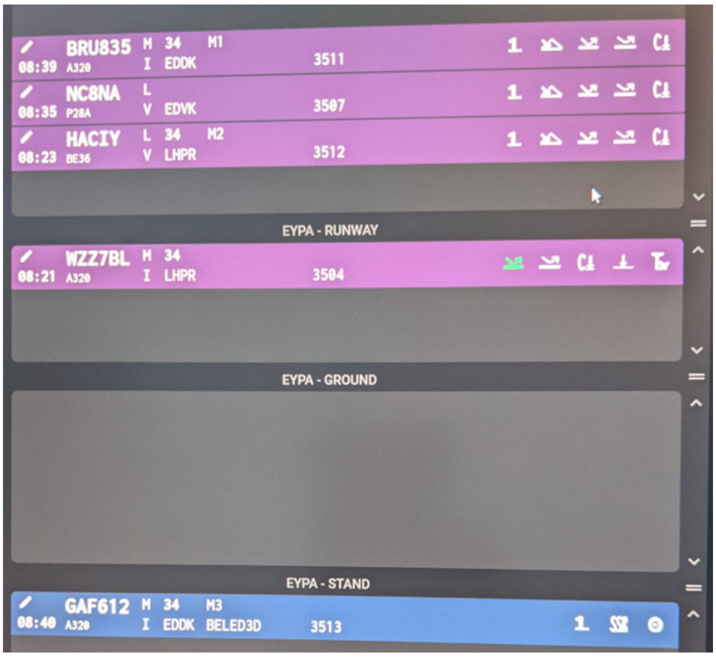
2.3.1. ATCo Primary Tasks in Baseline Condition without ABSR Support

During the simulation runs, ATCos primarily needed to control the relevant traffic at three remote airports (tower and ground), with the above-described hardware and software setup consisting of an outside view, radar displays, and the electronic flight strip system.

Hence, they mainly gave ATC clearances, allowed for startup and pushback, instructed taxi, lineup/vacate and takeoff/landing/touch-and-go clearances for the single runway in use at each airport, as well as approved to enter/leave the control zone and to contact adjacent sectors. They also had to handle special situations on the ground with aircraft and ground vehicles being involved, such as a bird strike following a runway check and an emergency landing with the disembarkation of a sick passenger. The ATCos instructed all commands to the relevant traffic verbally in the English language via an emulated radio system.

Three simulation pilots (one for each airport) in another room communicated with the ATCo to run air and ground traffic with the support of a simulation pilot interface (see Appendix D). The ATCos were instructed to speak as usual at their working position. This also implies that some ATCos stick closer to the ICAO phraseology than others. The only continuous additional content for each ATCo utterance was the name of the station the ATCos are representing with the current utterance, i.e., "vilnius/kaunas/palanga tower," in order to fulfill safety requirements of the multiple remote tower concept.

The ATCos were asked to enter the semantic content of all utterances in terms of changed flight status into the electronic flight strip system with an electronic touch pen. Thus, they had to touch the flight status icon PUSHBACK in case they verbally instructed a pushback clearance or TAXI and the name of the taxiway if there were multiple options in case they issued to taxi via a certain taxiway (see Figure 6).



BRU835 M 34 M1
08:39 A320 I EDDK 3511

NC8NA L
08:35 P28A V EDVK 3507

HACIY L 34 M2
08:23 BE36 V LHPR 3512

EYPA - RUNWAY

WZZ7BL M 34
08:21 A320 I LHPR 3504

EYPA - GROUND

EYPA - STAND

GAF612 M 34 M3
08:40 A320 I EDDK BELED3D 3513

Figure 6. Prototypic electronic flight strips (blue departures; violet arrivals) in different bays (air, runway, ground, stand) with relevant information on the left (estimated time, callsign, aircraft type and weight category, flight rules, runway, destination airport, stand, departure route, squawk) and status icons on the right (e.g., CLEARED TOUCH_GO in green, ENTER_CTR, etc.).

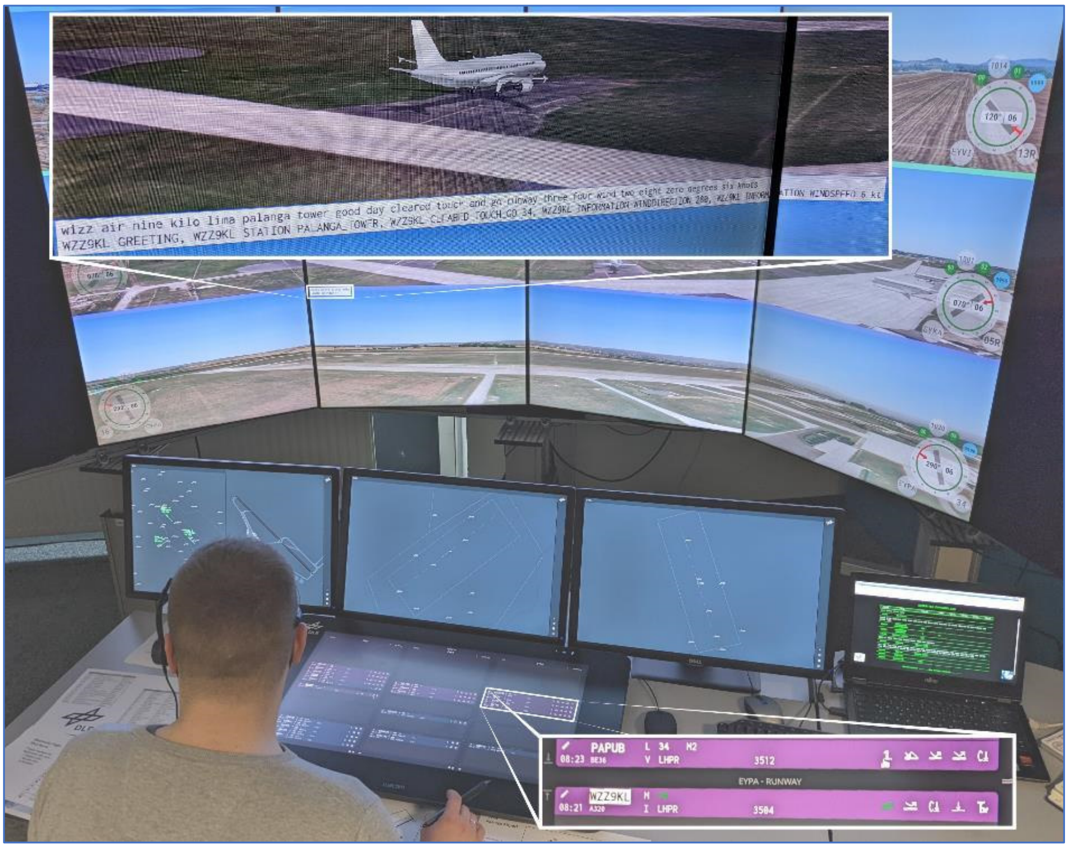
The ontology defines 80 different command types as relevant for tower ATCos if they also include the role of ground control. All of these command types have been implemented within our command extraction algorithm.

The airport topologies were rather simple, i.e., the two smaller airports (Kaunas, Palanga) had just one taxiway each from the apron to the lineup. They vacated the single runway, and only the biggest airport (Vilnius) had two taxiway alternatives each for lining up and vacating the single runway. No runway change occurred during the simulation time. The weather conditions at all three airports remained visual meteorological conditions in the daytime throughout the simulation.

The relevant traffic in the two different one-hour simulation scenarios comprised twelve flights in Vilnius (plus two ground vehicles), six flights in Kaunas (plus one ground vehicle), and five flights in Palanga—at the latter airport, including training flights with multiple approaches—so 23 flights plus three ground vehicles (the ground vehicles make 11.5% of total relevant traffic) in total. For later evaluation, the results refer to all 26 traffic vehicles (flights plus ground vehicles) as ATC communication took place between ATCos and pilots or ground vehicle drivers, respectively. The callsigns and timing of appearance of the flights in these two scenarios were slightly different in order to reduce learning effects.

2.3.2. ATCo Tasks in Solution Condition with ABSR Support

In the solution scenario, ATCos had the same hardware setup as in the baseline scenario. The only difference was the support of the ABSR system. ATCos could majorly resign from using the electronic pen to maintain flight strips and benefit from automatic maintenance through the ABSR system, i.e., the ABSR output was used to highlight the flight status icons and callsigns in electronic flight strips automatically (see lower zoomed white box in Figure 7). The ATCos only needed to check the automatically highlighted output, i.e., representing issued commands and thus changes in the aircraft flight status, and correct if needed. A video about the simulation environment in the solution runs can be downloaded from https://www.youtube.com/watch?v=Y76kQmo_ANU&cbrd=1 (accessed on 4 April 2023). The ABSR output was only shown to the ATCos in solution scenarios. However, recording of verbal utterances, automatic transcription and automatic annotation was also performed in the background in baseline runs. The flow of using speech recognition and understanding output in the flight strips can be traced in Figure 7.



wizz air nine kilo lima palanga tower good day cleared touch and go runway three four wind two eight zero degrees six knots

WZZ9KL GREETING, WZZ9KL STATION PALANGA_TOWER, WZZ9KL CLEARED TOUCH_GO 34, WZZ9KL INFORMATION WINDDIRECTION 280, WZZ9KL INFORMATION WINDSPEED 6 kt

Figure 7. ATCo in front of electronic flight strip display with highlighted callsign and flight status icons, as well as outside view with transcription and annotation of ABSR output.

The complete transcription of words (first line) and the relevant annotation of commands in the agreed ontology format (second line) have been displayed in the outside view of the human-machine interface as shown in Figure 7 (zoomed white box on the upper area of the figure) if the validation condition "with ABSR support" was active.

2.4. *Questionnaires and Further Tasks during and after Simulation Runs*

Every five minutes, the ATCos were requested to rate their workload on a displayed graphical interface for an instantaneous self-assessment of workload (ISA) scale [32]. This interface offered values from 1 (low workload) to 5 (high workload) and appeared in the EFS system (see Figure 8).

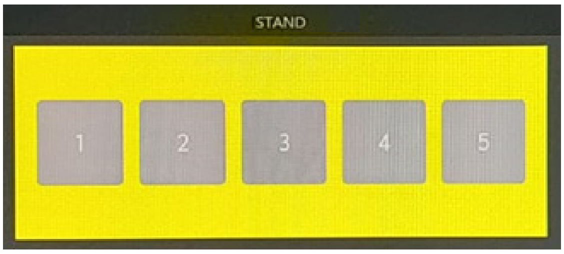


Figure 8. Instantaneous self-assessment of workload (ISA) scale to be responded to. "1" corresponds to "Under-utilized", "2" to "Relaxed", "3" to "Comfortable", "4" to "High Workload", and "5" to "Excessive Workload".

2.4.1. ATCo Secondary Tasks during Simulation Runs

Furthermore, the ATCos were asked to perform a secondary task next to their primary ATC task. After 10 and 40 min in the scenario, ATCos were requested to sort a deck of 48 cards and name one to four randomly missing cards (see Figure 9). This sorting of cards was repeated three times each or a maximum of 15 min (after 10 min) or 13 min (after 40 min), respectively. This secondary task is aimed to give a more objective impression about workload when comparing the time needed to sort and identify missing cards between baseline and solution scenarios. It is assumed that ATCos have more free cognitive capacity (less workload) if they can sort the cards quicker in one of the simulation conditions. The points in time (after 10 and 40 min) have been chosen as the ATCo workload should have been slightly increased due to the traffic situation at that time. The need to respond to ISA and to perform the card sorting remained identical in baseline and solution runs.

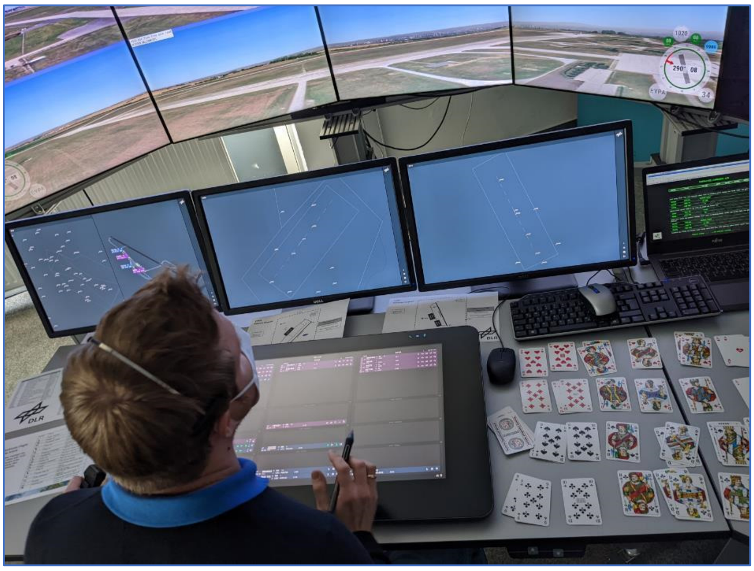


Figure 9. ATCo interrupts card sorting (secondary task) to check the outside view.

2.4.2. ATCo Post-Run Questionnaires after Simulation Runs

The post-run questionnaires needed to be filled by ATCos twice on each validation day, i.e., after each of the two simulation runs with the two different conditions. The well-established questionnaires cover the most important factors of air traffic controller work, such as situation awareness, workload, and trust [33] and are listed below:

- NASA-TLX (National Aeronautics and Space Administration Task Load Index) [34,35];
- Bedford Workload Scale [36];
- Three SHAPE questionnaires (Solutions for Human Automation Partnerships in European ATM) [37]:
    - ○ AIM-s (Assessing the Impact on Mental Workload);
    - ○ SASHA (Situation Awareness for SHAPE) ATCo;
    - ○ SATI (SHAPE Automation Trust Index);
- CARS (Controller Acceptance Rating Scale) [38];
- SUS (System Usability Scale) [39,40].

2.4.3. Statistical Analysis Approach

When reporting the results of data that has been measured for baseline and solution runs, there will also be a statistical significance analysis, e.g., of all the above-mentioned questionnaires. Usually, there is a learning effect if ATCos perform multiple simulation runs in a row, i.e., they will perform better in the later runs, because they are used to the overall environment. Hence, better performance cannot simply be assigned to possibly different simulation run conditions such as baseline or solution. The sequence of baseline and solution runs is also an independent variable.

Therefore, two measures have been taken to compensate for the sequence effects as much as possible. First, the order of simulation runs alternate, i.e., half of ATCos start with a baseline run and end with a solution run and vice versa for the other half. The performance usually is, of course, better in the later runs, but the effect on baseline and solution runs should average out. Nevertheless, the standard deviations will be higher than they would be without sequence effects. Hence, secondly, the sequence effects will be compensated by considering the performance difference between the two runs. This sequence effect compensation technique (SECT) is described in more detail in [41]. An example shall illustrate the application of SECT. If any performance in all first runs of ATCos is 50 s and in all second runs 30 s, i.e., 20 s better, the performance difference is calculated as 50–30 = 20. Half of this difference (20/2), i.e., 10 s, is subtracted from each result of a first run and half of the difference is added to each result of a second run. Afterwards, the averages per run are the same. Furthermore, the averages of baseline and solution keep the same. We had exactly half of the ATCos having a baseline run and a solution run as the first run, respectively. However, the standard deviation will decrease, i.e., statistical significance will increase. This was already shown for earlier project result analyses such as of AcListant®-Strips when analyzing workload benefits [18].

Unpaired t-Tests can only reject hypotheses with some probability α. Therefore, the so-called null hypothesis H_0 is usually the opposite of the effect to be validated, e.g., "*ABSR support does not reduce workload as measured with a secondary task*". The test value T is calculated as the product of (1) the difference between the mean value of the performance measurement and $\mu_{0,}$ which is set to zero, and (2) the square root of the number of performance measurements, i.e., ten study subjects, divided by the standard deviation of the performance measurement. If the measurement values follow a Normal Gaussian distribution, the value T obeys a t distribution with n-1 degrees of freedom. Therefore, the resulting value T is compared with the value of the inverse t-distribution at the position $t_{n\text{-}1,1\text{-}\alpha}$ with n-1 degrees of freedom. If the calculated value T is bigger than the $t_{n\text{-}1,1\text{-}\alpha}$ threshold, we can reject the null hypothesis with probability α. As this falsifies the null hypotheses, we could assume that "ABSR support does reduce workload as measured with a secondary task." Additionally, the minimum α will be calculated, i.e., so that the value T threshold is still exceeded. These calculations will be performed on all single rated statements and answered questions, respectively, as well as for the group of statements/questions that belong together in a single questionnaire, e.g., the aggregating of the six items of NASA-TLX.

2.4.4. ATCo Post-Validation Overall Questionnaire

The post-validation questionnaire requested to be filled by ATCos only once after finishing all simulation runs, i.e., there is an overall rating on the ABSR prototype instead of a rating on baseline and solution each. It contained 28 statements to be rated regarding human performance, safety, operating methods, and technical feasibility. If answers on the post-validation questionnaire of the ten ATCos are reported in the following, the scale ranges from 1 (fully disagree) to 10 (fully agree), i.e., the scale mean is 5.5.

2.5. Validation Schedule and Participants

Each validation day with an ATCo began with organizational tasks such as the signature of informed consent, a briefing, and a demographics questionnaire. It was followed by 60 min training run with low to medium traffic (30 min each with baseline and solution condition, i.e., without ABSR and with ABSR support). Then, two simulation runs of 60 min each with baseline and solution conditions, respectively, and medium traffic were carried out. One run included a bird strike, and the other run included a sick passenger in an aircraft as special situations that the ATCos needed to handle and coordinate with ground vehicles. In order to average out the influence of the learning effect, baseline and solution scenarios have been alternated for ATCos throughout the validation campaign. After each run, the ATCos were requested to fill the mentioned questionnaires regarding workload, situation awareness, etc., as sketched in Section 2.4.2 and gave comments and answers in a debriefing. Finally, ATCos needed to fill out an overall tailor-made questionnaire (see Section 2.4.4) on the ABSR system after the last debriefing.

It has to be noted that the technical team of the validation campaign replaced a laptop and made a software update regarding the allowed central processing unit (CPU) load for the automatic speech recognition (ASR) engine after the eighth ATCo in the simulation campaign. However, no significant change in ABSR accuracy was noted due to this.

The validation campaign took place at DLR TowerLab in Braunschweig, Germany, from 14 February to 3 March 2022 (8:30 a.m. to 4:30 p.m.). This study was conducted with one ATCo per day for exactly ten days with five ATCos from Oro Navigacija (ON, Lithuania) and five ATCos from AustroControl (ACG, Austria). All participants were holders of an active tower ATCo license. The ten ATCos were not involved in the project in terms of participation in previous work sessions.

The nine male and one female ATCo had an arithmetic mean age of 31.9 years (standard deviation, SD: 5.5 years). The ATCos had 7.4 years of professional working experience as an ATCo (SD: 5.8 years), while ON ATCos were already longer on duty (9 years, SD: 7.3 years) compared to ACG ATCos (5.7 years, SD: 3.9 years).

3. Results

Each of the ten ATCos participated in a baseline run without ABSR support and a solution run with ABSR support, i.e., the data of twenty simulation runs with their succeeding post-run questionnaires as well as the final ten post-validation questionnaires' answers are analyzed in the following subsections. This section details:

(1) Objectively measured speech recognition performance;
(2) Objectively measured speech understanding performance;
(3) Perceived speech recognition and understanding performance;
(4) Operational and technical questions;
(5) Overall ratings on perceived workload, perceived situation awareness, satisfaction, acceptance, trust, and usability;

(6) Ratings per simulation run on perceived and more objectively measured workload, perceived situation awareness, satisfaction, acceptance, trust, and usability;
(7) General debriefing feedback.

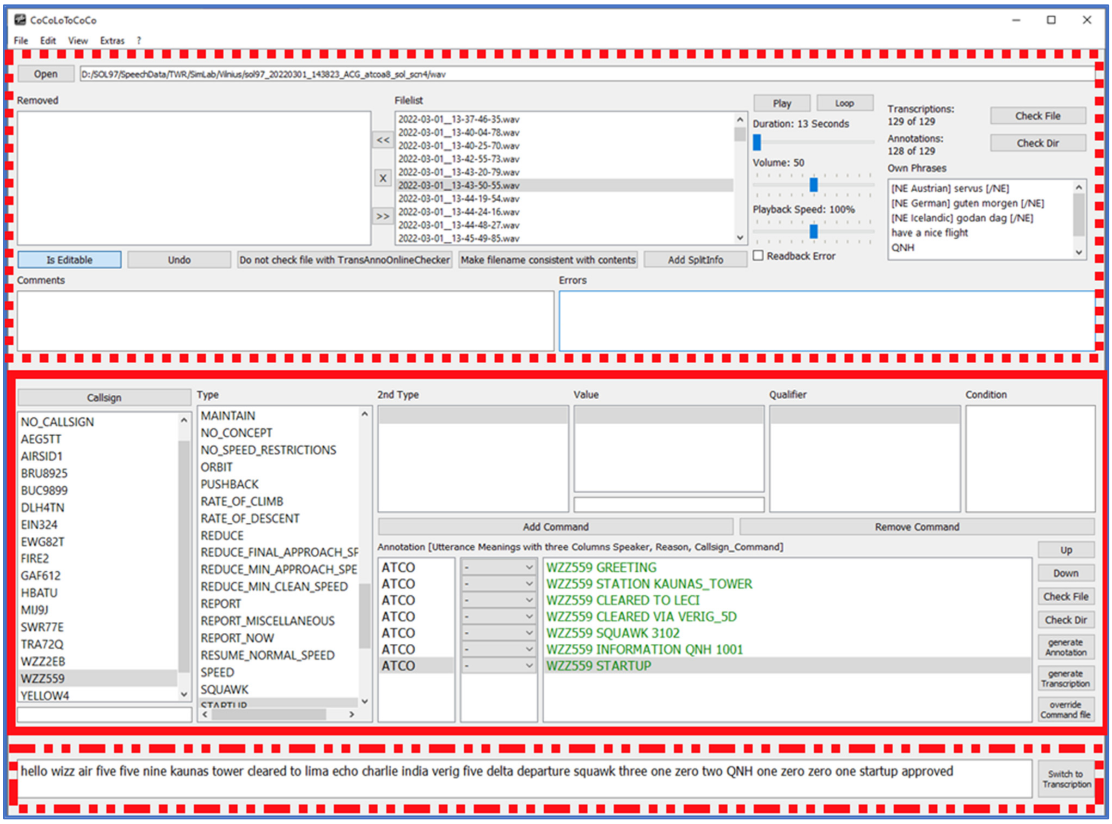
The tailor-made statements of the questionnaires to be rated by ATCos described in the following contained the term ASR for brevity, even if automatic speech recognition and understanding was meant and experienced by the ATCos. Furthermore, the ABSR performance and the effect on subjective, as well as objective results are shown in more detail on a per-case basis by comparing ON and ACG ATCos for two reasons. First, the amount of training data differs by a factor of four between ON and ACG ATCos which influences the speech-to-text and text-to-concept performance. Second, the three controller working positions that (1) the Lithuanian ATCos are used to, (2) the Austrian ATCos are used to, and (3) is used as a prototypic environment in the simulation differ so that the familiarization with the system differs as well.

3.1. *Results of Speech-to-Text Analysis*

3.1.1. Audio Recordings with Transcriptions and Annotations

Verbal utterances of ATCos that were triggered with the push-to-talk button during twenty hours of simulation runs (radar data duration) have been recorded as wav-files. For each wav-file of the twenty simulation runs (baseline and solution) exists an automatic transcription and an automatic annotation. We recorded 2427 wav files with a net speech time of 4.5 h (i.e., when ATCos speak) during 20 h of radar simulation, i.e., the frequency load by ATCos was roughly 22%. The average duration per utterance was 6.6 s.

All wav-files have been manually transcribed and annotated ("gold") with DLR's Controller Command Logging Tool for Context Comparison (CoCoLoToCoCo, see Figure 10) to enable comparison and calculations about recognition and error rates on the word level and semantic level.



D:/SOL97/SpeechData/TWR/SimLab/Vilnius/sol97_20220301_143823_ACG_atcoa8_sol_scn4/wav

2022-03-01__13-37-46-35.wav
2022-03-01__13-40-04-78.wav
2022-03-01__13-40-25-70.wav
2022-03-01__13-42-55-73.wav
2022-03-01__13-43-20-79.wav
2022-03-01__13-43-50-55.wav
2022-03-01__13-44-19-54.wav
2022-03-01__13-44-24-16.wav
2022-03-01__13-44-48-27.wav
2022-03-01__13-45-49-85.wav

[NE Austrian] servus [/NE]
[NE German] guten morgen [/NE]
[NE Icelandic] godan dag [/NE]
have a nice flight
QNH

NO_CALLSIGN
AEG5TT
AIRSID1
BRU8925
BUC9899
DLH4TN
EIN324
EWG82T
FIRE2
GAF612
HBATU
MIJ9J
SWR77E
TRA72Q
WZZ2EB
WZZ559
YELLOW4

MAINTAIN
NO_CONCEPT
NO_SPEED_RESTRICTIONS
ORBIT
PUSHBACK
RATE_OF_CLIMB
RATE_OF_DESCENT
REDUCE
REDUCE_FINAL_APPROACH_SP
REDUCE_MIN_APPROACH_SPE
REDUCE_MIN_CLEAN_SPEED
REPORT
REPORT_MISCELLANEOUS
REPORT_NOW
RESUME_NORMAL_SPEED
SPEED
SQUAWK
STARTUP

Annotation [Utterance Meanings with three Columns Speaker, Reason, Callsign_Command]

ATCO	-	WZZ559 GREETING
ATCO	-	WZZ559 STATION KAUNAS_TOWER
ATCO	-	WZZ559 CLEARED TO LECI
ATCO	-	WZZ559 CLEARED VIA VERIG_5D
ATCO	-	WZZ559 SQUAWK 3102
ATCO	-	WZZ559 INFORMATION QNH 1001
ATCO	-	WZZ559 STARTUP

hello wizz air five five nine kaunas tower cleared to lima echo charlie india verig five delta departure squawk three one zero two QNH one zero zero one startup approved

Figure 10. Software tool CoCoLoToCoCo to support transcription and annotation of ATC utterances.

The upper area of CoCoLoToCoCo (red dotted line) lists all audio files of a selected folder, has buttons and sliders to adjust the playback of the files, has a comment window and an error output window, as well as offers some further file-checking opportunities. The middle area (red solid line) shows the annotation view with a column per element of a controller command, the resulting annotation of an audio file in ontology format [9] (green font), and further buttons for rearranging and checking. The lower area (red point-dash line) visualizes the transcription of a selected audio file following defined transcription rules.

The gold transcriptions of the validation trials contain in total 37,238 words without words that are not fully uttered and thus contain a "*" such as "*lufthan**" due to our transcription rules, i.e., each ATCo utterance contains roughly 15 words. Table A3 shows the top-25 1-grams, i.e., the uttered words with their absolute and relative frequency. The most often occurring words, "one" (6.43%) and "zero" (3.97%), are usually in the top three for other ATC communication corpora as well. However, the word at rank three, "tower" (3.96%), is specific for the multiple remote tower environment, in which the transmitting entity should always be named and, therefore, appears quite often. Normally, the digits from zero to nine fill the first ten ranks in ATC communication corpora.

Furthermore, the words "*runway*," "*to*," and "*cleared*" appear in the top 12 as runway clearances and "*cleared to*" are often uttered. This latter result is confirmed by analyzing two real-life ATCo utterance corpora from Vilnius tower, as well as from Vienna tower, with roughly 7500 words in total each. This shows that the simulation setup and the challenges for the speech-to-text engine were quite realistic.

Table A4 lists the number of different words to reach a relevant portion of all uttered words, i.e., if speech-to-text performs well on the 100 most often occurring words, almost 90% of the total number of words are covered.

3.1.2. Speech-To-Text Performance

Some abbreviations that are used for analyzing purposes in the following and in the Appendices A and B are introduced:

- Onl = online (analysis as experienced by ATCos during simulation runs);
- Off = offline (analysis of audio files after the simulation runs);
- WER = Word Error Rate;
- Subs = Substitutions;
- Del = Deletions;
- Ins = Insertions;
- LevenDist = Levenshtein Distance [42] between automatic and gold transcription;

The speech-to-text accuracy is presented with details per each simulation run in the tables of Appendix A (see Tables A1 and A2). Table A1 visualizes the WER for offline recognition (Off) as evaluated after the end of the validation trials. It shows what results would be already achievable when the technical setup is improved to deliver the offline performance during the simulation runs. Table A2 visualizes the WER for online (i.e., real-time) recognition from the voice stream (Onl) as evaluated during the simulation runs, i.e., the WER are usually worse than for Off.

There were some technical problems with the ABSR setup: (1) the audio device continuously disconnected in one simulation run resulting in the loss of some data, and (2) there was partly CPU overload, especially for the first eight ATCos. The performance of the ASR engine was much worse in the online mode (as experienced by ATCos) than in the later offline analysis of recorded audio files. Worse speech-to-text performance, i.e., a higher WER being the sum of substitutions, insertions, and deletions regarding two-word sequences divided by the total number of correct words, of course also led to worse text-to-concepts performance. Some average and some specific results from these tables are analyzed deeper in the following.

The average WER for all twenty runs was 5.1% in Off mode. When just considering solution runs, the average WER even reached 4.4%, while baseline runs have an average

WER of 5.7%. When omitting the single run with audio device problems, the maximum WER was below 8% for all other 19 simulation runs in Off mode, i.e., the highest WER in that single run was 11.5%, and the lowest WER for any run was 1.3%. It needs to be admitted that the training data already contained a few speech samples from some ATCos that also participated in the final validation trials.

In Onl mode, the average WER was 13.6%, while the average WER for solution runs was 9.8% and for baseline runs 17.4% (see Table A2). There is a remarkable difference in the WER of ON ATCos (6.8%) compared to ACG ATCos (12.8%) in solution runs. This probably goes back to the amount of training data in the identical recording environment to the later validation trials, which was only 3.6 h for ON and even 0.9 h for ACG.

Four of twenty runs still achieved good performance with WER < 3%. However, three other runs that were affected by technical problems achieved a WER > 23%. Still, the Onl performance was sufficient in almost all solution runs to produce an acceptable text-to-concept quality. Nevertheless, the degradation of the speech-to-text performance is higher from offline mode to online mode than expected and offers room for improvement.

3.2. Text-To-Concept Quality

3.2.1. Description of Gold Annotation Data Set

All twenty simulation runs consist of 7560 commands (ALL), whereof 3701 are from baseline runs (BAS), and 3859 are from solution runs (SOL), respectively. Hence, there were 3.1 commands per ATCo utterance and 5.1 words per command if we assume that all words of an utterance are relevant to form a command.

However, it has to be noted that there are some word sequences annotated as commands that do neither influence the aircraft status nor include any request, report or traffic information from the ATCo side:

- First, the annotations GREETING (e.g., "hello"), FAREWELL (e.g., "bye"), and NO_ CONCEPT (e.g., "thanks;" no relevant ATC command in the utterance) that are summing up to 9.8% of commands during this study. These command types can indicate that the ATCo workload might not be assumed as overwhelmingly high if they still have time for welcoming, saying goodbye, and thanking anybody.
- Second, the annotation CORRECTION and CALL_YOU_BACK (e.g., "standby") that sum up to 1% of the commands might indicate a higher workload as ATCos often correct themselves, are asking for repetition of the transmission or are telling to wait for further information. The annotation SAY_AGAIN, which also belongs to this command group, has not been used.
- Third, the annotation AFFIRM and one annotation of DISREGARD that sum up to 4.1% of the commands have ATC communication relevant content, even if they are no commands in a classical sense. The annotation NEGATIVE, that also belongs to this command group, has not been used.

Though, the above-listed annotations enable a workload analysis of human ATC operators that will be published in another paper. 15 of the 80 possible command types for tower ATCos as defined in the ontology, such as GO_AROUND and ABORT TAKEOFF, did not occur at all in the 7560 commands. This means 65 different command types have been used by the ten ATCos, e.g., PUSHBACK, TAXI TO, CLEARED TAKEOFF/LANDING, ENTER_CTR, etc. Table A5 lists the relative occurrence of all command types greater than 1%. The last type, "others", groups all command types that occurred between 0.33% and 1%, such as CONTACT, ENTER_CTR, LINEUP_BEHIND, CLIMB, and DIRECT_TO. In total, there are 36 different command types that appeared more than 25 times, i.e., more than 0.33%.

The most often used command type is—unsurprisingly—STATION, as ATCos were asked to utter it in each radio transmission. However, 1529 occurrences (20.2% of commands) in 2427 utterances mean that ATCos did not follow this multiple remote tower safety-related request in 37% of all utterances. This might not be critical if ATCos just uttered "bye," but in any case, it should be considered for the multiple remote tower concept. The (CONTINUE) TAXI TO/VIA commands sums up to 11.5% of commands. The INFORMATION WINDSPEED/DIRECTION even sum up to 15% of the commands as they were instructed for all takeoffs and landings/touch-and-gos. The exclusive runway clearances CLEARED TAKEOFF/LANDING/TOUCH_GO/VISUAL sum up to 6.8% of commands. The runway usage clearances LINEUP, LINEUP_BEHIND, VACATE (VIA), and BACKTRACK sum up to 4% of commands.

A total of 29 of those 65 used command types occurred a maximum of 25 times for all ATCos in total such as BACKTRACK, CLEARED VISUAL, HOLD_SHORT, JOIN_TRAFFIC _CIRCUIT, LEAVE_CTR VIA, and ORBIT. For the above considerations, we neglect that only 87% of all words that are available in the gold transcriptions have been used by the automatic command recognition algorithm to classify commands (see column "*Unknown Classified Rate*" in Tables A6, A8 and A10).

It needs to be mentioned that our prototype follows a more holistic approach than some very basic prototypes of other actors in the field of speech recognition and understanding [43]. Our command extraction algorithm does not only extract callsigns (DLH4TN), basic types (TAXI), and values, but more sophisticated command types of multiple parts (TAXI TO/VIA), units, qualifiers, conditions (WHEN READY), chain commands with multiple callsigns, tackles many types of corrections through the ATCo and even robustly recognizes elements of the ontology if there are minor and major (acceptable) deviations from ICAO phraseology [44] in the utterances. Furthermore, we support a bigger number of command types (from the agreed ontology) as defined by the different actors themselves. The execution time of the command extraction per utterance in offline mode on a standard laptop, i.e., on a complete transcription, has an arithmetic mean of 2 ms and a median of 1.2 ms with a minimum execution time below 0.1 ms and a maximum execution time below 40 ms independent of performing command extraction on gold, offline or online transcription files. In addition, our prototype is—to the best of our knowledge—the first to support multiple remote towers at the same time (not just one) and delivers recognition error rates on an acceptable level despite all the above-mentioned complex add-ons.

3.2.2. Description of Results of Automatically Extracted Commands on Different Versions of Speech-To-Text Transcriptions

The following three subsections present recognition and error rates on callsign and command level, as well as the portion of words from the utterances that have not been used for ATC concept extraction while referring to Appendix B. More details on the semantic level metrics can be found in [45]. The command extraction results will also be presented by comparing the different command type groups:

- "All;"
- "Relevant" if appearing more than 25 times in all 20 runs;
- "EFS" has a visible effect on the electronic flight strips;
- "Status" that changed the aircraft status in the electronic flight strips;
- "Outside" is just shown on the monitors for the outside view;
- "Hypo-EFS" could have been highlighted in the flight strips but have not been during the trials, such as recognizing the active runway in an utterance.

3.2.3. Speech Understanding Performance on Gold Transcriptions

In total, 65 different command types have been automatically extracted from the gold transcriptions, i.e., the same number as in gold annotations. Table A6 shows how well the ontology-conform automatic recognition of ATC commands is modeled. The command recognition rate is around 96% with an error rate below 2.5%; the rejection rate (not reported herein) causes a difference to 100% in the total sum of command rates. The callsign recognition rate even achieved 99.8% with an error rate of 0.2%. The command recognition rates in solution runs were 96.6% for ON and 95.4% for ACG.

A total of 18.3% of all problematic annotations (recognized commands) go back to the three ground vehicles in the scenario that make up 11.5% of all relevant traffic. Further, 7.3% of problematic annotations go back to the emergency aircraft, even if this makes up 3.8% of the flights.

18 of the 80 defined command types from the ontology had visible effects in the flight status icons of the electronic flight strips—hereinafter referred to as command type group *Status*. Three further commands had a visual effect on the textual data of the electronic flight strips. These 21 commands that influenced the appearance of the electronic flight strips are grouped in the command type group *EFS*. Three supported commands contained weather information from the *Outside* view (QNH, INFORMATION WINDDIRECTION and WINDSPEED); the values of four further supported commands could have been displayed in the relevant field of the electronic flight strip. However, this highlighting has not been fully implemented yet (command group *Hypo-EFS*), i.e., STATION, INFORMATION ATIS, INFORMATION ACTIVE_RWY, and HOLD_SHORT for all possible airfield elements such as taxiways. The command type group *Relevant* includes all commands that have been automatically extracted more than 25 times. Table A7 shows the command recognition performance on the above-mentioned command type groups, i.e., presenting command recognition rates of 96% and more.

3.2.4. Speech Understanding Performance on Offline Transcriptions

The command recognition results of Table A8 are based on the output of the speech recognition engine, i.e., the transcription from Off mode. The command recognition rate is above 91%, with an error rate below 5%. The callsign recognition rate achieved almost 98.5% with an error rate below 1%. The command recognition rate of command type group *EFS* is beyond 93%, as Table A9 shows. 16.2% of all problematic annotations go back to the three ground vehicles that comprise 11.5% of all relevant traffic.

3.2.5. Speech Understanding Performance on Online Transcriptions

Tables A10 and A11 present the command recognition results on transcriptions from Onl mode. The command recognition rates are roughly 10% worse than in Off mode. The command recognition rate for solution runs in which the ATCos saw the ABSR output was 82.9%, with an error rate of 6.6%. However, there is a huge difference in the command recognition rate for ON ATCos (88.0% based on WER of 6.8%) compared to ACG ATCos (77.7% based on WER of 12.8%). As the command recognition rates for ON and ACG ATCos were both close to 96% on gold transcriptions, the high WER resulting from the mentioned low amount of available training data was a major impact on the online command recognition next to some deviations of ATCos from ICAO phraseology. The online callsign recognition rate achieved 94.2% with an error rate of 2.4%. This again shows the influence of the high WER on the ATC concept extraction.

The following measurements, especially the questionnaire ratings of ATCos, are based on the Onl mode, as this performance was "experienced" by ATCos during simulation runs.

3.2.6. Subjectively Perceived Speech Recognition and Understanding Performance and Functionality (Post-Validation)

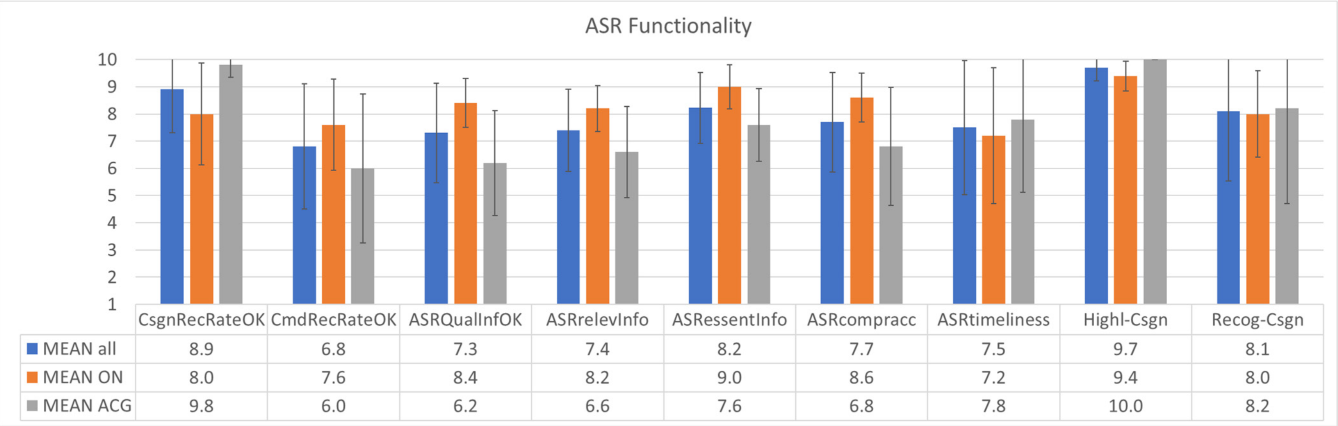
The post-validation questionnaire contained nine statements about technical feasibility with respect to the recognition and error rate of callsigns and commands as well as the ASR functionality:

1. The recognition rate and recognition error rates for callsigns by ASR were at an acceptable level. [CsgnRecRateOK];
2. The recognition rates and recognition error rates for commands by ASR were at an acceptable level. [CmdRecRateOK];
3. Overall, the level and quality of information provided by ASR were an acceptable level. [ASRQualInfOK];

The post-validation questionnaire contained four statements about the ASR interface:

4. The ASR tool interface (HMI) provides suitable access to relevant information in all situations. [ASRrelevInfo];
5. The ASR tool interface (HMI) does not display any non-essential information (clutter). [ASRessentInfo];
6. The ASR tool display is both comprehensible and acceptable. [ASRcomprehaccep];
7. The timeliness of the ASR tool display is within acceptable limits. [ASRtimeliness];
8. Automatic Speech Recognition (ASR) highlighting aircraft callsigns in the electronic flight strip display technically worked well. [Highl-Csgn];
9. Automatic Speech Recognition (ASR) highlighting aircraft callsigns in the electronic flight strip display supports recognizing which aircraft callsign has been (speech) recognized quickly. [Recog-Csgn].

The results are shown in Figure 11. ATCos rated the recognition of callsigns as almost perfect, with a mean value of around 9 on a scale from 1 to 10. The recognition rates of ATC commands were also perceived as good, with a mean value of around 7. The general quality level of information presentation from ASR was rated to be at an acceptable level with a mean value of slightly beyond 7. It has to be noted that the command recognition and overall ASR information displayed were rated much higher from ON than from ACG ATCos. This is most probably due to the underlying WER of 13% for ACG ATCos and 7% for ON ATCos, which is, however, still improvable to reach the 4% WER of offline analysis. Relevant information about the ABSR system can be assessed (mean value 7.4, but more than 1.5 points rated higher by ON than by ACG). The ASR tool seems to only present essential information with a mean value of 8.2 (again, ON rated almost 1.5 points higher than ACG). The ASR visualization is perceived as comprehensible with a mean value of 7.7 (again, ON rated almost 2 points higher than ACG). Finally, the output of the ABSR system was shown timely (mean value 7.5) due to the ATCo feedback.



	CsgnRecRateOK	CmdRecRateOK	ASRQualInfOK	ASRrelevInfo	ASRessentInfo	ASRcompracc	ASRtimeliness	Highl-Csgn	Recog-Csgn
■ MEAN all	8.9	6.8	7.3	7.4	8.2	7.7	7.5	9.7	8.1
■ MEAN ON	8.0	7.6	8.4	8.2	9.0	8.6	7.2	9.4	8.0
■ MEAN ACG	9.8	6.0	6.2	6.6	7.6	6.8	7.8	10.0	8.2

Figure 11. Subjective ATCo ratings on ASR accuracy and functionality.

The highlighting of callsigns in the electronic flight strip display (*Highl-Csgn*) was perceived as working technically very well, with a mean of 9.7 on a 10-point scale and a low standard deviation of 0.5. The second statement *Recog-Csgn* rated with a mean value of 8.1, helped the ATCos to detect which aircraft callsign has been recognized by the ABSR system. This information is needed to decide whether the following recognized ATC commands are highlighted for the correct callsign. The interesting part of these answers is the comparison with the objective measurements, i.e., the online callsign recognition rates, which are 92.1% for Lithuanian ATCos and 91.3% for Austrian ATCos (see Table A10). The same applies to the callsign recognition error rates, which are 3.9% for ACG, and also much higher than the 2.4% for ON ATCos. We have no real answer for this discrepancy between subjective rating and objective measurement.

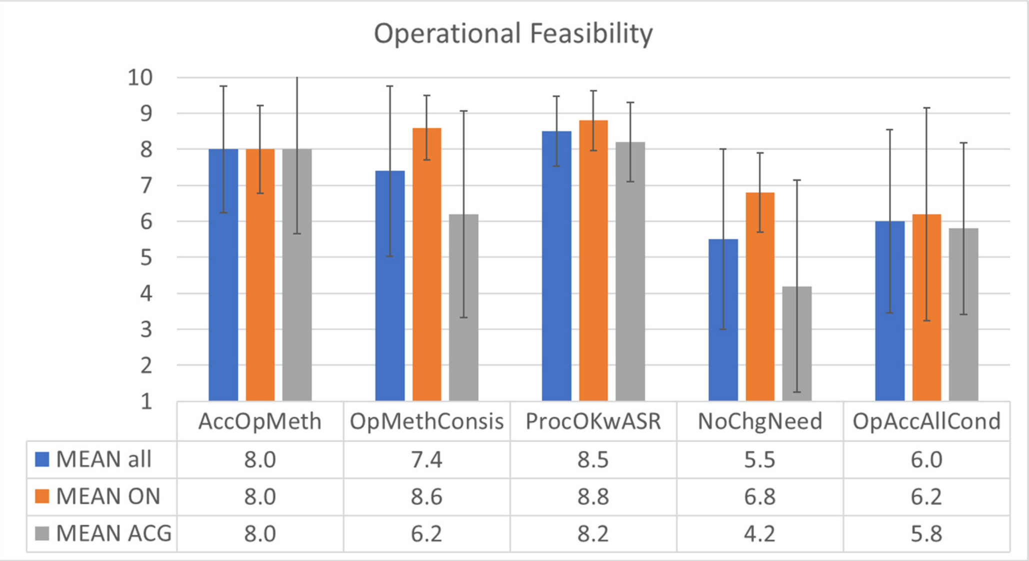
3.3. Answers to Subjective Post-Validation Questionnaires

3.3.1. Operational Use of ASR (Post-Validation)

The post-validation questionnaire contained five statements about the operational feasibility of the ASR system:

1. I can apply operating methods in an accurate, efficient, and timely manner with ASR. [AccOpMeth];
2. I think that operating methods are clearly identified and consistent in all operating conditions. [OpMethConsis];
3. Procedures and operating methods are acceptable when using the ASR tool. [ProcOK-wASR];
4. There are no changes needed to current working methods/procedures to fully support the use of the ASR tool. [NoChgNeed];
5. The ASR tool would be operationally acceptable under either nominal or non-nominal conditions. [OpAccAllCond].

The results are shown in Figure 12. The operating methods with ASR seem to be accurate, efficient, timely, and consistent in different conditions, with mean values of 8 and 7.4, respectively. Procedures and operating methods seem to be fine, with a mean value of 8.5 and a standard deviation of only 1.0. There are some changes to current working methods needed to fully support the use of the ASR tool, as the mean value equals the scale mean value of 5.5. However, ON ATCos rated this statement with almost 7, while ACG ATCos rated it with slightly above 4 points. The ASR seems to be operationally acceptable under different conditions, most probably under the majority of nominal and a few non-nominal conditions, as the ATCo rating was just slightly beyond the scale mean value.



	AccOpMeth	OpMethConsis	ProcOKwASR	NoChgNeed	OpAccAllCond
🟦 MEAN all	8.0	7.4	8.5	5.5	6.0
🟧 MEAN ON	8.0	8.6	8.8	6.8	6.2
⬜ MEAN ACG	8.0	6.2	8.2	4.2	5.8

Figure 12. Subjective ATCo ratings on operational feasibility and operating methods.

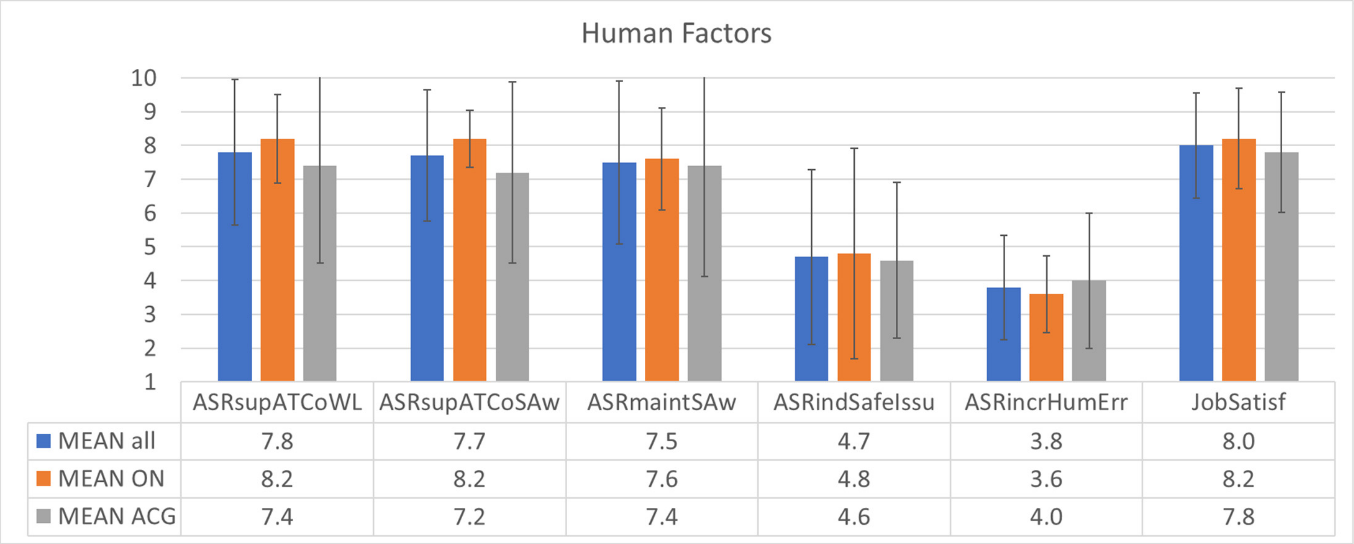
3.3.2. Human Factors Questions (Post-Validation)

The post-validation questionnaire contained six statements on human factors:

1. I think that ASR supports me in maintaining my workload at an acceptable level. [ASRsupATCoWL];
2. I think that ASR supports me in maintaining an adequate level of situational awareness. [ASRsupATCoSAw];
3. My situational awareness is maintained at an acceptable level with Automated Speech Recognition (ASR). [ASRmaintSAw];

4. I see many safety-related issues to be solved regarding automatic speech recognition implementation. [ASRindSafeIssu];
5. I think that ASR did increase the potential for human errors. [ASRincrHumErr];
6. Overall, I was satisfied performing my task with ASR. [JobSatisf].

The results are shown in Figure 13.



	ASRsupATCoWL	ASRsupATCoSAw	ASRmaintSAw	ASRindSafeIssu	ASRincrHumErr	JobSatisf
■ MEAN all	7.8	7.7	7.5	4.7	3.8	8.0
■ MEAN ON	8.2	8.2	7.6	4.8	3.6	8.2
■ MEAN ACG	7.4	7.2	7.4	4.6	4.0	7.8

Figure 13. Subjective ATCo ratings on human factors.

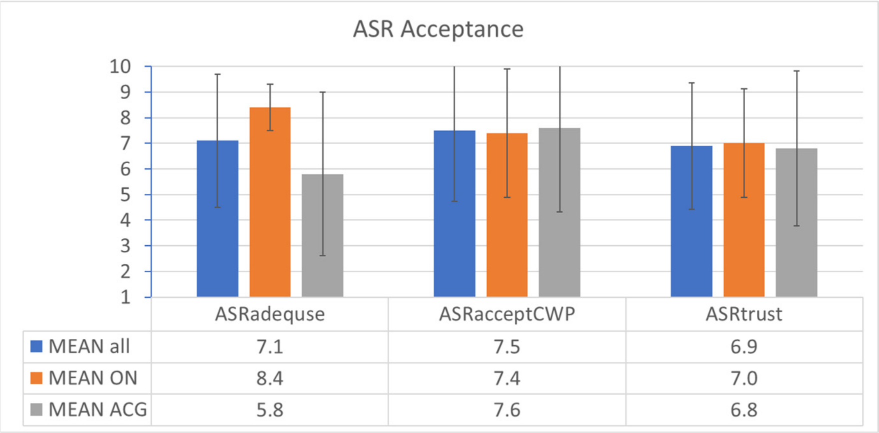
ASR seems to support maintaining situation awareness and workload of ATCos at an acceptable level with mean values of 7.5 and beyond on a 10-point scale. The *ASRsupATCoWL* statement was rated with 7.8 on a 10-point scale (90% of ATCos rated this item with 7 or above). The *ASRsupATCoSAw* statement was rated with 7.7 on a 10-point scale (90% of ATCos rated this item with 7 or above). The statement, if ASR induced safety issues or increased the potential for human errors, was rated with mean values below the scale mean of 5.5. ATCos rated their job satisfaction with using ASR high (mean value of 8 on the 10-point scale).

3.3.3. Acceptance (Post-Validation)

The post-validation questionnaire contained three statements about acceptance of and trust in the ASR system:

1. I think that the ASR system is adequately usable. [ASRadequse];
2. I would accept such an ASR system in my future tower CWP. [ASRacceptCWP];
3. My trust in the ASR system is at an acceptable level. [ASRtrust].

The results are shown in Figure 14. ATCos rated the adequate usage of ASR with a mean value of around 7. However, it has to be noted that it was rated much higher by ON than by ACG ATCos. All ATCos would accept such an ASR system in their future tower CWP with a mean value of 7.5. They trusted the ASR system with a mean value of around 7.



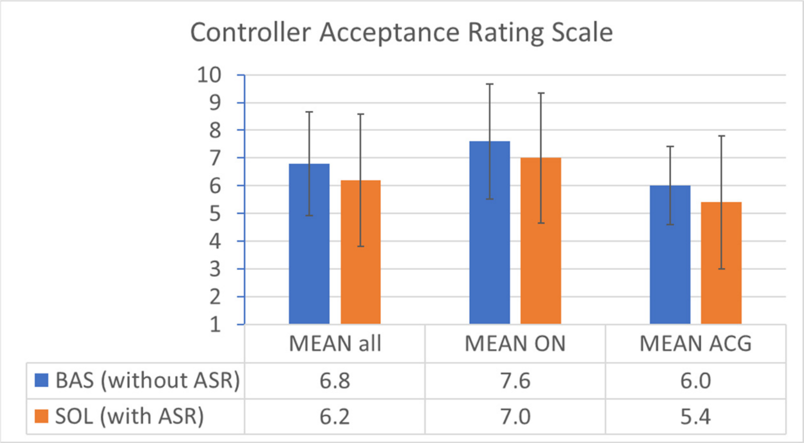
	ASRadequse	ASRacceptCWP	ASRtrust
■ MEAN all	7.1	7.5	6.9
■ MEAN ON	8.4	7.4	7.0
■ MEAN ACG	5.8	7.6	6.8

Figure 14. Subjective ATCo ratings on technical ASR acceptance.

3.4. Answers to Subjective Post-Run Questionnaires

3.4.1. Controller Acceptance Rating Scale (CARS) (Post-Run)

The post-run questionnaires contained the CARS statement to be rated on a scale from 1 to 10, with 10 being the best value, as listed in Appendix C.1. The results of the CARS questionnaire are shown in Figure 15. The acceptance was, on average, 0.6 points higher on the CARS scale for the baseline condition compared to the solution. The CARS questionnaire was filled out by each ATCo twice, once after the run with ABSR support and once after the run without ABSR support. Therefore, we are able to perform a paired *t*-test. After compensating sequence effects, the α was 0.1 to reject the inverse hypothesis that ABSR support reduces the controller acceptance due to CARS. The absolute value was 6.8 versus 6.2 (0.8 points higher for ON on average and 0.8 points lower for ACG on average).

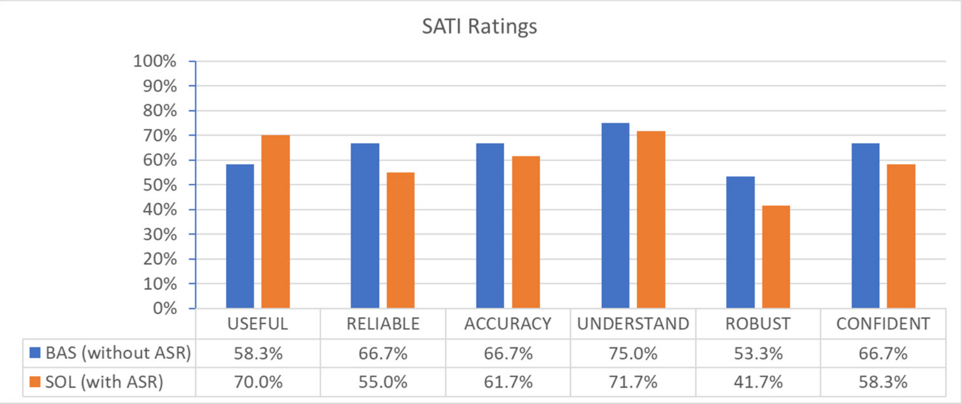


	MEAN all	MEAN ON	MEAN ACG
■ BAS (without ASR)	6.8	7.6	6.0
■ SOL (with ASR)	6.2	7.0	5.4

Figure 15. Subjective ATCo ratings on CARS.

3.4.2. Trust (SATI) (Post-Run)

The post-run questionnaires contained the six statements of SATI, as listed in Appendix C.2. The seven-item answer scale ranged from “Never, Seldom, Sometimes, Often, More Often, Very Often, and Always.” To present the results in a bar diagram, “Never” is translated to 0%, “Seldom” to 1/6 %”. . . “Very Often” to “5/6 %” until “Always” to 100%. The results are shown in Figure 16.



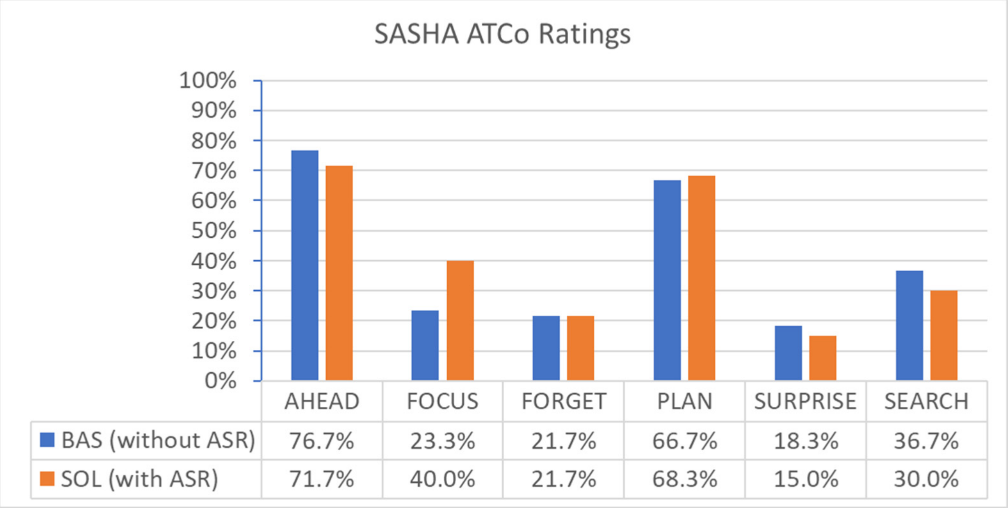
	USEFUL	RELIABLE	ACCURACY	UNDERSTAND	ROBUST	CONFIDENT
■ BAS (without ASR)	58.3%	66.7%	66.7%	75.0%	53.3%	66.7%
■ SOL (with ASR)	70.0%	55.0%	61.7%	71.7%	41.7%	58.3%

Figure 16. Subjective ATCo ratings on SATI questionnaire.

ABSR support reduced trust in automation due to SATI (α = 0.25). However, the usefulness of the system (*USEFUL* in Figure 16) was rated much better for SOL than for BAS (α = 0.05). The other five mean values are better for BAS than for the SOL condition. It is noteworthy that the four statements *RELIABLE*, *ACCURACY*, *UNDERSTAND*, and *ROBUST* from ON ATCos have better ratings for SOL than for BAS condition on average. The ambivalence of results will be discussed in Section 4.

3.4.3. Perceived Situational Awareness (SASHA ATCo) (Post-Run)

The post-run questionnaires contained the six statements of the SASHA ATCo, as listed in Appendix C.3. The seven-item answer scale ranged from "Never, Seldom, Sometimes, Often, More Often, Very Often, and Always." To present the results in a bar diagram, "Never" is translated to 0%, "Seldom" to 1/6 %"…"Very Often" to "5/6 %" until "Always" to 100%. The results are shown in Figure 17.



	AHEAD	FOCUS	FORGET	PLAN	SURPRISE	SEARCH
■ BAS (without ASR)	76.7%	23.3%	21.7%	66.7%	18.3%	36.7%
■ SOL (with ASR)	71.7%	40.0%	21.7%	68.3%	15.0%	30.0%

Figure 17. Subjective ATCo ratings on SASHA ATCo questionnaire.

ABSR support reduced the situation awareness of ATCos due to SASHA (α = 0.33). However, "searching for information" was less needed in the SOL condition (α = 0.15). The mean values of the first two items, *AHEAD* and *FOCUS*, are better for BAS than for SOL conditions. The mean values of the last four items, *FORGET*, *PLAN*, *SURPRISE*, and *SEARCH*, are equal or better for the SOL condition compared to the BAS condition without analyzing standard deviations, as differences in mean values are rather small.

3.5. Perceived Workload (High Workload Contribution) (Post-Run)

The post-run questionnaires contained a free-text question about high workload: “Which factors/events/conditions have contributed to potentially high workload?”.

The structured answers and the number of ATCos noting this after each conducted simulation run (multiple notions in one questionnaire answer possible) were as follows:

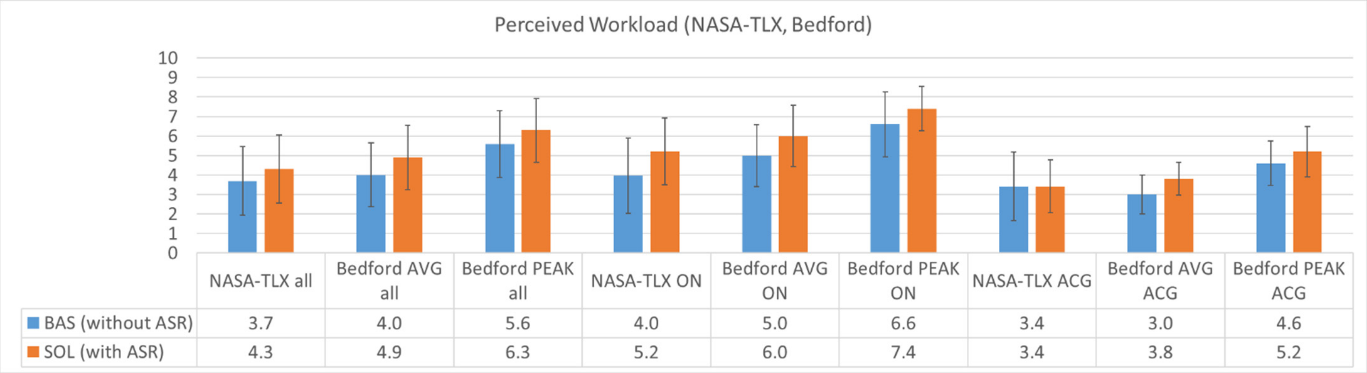
- New/unknown airspace/airport layout (especially multiple remote towers): 15 times;
- New/unknown equipment/hardware/software/electronic flight strips: 7 times;
- Checking of ABSR output (only in solution condition): 4 times;
- Unexpected/unusual air traffic situations: 3 times;
- Other: Secondary task (2 times), tower view/runway perspective (2 times), slightly different phraseology to always name the calling tower (2 times), miscommunication, system errors.

Interpreting the above results, 15 of 20 ATCo answers stated that the unknown multiple remote tower environment with unknown airport layouts induced a higher workload. Furthermore, many ATCos remarked that the flight strip handling was difficult (as some details were different from “home”). This means that the majority of workload-increasing factors can be assigned to environmental aspects that should normally not be tested in the ABSR validation trials. The above-listed checking of ABSR output, as well as unexpected situations and some further aspects, seem to have been only a minor factor for the higher workload.

3.6. Perceived Workload (NASA-TLX and Bedford Workload Scale) (Post-Run)

The post-run questionnaires contained the six statements of NASA-TLX (National Aeronautics and Space Administration—Task Load Index) as listed in Appendix C.4 and the two statements of the Bedford workload scale to rate the average workload (AVG) and peak workload (PEAK) on a scale from 1 to 10 with 10 being the highest workload. In addition, the 15 pair-wise comparisons of workload contributing factors (as the other part of the weighted NASA-TLX questionnaire) were assessed with ATCos once.

The results of the weighted NASA-TLX and the Bedford workload scale are shown in Figure 18. Figure A1 in Appendix C shows the weight per each of the six dimensions for NASA-TLX, which is almost equally distributed except for more weight for mental workload than for physical workload. The overall weighted workload (OW) due to NASA-TLX was higher for the solution than for the baseline condition: 43.1 and 38.9 (α = 0.02), respectively, with huge standard deviations around 17.5. However, the general difference between baseline and solution was only induced by the ON ATCo ratings, as the OW for ACG remained the same in baseline and solution.



	NASA-TLX all	Bedford AVG all	Bedford PEAK all	NASA-TLX ON	Bedford AVG ON	Bedford PEAK ON	NASA-TLX ACG	Bedford AVG ACG	Bedford PEAK ACG
🟦 BAS (without ASR)	3.7	4.0	5.6	4.0	5.0	6.6	3.4	3.0	4.6
🟧 SOL (with ASR)	4.3	4.9	6.3	5.2	6.0	7.4	3.4	3.8	5.2

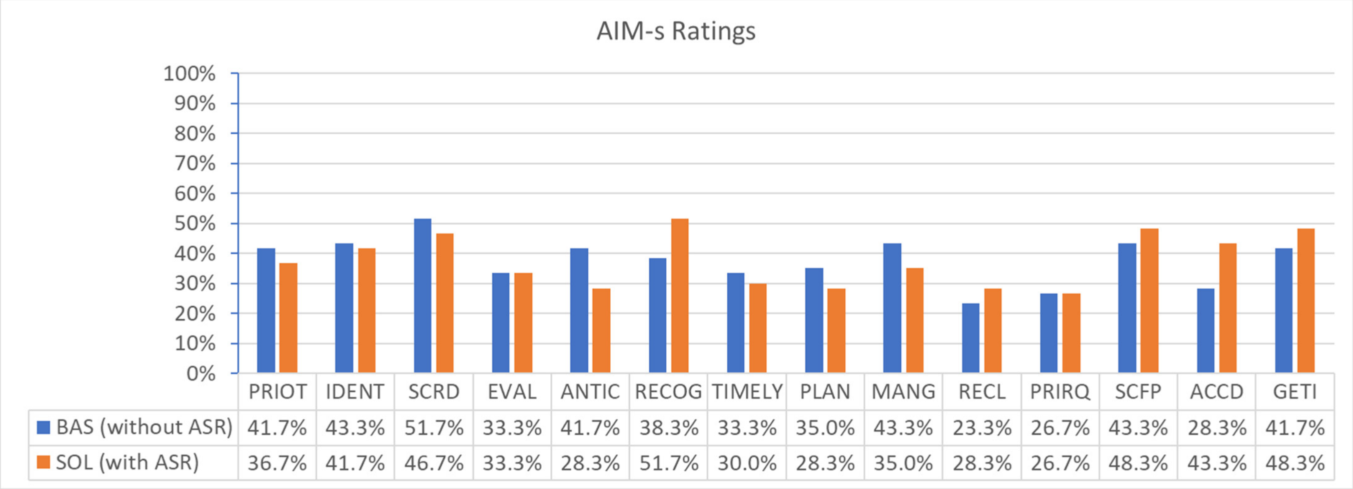
Figure 18. Subjective ATCo ratings on NASA-TLX (Weighted Overall Workload).

Furthermore, a clear learning effect during the validation day in terms of NASA-TLX OW can be seen. Those five ATCos who started with a baseline, rated the baseline (their first run) with an OW of 41.9; those five ATCos who started with a solution, rated the baseline (their second run) with an OW of 32. Those five ATCos who started with the solution, rated the solution (their first run) with an OW of 48.9; those five ATCos who started with baseline, rated the solution (their second run) with an OW of 37.2.

The average and peak Bedford workload were 0.9 and 0.7 points higher, respectively, in the solution condition with ABSR support compared to the baseline condition (α = 0.001). The peak workload was roughly 1.5 points higher than the average workload. The workload level, in general, was roughly two points lower for ACG than for ON ATCos.

3.7. Perceived Workload through Automation Impact (AIM-s) (Post-Run)

The post-run questionnaires contained the sixteen statements of AIM-s as listed in Appendix C.5. The seven-item answer scale ranged from "None, Very Little, Little, Some, Much, Very Much, Extreme." To present the results in a bar diagram, "None" is translated to 0%, "Very Little" to 1/6 %"... "Very Much" to "5/6 %" until "Extreme" to 100%. The statements SHARE and TMN are not analyzed further as there were no team members during the simulation runs (fourteen statements remain). Figure 19 shows the results.



	PRIOT	IDENT	SCRD	EVAL	ANTIC	RECOG	TIMELY	PLAN	MANG	RECL	PRIRQ	SCFP	ACCD	GETI
🟦 BAS (without ASR)	41.7%	43.3%	51.7%	33.3%	41.7%	38.3%	33.3%	35.0%	43.3%	23.3%	26.7%	43.3%	28.3%	41.7%
🟧 SOL (with ASR)	36.7%	41.7%	46.7%	33.3%	28.3%	51.7%	30.0%	28.3%	35.0%	28.3%	26.7%	48.3%	43.3%	48.3%

Figure 19. Subjective ATCo ratings on AIM-s questionnaire.

After compensating sequence effects, the overall perceived workload due to AIM-s is not statistically better with or without ABSR support. We measured an α of 0.49, which is not better than throwing a coin. However, the anticipation of the future air traffic situation was much better for SOL than for BAS (α = 0.02). Nine of the fourteen statements have been rated better on average (less) for the SOL condition than for the BAS condition. Only the five statements related to information *RECOG*, *RECL*, *SCFP*, *ACCD*, and *GETI* have been rated worse for SOL condition compared to BAS condition.

3.8. Perceived Workload (Instantaneous Self-Assessment of Workload (ISA)) (Within-Run)

During each simulation run, ATCos needed to rate their workload of the recent five minutes on a scale from 1 (bored) to 5 (almost overloaded). The results are shown in Figure 20. The average ISA workload was 0.1 points less, i.e., better, in solution condition with ASR support compared to baseline condition with α = 0.15 (2.1 and 2.0 points, respectively).

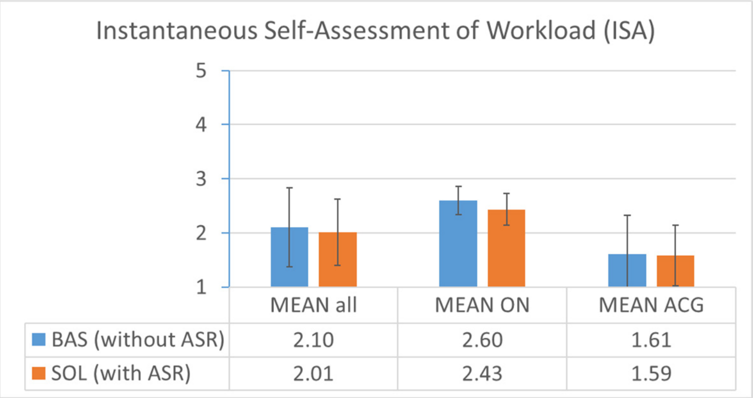


Figure 20. Subjective ATCo workload self-assessment (ISA).

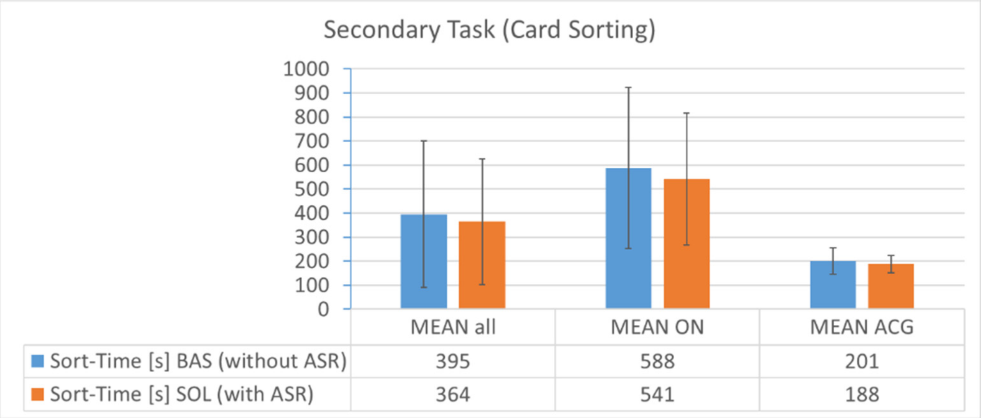
The ISA of ON ATCos was on a higher level with 2.6 and 2.4, respectively, and had a much lower standard deviation of below 0.3. The ISA score of ACG ATCos was around 1.6, with a standard deviation more than twice as much as of ON ATCos.

3.9. Objectively Measured Workload with Secondary Task (Card Sorting) (Within-Run)

The ATCos always needed to make sure that their primary task of doing ATC remains safe and efficient. However, if they had time for a secondary task, i.e., free mental capacity, they should sort cards. This method has already been used in earlier ASR projects to generate a more objective measure of mental workload than just via self-ratings.

ATCos needed to sort 48 cards of a German Doppelkopf deck into six decks (Aces, Kings, Queens, Jacks, Tens, and Nines). In the beginning, all 48 cards are on one stack, with the picture side of the cards looking downwards. Each card needed to be turned around in a single move with just one hand to put it onto the correct of the six decks. After sorting, ATCos should name one to four randomly missing cards that the supervisor took out of the 48 cards deck prior to starting sorting. If there was an error in naming the missing cards, e.g., not all missing cards are named, ATCos must try again until all missing cards are named correctly. The time measurement in seconds started when the deck of 48 cards was put next to the electronic flight strip display. The time measurement ended when all missing cards were named correctly. Sorting cards were trained once in each of the thirty minutes training runs. Card sorting in the baseline and solution runs started after 10 min (for at least 15 min or at least three rounds) and again after 40 min (for at least 13 min or at least three rounds). Those time frames comprised higher traffic density to measure any difference in workload through ASR support.

The results are shown in Figure 21. ATCos finished their secondary task 8% slower in baseline runs when not being supported by ASR (395 s vs. 364 s with a standard deviation of 305 s and 262 s). This difference was 9% for ON and 7% for ACG ATCos. When compensating sequence effects with the SECT technique, ATCos were even 9% slower in baseline runs compared to solution runs. After compensating sequence effects, the α was 0.24 to reject the hypothesis that ABSR support does not reduce the workload of ATCos.



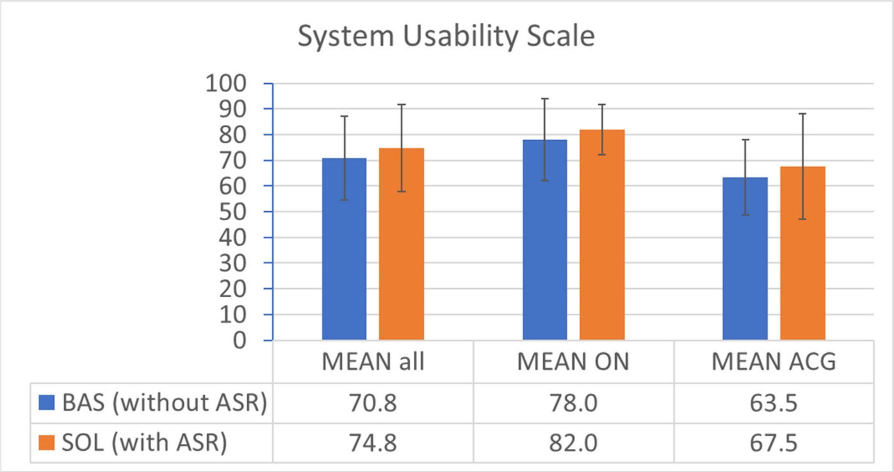
	MEAN all	MEAN ON	MEAN ACG
🟦 Sort-Time [s] BAS (without ASR)	395	588	201
🟧 Sort-Time [s] SOL (with ASR)	364	541	188

Figure 21. ATCo performance in the secondary task (card sorting).

When translating the timing results into workload, again, ON ATCos experienced a higher workload level (around 9 min sorting average) than ACG ATCos (around 3 min sorting average with more task repetitions than ON ATCos), but workload in solution condition seems to be lower than in baseline regarding the secondary task of card sorting. Additionally, the secondary task showed a great learning curve, i.e., ATCos were almost 19% slower in sorting the cards in their first simulation run compared to their second simulation run (baseline and solution alternated).

3.10. System Usability (Post-Run)

The post-run questionnaire contained the ten statements of the System Usability Scale (SUS), as listed in Appendix C.6. The results are shown in Figure 22 (one ATCo did not answer one of his ten statements both in baseline (without ASR) and solution (with ASR) condition. Therefore, the scale mean "3" ((5-1)/2) was chosen as a replacement to not heavily influence the overall result). ABSR support increases the system usability due to SUS ratings ($\alpha = 0.16$). There were three statements rated in the expected direction with an $\alpha < 0.075$, i.e., ATCos like to use the system, they do not deem it complex, and they hardly need support to use it.



	MEAN all	MEAN ON	MEAN ACG
🟦 BAS (without ASR)	70.8	78.0	63.5
🟧 SOL (with ASR)	74.8	82.0	67.5

Figure 22. Subjective ATCo ratings on system usability.

Considering all ATCos, the SUS score was 4 percent absolute (5.7% relative) higher in the solution condition (SOL) with ABSR support compared to the baseline condition (BAS) without ABSR support. The difference of 4 percent remains when just analyzing the ON score or ACG score independently. However, the score itself is 14.5%, absolutely higher for ON than for ACG. This is probably due to the fact that ON really liked the electronic flight strip display (also in the baseline version), whereas ACG ATCos needed to adapt themselves more to the strip system due to the difference in their daily-life system.

3.11. Debriefing Feedback (Post-Validation)

The debriefing was conducted as a semi-structured interview with some pre-defined questions and some options for further thoughts and inputs. The feedback of ATCos is semantically reported per category in the following subsections—the most important feedback relevant for future usage of ABSR is listed after arrow symbol bullets. However, also the remaining feedback helps to improve future simulation planning, i.e., to know which aspects that are not the core part of the study do influence the subject's experience and study results. For example, the prototypic flight strip system induced a row of effects on how the ABSR output is perceived. The last question outlines further research or usage of ABSR systems.

3.11.1. Study Preparation and Conduction

- Briefing slides via e-mail two weeks before the trials and briefing at DLR was very good;
- All ATCos felt well-trained for the purpose of the validation after one hour of training;
- Simulation pilots performed well;
- Air traffic scenarios were rated to be fine for the study purpose;
- On the one hand, the baseline condition (manual work) was similar to everyday work, so performance might be better, therefore (2 ATCos);
- ➢ On the other hand, ASR in solution condition was good because it supported using a flight strip system that ATCos were not used to.

3.11.2. ABSR Functionality (also Related to Electronic Flight Strip Display)

- ➢ ABSR concept and implementation were found to be good by many ATCos;
- ➢ Checking ABSR output in the flight strip display slows some ATCos because, in the baseline mode, ATCos tick while speaking;
- ➢ Some ATCos judged the speed of ABSR output while speaking as sufficient; two ATCos wanted to have faster output;
- ➢ Non-standard situations should be covered well, i.e., better, by ASR;
- ➢ Speech understanding (annotation process) was good for covering errors in speech recognition (transcription process);
- ➢ Highlighting of callsigns and status icons (in green) and the 10s-highlighting mechanism in electronic flight strips were fine for all ATCos;
- ➢ When ASR worked fine, a tendency to over-rely on automatism existed;
- ➢ In case of non-recognition, a double effort to manually recognize the error and correct it compared to pen input (2 ATCos);
- ABSR output in outside view (complete transcription and annotation in solution condition) was just checked for curiosity by all ATCos.

3.11.3. Feedback to Colleagues Not having participated

When I am home in Lithuania/Austria, I tell my colleagues that working with DLR's speech recognition was:

- ➢ Interesting (said by all ON ATCos);
- ➢ Worked pretty well (2 ATCos);
- ➢ Positively surprising (even when speaking fast);
- ➢ Very good even if not being an early adaptor of new technologies and being very safety critical.

3.11.4. Usefulness of ASR

If you would use it tomorrow in your tower controller working position (not multiple remote towers), would ASR help?

- ➢ Yes, that would be great (3);
- ➢ Nothing to be changed to be used tomorrow (1);
- ➢ Great support is possible if some/many aspects are improved (4).

3.11.5. Used Phraseology in Baseline and Solution Runs

Did you think you have spoken differently in baseline and solution conditions?

- ➢ In baseline less carefully spoken because only simulation pilots needed to understand (3 ATCos);
- ➢ Spoken closer to phraseology in solution as being better supported (2 ATCos);
- ➢ Some stated that there was no difference in speaking;
- ➢ "ATCos automatically become more phraseology conform: That is one of the greatest advantages of such a technology."

3.11.6. Flight Strip System (More Related to 'Multiple Remote Tower" than the Core Study Purpose 'ABSR Support')

- Runway bay handling needs to be improved (sorting, highlighting, timing, etc.);
- Drag-and-drop functionality over the borders of flight strip bays for individual planning purposes was needed;
- Handling training flights (touch-and-go/low approach) that do not switch from an arrival flight strip to a departure flight strip were slightly difficult;
- Strip handling for aircraft crossing the control zone is difficult with status options;
- Visual flagging of strips (left/right) would be beneficial;
- Hide some non-frequent status icons;
- "Takeoff" status should include "lineup"-status (if not given explicitly);
- A combination of the selection of taxi status and taxiway would be easier;
- Suggestions for colors, e.g., ground vehicles, consistency with other systems;
- One ATCo loved the flight strip system; the majority of ATCos were ok with it;
- Many ATCos liked the fade-away functionality of flight strips;
- The portion of gazes at the three areas 'flight strip display,' 'outside view,' and 'radar view': too much on flight strips and too few on outside view where one can hardly identify small objects.

3.11.7. Further Applications/Ideas/Things to Be Changed?

- ➢ Callsign highlighting in flight strip display from pilot utterance would help to identify the communication partner;
- ➢ Speech log for pilot utterances (especially in emergency situations) anywhere on the controller screen;
- Connect ABSR output with:
    a. Radar information for automatic setting of landed/departed status;
    b. Lighting system to turn off stop bar lights in case of lineup clearance;
    c. Follow the greens for correct lighting;
    d. Airport phone conversation to automatically extract and include stand numbers given by the airport;
    e. Safety net functionality for dedicated aspects in case of good error rates, e.g., readback error detection;
    f. Transcription for incident analysis and searching for callsigns; other analysis on transcribed data;
    g. Great technology for on-the-job training.

4. Discussion on Major Study Results

The results on mental workload, situation awareness, satisfaction, acceptance, trust, and usability are ambivalent. The subjective post-run ratings on NASA-TLX, Bedford workload scale, and AIM-s, when interpreted as a whole, indicate a worse performance in solution runs with ABSR support compared to baseline runs without ABSR support.

However, the subjective post-validation rating on ABSR support for workload, the self-assessed workload ratings during the simulation runs by ISA, and the performance measurement of the objective secondary task indicate that ABSR support positively influences ATCo workload.

There might also be an influence through the usage of standardized and tailor-made questionnaires. The general low to medium workload level, as rated with roughly two on average on the five-point instantaneous self-assessment of workload scale, causes that it is hard to unambiguously measure a workload effect. Hence, the necessity for controller support functionalities might also be low in such a multiple remote tower environment.

The complexity of the task came with supervising three airports remotely at the same time with a working position the ATCos had not seen before. This could be the reason why especially the callsign highlighting was well-acknowledged by ATCos in order to reduce search times at the different displays. A workload reduction, especially in low workload conditions, is not always beneficial. Hence, it is also a success if the mental workload of ATCos is balanced at a medium level without peaks and boredom.

Similarly, the post-run rating on situation awareness (SASHA) indicates a negative influence, whereas the two rated post-validation statements on situation awareness at an acceptable level with ABSR support have answer values in the most positive scale third. Very similar effects were also seen for satisfaction, acceptance, and trust when comparing post-run ratings with overall post-validation answers.

The usability ratings (post-run and post-validation) seem to all indicate favor for ABSR support. The score of the system usability scale was four points better for the solution (with ABSR support) compared to the baseline (without ABSR support). A total of 80% of ATCos (with 8/10 or more points on the questionnaire scale) stated that they would accept such an ABSR system in their usual working position and that they could apply operating methods in a timely manner. Though, a row of adjustments were encouraged by ATCos, i.e., to make ABSR also reliable under non-nominal conditions where the pressure on ATCos is already high. The need for changes was rated very inhomogeneous by the different ATCos, i.e., some had already seen good support with the prototype's current technology readiness level, and others wanted to increase the number of covered situations and examples.

However, the comparison of a further objective measure with a subjective measurement again shows the ambivalence of some ATCo ratings: While ACG ATCos rated the perceived callsign recognition quality with 1.8 points higher than ON ATCos on a 10-point scale and the perceived command recognition quality with 1.6 points lower than ON ATCos such an effect cannot be seen in the online recognition rates where the callsign recognition rate and the command recognition rate in solution runs of ON ATCos was 2% and 10% (consistently both) better than of ACG ATCos, respectively.

Our study results based on text-to-concept analysis also revealed a potential safety issue for multiple remote towers: Even if ATCos were asked to utter the name of their current transmission station in each radio transmission, the station name, e.g., *vilnius tower*, was missing in every fifth utterance. This might confuse listening to cockpit crews being on or flying to one of the other two airports.

The subjective feedback through questionnaires etc., and the results from objective measurements at least are not consistent or even contradictory. This is a hint that ABSR's performance does not match with ATCos expectations. Objectively a word error rate of 10% with a command recognition rate of 80% might be sufficient to already have positive effects on workload. The ATCos are then, however, not trusting the system, which will be a showstopper. Objective improvements are not enough. ATCos also need to be convinced by their subjective feelings. Previous validation trials for Frankfurt airport to support apron controllers by ABSR to reduce workload for pre-filling electronic flight strips [12] and for Vienna approach controllers [41] indicate that a command recognition rate greater than 90% is needed.

5. Conclusions and Outlook

5.1. Conclusions on ABSR Study in Multiple Remote Tower Environments

Human-in-the-loop trials were conducted with five Austrian and five Lithuanian air traffic controllers (ATCos) to validate whether an assistant-based speech recognition (ABSR) system can support air traffic controllers in a multiple remote tower environment. In baseline runs, controllers needed to manually maintain electronic flight strips without ABSR support, whereas in solution runs, they were supported by ABSR through callsign highlighting and automatically inputting recognized commands from ATCo utterances into electronic flight strips.

This study recorded a huge amount of data with results analyses that are shared with other researchers by this article. The chosen "within-subject design" [46] assessed the dependent variables mental workload, situation awareness, satisfaction, acceptance, trust, and usability with the independent variable "availability of ABSR support". Further qualitative feedback was gathered on ABSR accuracy, technical functionality, and operating methods. Although a very small number of training data of 3.6 and 0.9 h, respectively, was available for the adaption of the ABSR models to Lithuanian and Austrian tower phraseology, some results show statistical significance and are in line with findings of earlier ABSR projects from an approach environment [8]. The text-to-concept accuracy of the speech understanding module performed well, i.e., correcting wrong word recognition by context information. A callsign recognition rate of 94.2% and a command recognition rate of 82.9% were achieved, although each 10th word was wrongly recognized due to the observed word error rate of 9.8%. Given an independent distribution of word errors and an average callsign length of five words, a word error rate of 10% would result in a callsign recognition rate of below 60%, i.e., $(1–0.1)^5$. For an average command length of six words, including values, qualifiers, and conditions plus the five words for the callsign, the expected command recognition rate would be below 35%, i.e., $(1–0.1)^{11}$. These theoretical values were outperformed by our speech understanding module (command recognition) by using context information.

The study results on human factors comprised subjective ratings on mental workload, situation awareness, satisfaction, acceptance, trust, and usability via standardized and tailor-made questionnaires, the self-assessed workload during simulation runs, and an objective method to assess workload based on a secondary task.

The analysis results on the dependent variables were ambivalent. The reasons are the small number of study subjects, the prototype of a non-operational user interface, and the low workload resulting from low to medium traffic in the multiple remote tower environment of the chosen airports. A positive influence on workload was found with the self-assessed workload ratings during the simulation runs and the performance in the secondary task as a more objective measurement during simulation runs. Future validation trials involving ATCos should focus more on objective or live measurements than on retrospective ratings.

Our study results with ATCos reporting on benefits and drawbacks raise detailed awareness and give recommendations on which aspects of automatic speech recognition and understanding for a multiple remote tower environment are already solved and which aspects require deeper research to go beyond the now achieved technology readiness level four.

The speech-to-text performance is a prerequisite to enable good text-to-concept performance. An error analysis after the validation trials revealed processor overload as a factor in decreasing our speech-to-text performance. When applying our command extraction on offline speech-to-text analysis results having a word error rate of 4.4%, we achieve a command recognition rate of 91.8% and a callsign recognition rate of 98.2%. The data analysis showed that ABSR support has a statistically significant positive effect on the usage of ICAO phraseology: The above-reported solution runs have higher command recognition rates than baseline runs because ATCos obtain better support if recognition rates are higher. If ATCos are sticking closer to ICAO phraseology just by the pure presence of an ABSR system, that will already be a safety feature. Some ATCos, i.e., the human operators that would use the operating system later on, highlighted that such an ABSR system would be a great support in their working position.

5.2. Outlook on Future Work

The amount of training data must be further increased, given representative samples. Furthermore, a large amount of data must be recorded from operations rooms (not from labs) because this is the operational environment. The European-wide agreed ontology for the annotation of ATC utterances was successfully used and enhanced in this study and should be further exploited or standardized. The continuous mutual enhancements of the ontology for en-route/oceanic, approach, tower, and apron traffic within the ASR projects HAAWAII (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration (HAAWAII), Homepage: https://www.haawaii.de (accessed on 4 April 2023)) (as the successor of MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance (MALORCA), Homepage: https://www.malorca-project.de (accessed on 4 April 2023)), and STARFiSH (Safety and Artificial Intelligence Speech Recognition (STARFiSH), Homepage: https://www.dlr.de/fl/desktopdefault.aspx/tabid-1149/1737_read-74905/ (accessed on 4 April 2023)) tremendously build a base for interoperability of systems. Hence, following ASR activities can build on strong shoulders and reuse the achieved good results and methods of such ABSR projects.

For the specific case of electronic flight strips, eye tracking technology could be of further help to make sure that ATCos checked the ABSR output [47]. This technology could also be used to assess the time to recognize and correct an ABSR error (Times to correct ABSR errors in an ATM environment have been investigated in "Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload" of Helmke et al. presented at the 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023), Savannah, GA, USA, 5–9 June 2023). Furthermore, the support through callsign highlighting when recognized from pilot utterances should be investigated and potentially feed attention guidance systems at the controller working position. To summarize, the validation trials have shown the potential of using the output of an ABSR system in the multiple remote tower environment and revealed aspects to be considered when moving forward to higher technology readiness levels.

Author Contributions: Conceptualization, O.O.; Methodology, H.H., S.S. (Shruthi Shetty), M.K. and H.E.; Software, O.O., S.S. (Shruthi Shetty), M.K., H.E., S.S. (Saeed Sarfjoo) and P.M.; Validation, O.O., H.H. and S.S.-M.; Formal analysis, O.O. and M.M.; Investigation, O.O.; Resources, Š.M., T.P., H.U. and A.Č.; Data curation, O.O.; Writing—original draft, O.O.; Writing—review and editing, H.H., S.S. (Shruthi Shetty), M.K., H.E., S.S.-M., S.S. (Saeed Sarfjoo), P.M., Š.M., T.P., H.U., M.M. and A.Č.; Supervision, O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation program under grant agreement No 874470.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We like to thank the ten air traffic controllers of ACG and ON who participated in the validation study in spring 2022 under a strict hygienic protocol due to COVID-19 as well as roughly one dozen of air traffic controllers from ON, ACG, and the Polish Air Navigation Services Agency (PANSA) who contributed to the recording of training data (surveillance data and speech utterances) in earlier multiple remote tower trials at DLR Braunschweig. Further, we thank the air traffic management simulation department and the simulation pilots at DLR's Institute of Flight Guidance as well as our colleague Lennard Nöhren for supporting software preparation of simulation environment and conduction of a row of multiple remote tower human-in-the-loop validation studies in the course of our automatic speech recognition activities.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of this study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

ABSR	Assistant Based Speech Recognition
ACG	Austro Control
AIM-s	Assessing the Impact on Mental Workload
ASR	Automatic Speech Recognition
ATC	Air Traffic Control
ATCo	Air Traffic Controller
ATIS	Automatic Terminal Information Service
ATM	Air Traffic Management
BAS	Baseline Runs
CARS	Controller Acceptance Rating Scale
CoCoLoToCoCo	Controller Command Logging Tool for Context Comparison
CPU	Central Processing Unit
CWP	Controller Working Position
Del	Deletions
DLR	German Aerospace Center
DTT	Digital Tower Technologies
EASA	European Union Aviation Safety Agency
EFS	Electronic Flight Strip System
EUROCAE	European Organization for Civil Aviation Equipment
HMI	Human Machine Interface
ICAO	International Civil Aviation Organization
Ins	Insertions
ISA	Instantaneous Self-Assessment
LevenDist	Levenshtein Distance
NASA-TLX	National Aeronautics and Space Administration Task Load Index
Off	Offline (analysis of audio files after the simulation runs)
ON	Oro Navigacija
Onl	Online (analysis as experienced by ATCos during simulation runs)
OW	Overall Weighted Workload
SASHA	Situation Awareness for SHAPE
SATI	SHAPE Automation Trust Index
SD	Standard Deviation
SECT	Sequence Effect Compensation Technique
SHAPE	Solutions for Human Automation Partnerships in European ATM
SOL	Solution Runs
Subs	Substitutions
SUS	System Usability Scale
TWR	Tower
WER	Word Error Rate

Appendix A. Speech-To-Text Accuracy

The following tables in this Appendix A show the speech recognition performance on the word level, i.e., the word error rates (WER). The first row must be read like this; 1,944 words were spoken. Ninety-seven errors occurred, i.e., 43 words were substituted by another word, 38 words were not recognized at all (deleted), and 16 words were inserted, i.e., not said, but a word was recognized. This results in a word error rate of 5.1% (97/1944).

Table A1. Speech-To-Text performance for offline recognition on audio files (Off).

Sample	# Words	LevenDist	# Subs	# Del	# Ins	% WER
MEAN all	1944	97	43	38	16	5.1
MEAN ON	1966	94	38	36	20	5.0
MEAN ACG	1921	99	48	39	13	5.1
MEAN w/o outlier run	1971	90	40	34	16	4.5
MEAN BAS all	1902	104	46	43	15	5.7
MEAN BAS ON	1891	100	41	43	16	5.7
MEAN BAS ACG	1913	109	51	44	14	5.7
MEAN BAS w/o outlier run	1961	98	44	39	15	5.0
MEAN SOL all	1985	89	40	32	17	4.4
MEAN SOL ON	2041	88	36	30	23	4.3
MEAN SOL ACG	1929	90	44	34	11	4.6
MEAN SOL w/o outlier run	1980	81	36	28	17	4.1

Rows are shaded, when containing all ATCos, i.e., both from ACG and ON.

Table A2. Speech-To-Text accuracy for real-time online recognition from voice stream (Onl).

Sample	# Words	LevenDist	# Subs	# Del	# Ins	% WER
MEAN all	1936	245	46	175	24	13.6
MEAN ON	1954	199	38	140	21	11.9
MEAN ACG	1918	290	54	209	27	15.3
MEAN w/o outlier run	1967	212	41	152	19	11.0
MEAN BAS all	1891	300	54	219	27	17.4
MEAN BAS ON	1871	261	42	196	23	17.1
MEAN BAS ACG	1911	339	66	241	32	17.8
MEAN BAS w/o outlier run	1959	254	50	181	23	13.2
MEAN SOL all	1980	189	38	131	21	9.8
MEAN SOL ON	2037	136	34	83	19	6.8
MEAN SOL ACG	1924	242	42	178	22	12.8
MEAN SOL w/o outlier run	1976	171	32	123	15	8.9

Rows are shaded, when containing all ATCos, i.e., both from ACG and ON.

The following two tables show the frequency of certain words appearing in the gold transcriptions and the number of unique words needed to reach a certain portion of all words in the gold transcriptions, respectively.

Table A3. 1-grams of gold transcriptions.

Rank	Word	Count	Portion
1	one	2393	6.43%
2	zero	1479	3.97%
3	tower	1473	3.96%
4	three	1356	3.64%
5	runway	1154	3.10%
6	five	1085	2.91%
7	seven	925	2.48%
8	two	923	2.48%
9	four	898	2.41%
10	to	888	2.38%
11	cleared	808	2.17%
12	right	795	2.13%
13	vilnius	747	2.01%
14	eight	721	1.94%
15	nine	720	1.93%

Table A3. *Cont.*

Rank	Word	Count	Portion
16	via	601	1.61%
17	air	571	1.53%
18	degrees	556	1.49%
19	and	539	1.45%
20	knots	531	1.43%
21	bravo	465	1.25%
22	wind	456	1.22%
23	alfa	409	1.10%
24	taxi	408	1.10%
25	kaunas	390	1.05%
	others	15,947	42.8%
1-505	SUM	37,238	100%

Table A4. The number of different words needed to reach a certain portion of all uttered words.

Count	Portion
61	80%
101	90%
145	95%
283	99%
505	100%

Appendix B. Text-To-Concept Accuracy

The following tables lists the relative frequency of supported air traffic control command types from the gold annotations.

Table A5. Percentage of used command types in gold annotations occurring more often than 1% (7560 commands in total).

Command Type	Portion of All Commands
STATION	20.2%
INFORMATION WINDSPEED	7.5%
INFORMATION WINDDIRECTION	7.5%
TAXI TO	6.4%
GREETING	5.6%
TAXI VIA	4.8%
AFFIRM	4.0%
INFORMATION QNH	3.3%
CLEARED VIA	2.9%
STARTUP	2.9%
CLEARED TO	2.9%
CLEARED TAKEOFF	2.8%
FAREWELL	2.8%
CLEARED LANDING	2.8%
SQUAWK	2.8%
LINEUP	2.4%
REPORT	1.5%
PUSHBACK	1.4%
INFORMATION ACTIVE_RWY	1.4%
NO_CONCEPT	1.4%
REPORT_MISCELLANEOUS	1.4%
VACATE VIA	1.2%
CLEARED TOUCH_GO	1.1%
others	8.9%

The following six tables present the speech understanding performance per study subject group and per command type group for gold, offline, and online transcriptions, respectively.

Table A6. Text-to-concept quality for gold transcriptions (assumed to be 100% correct).

Gold Transcription	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	95.9%	2.4%	99.8%	0.2%	13.3%	100.0%
ON ATCos ALL	97.1%	1.5%	99.7%	0.2%	12.5%	49.9%
ACG ATCos ALL	94.8%	3.2%	99.9%	0.1%	14.2%	50.1%
ATCos ALL w/o outlier run	95.8%	2.5%	99.8%	0.2%	13.2%	91.8%
all ATCos BAS	95.9%	2.4%	99.7%	0.3%	13.8%	49.0%
ON ATCos BAS	97.6%	1.3%	99.7%	0.3%	13.0%	24.1%
ACG ATCos BAS	94.1%	3.5%	99.8%	0.2%	14.7%	24.8%
all ATCos SOL	96.0%	2.3%	99.8%	0.1%	12.8%	51.0%
ON ATCos SOL	96.6%	1.8%	99.7%	0.2%	12.0%	25.8%
ACG ATCos SOL	95.4%	2.9%	100.0%	0.0%	13.7%	25.3%

Table A7. Text-to-concept quality for gold transcriptions (assumed to be 100% correct) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	34	97.3%
EFS	21	97.4%
Status	18	96.7%
Outside	3	96.0%
Hypo-EFS	4	99.2%

Table A8. Text-to-concept quality for Off transcriptions (current best word error rates of automatic speech-to-text with callsign boosting on audio files).

Offline	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	91.4%	4.5%	98.4%	0.9%	14.0%	100.0%
ON ATCos ALL	92.7%	3.9%	98.6%	0.6%	12.8%	49.9%
ACG ATCos ALL	90.1%	5.1%	98.2%	1.2%	15.2%	50.1%
ATCos ALL w/o outlier run	91.7%	4.4%	98.7%	0.9%	13.9%	91.8%
all ATCos BAS	91.0%	4.6%	98.6%	0.8%	14.5%	49.0%
ON ATCos BAS	92.8%	3.6%	99.0%	0.3%	13.2%	24.1%
ACG ATCos BAS	89.3%	5.5%	98.1%	1.2%	15.8%	24.8%
all ATCos SOL	91.8%	4.5%	98.2%	1.1%	13.6%	51.0%
ON ATCos SOL	92.7%	4.1%	98.1%	0.9%	12.6%	25.8%
ACG ATCos SOL	90.9%	4.8%	98.3%	1.2%	14.6%	25.3%

Table A9. Text-to-concept quality for Off transcriptions (current best word error rates of automatic speech-to-text with callsign boosting on audio files) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	31	92.4%
EFS	21	93.4%
Status	18	92.7%
Outside	3	90.5%
Hypo-EFS	4	96.3%

Table A10. Text-to-concept quality for Onl transcriptions (automatic speech-to-text with callsign boosting from continuous stream).

Online	Command Recognition Rate	Command Error Rate	Callsign Recognition Rate	Callsign Error Rate	Unknown Classified Rate	Amount of Data
all ATCos ALL	79.4%	7.0%	91.7%	3.1%	15.4%	100.0%
ON ATCos ALL	84.2%	5.5%	92.1%	2.4%	13.8%	49.9%
ACG ATCos ALL	74.6%	8.6%	91.3%	3.9%	17.0%	50.1%
ATCos ALL w/o outlier run	81.2%	6.6%	94.0%	2.5%	14.9%	91.8%
all ATCos BAS	75.7%	7.5%	89.1%	3.8%	16.2%	49.0%
ON ATCos BAS	80.1%	5.6%	88.9%	2.8%	14.6%	24.1%
ACG ATCos BAS	71.4%	9.3%	89.3%	4.8%	17.9%	24.8%
all ATCos SOL	82.9%	6.6%	94.2%	2.4%	14.5%	51.0%
ON ATCos SOL	88.0%	5.4%	95.2%	2.0%	13.2%	25.8%
ACG ATCos SOL	77.7%	7.9%	93.2%	2.9%	16.1%	25.3%

Table A11. Text-to-concept quality for Onl transcriptions (automatic speech-to-text with callsign boosting from continuous stream) per command type groups.

Command Type Group	# Command Types	Command Recognition Rate
Relevant	31	80.7%
EFS	21	79.2%
Status	18	80.0%
Outside	3	81.0%
Hypo-EFS	4	87.2%

Appendix C. Questions and Statements of Questionnaires

The following full-text questions and statements were contained within the listed post-run questionnaires:

Appendix C.1. Statement and Answer Scale from CARS

The color coding shows worse answers in red and good answers in green.

"Please read the descriptors and score your overall level of user acceptance experienced during the run. Please check the appropriate number."

▪	Improvement mandatory. Safe operation could not be maintained.
▪	Major Deficiencies. Safety not compromised, but system is barely controllable and only with extreme controller compensation.
▪	Major Deficiencies. Safety not compromised but system is marginally controllable. Considerable compensation is needed by the controller.
▪	Major Deficiencies. System is controllable. Some compensation is needed to maintain safe operations.
▪	Very Objectionable Deficiencies. Maintaining adequate performance requires extensive controller compensation.
▪	Moderately Objectionable Deficiencies. Considerable controller compensation to achieve adequate performance.
▪	Minor but Annoying Deficiencies. Desired performance requires moderate controller compensation.
▪	Mildly unpleasant Deficiencies. System is acceptable and minimal compensation is needed to meet desired performance.
▪	Negligible Deficiencies. System is acceptable and compensation is not a factor to achieve desired performance.
▪	Deficiencies are rare. System is acceptable and controller does not have to compensate to achieve desired performance.

Appendix C.2. Statements from SATI Questionnaire

1. In the previous working period, I felt that the system was useful. [USEFUL]
2. In the previous working period, I felt that the system was reliable. [RELIABLE]
3. In the previous working period, I felt that the system worked accurately. [ACCURACY]
4. In the previous working period, I felt that the system was understandable. [UNDERSTAND]
5. In the previous working period, I felt that the system worked robustly (in difficult situations, with invalid inputs, etc.). [ROBUST]
6. In the previous working period, I felt that I was confident when working with the system. [CONFIDENT]

Appendix C.3. Statements from SASHA Questionnaire

1. In the previous run, I was ahead of the traffic. [AHEAD]
2. In the previous run, I started to focus on a single problem or a specific aircraft. [FOCUS]
3. In the previous run, there was a risk of forgetting something important (such as inputting the spoken command values into the labels). [FORGET]
4. In the previous run I was able to plan and organize my work as wanted. [PLAN]
5. In the previous run I was surprised by an event I did not expect (such as an aircraft call). [SURPRISE]
6. In the previous run I had to search for an item of information. [SEARCH]

Appendix C.4. Questions from NASA-TLX Questionnaire

1. How mentally demanding was the task? [Mental Demand, MD]
2. How physically demanding was the task? [Physical Demand, PD]
3. How hurried or rushed was the pace of the task? [Temporal Demand, TD]
4. How successful were you in accomplishing what you were asked to do? [Operational Performance, OP]
5. How hard did you have to work to accomplish your level of performance? [Effort, EF]
6. How insecure, discouraged, irritated, stressed, and annoyed were you? [Frustration, FR]

Furthermore, the 15 pairwise comparisons of workload contributing factors have been analyzed. When looking at the subscores for all six NASA-TLX dimensions, half of them (three) were rated equal or better in SOL compared to BAS (PD, EF, FR), and the other half was rated vice versa (MD, TD, OP). In general, physical demand (PD, 3.3%) was rated as being a less important contributor to workload, and mental demand (MD, 23.3%) was the most important contributor to workload. The other four dimensions were rather equally important contributors to the overall workload (TD 22%, OP 18%, EF 16.7%, FR 16.7%). The horizontal axis in Figure A1 shows the weight; the area shows the contribution of this very dimension to the OW of BAS and SOL conditions, respectively.

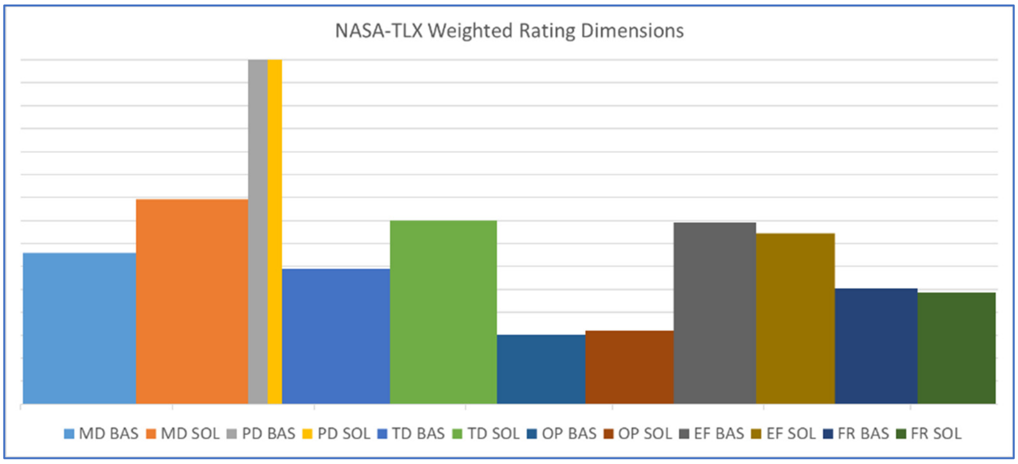


Figure A1. ATCo ratings on NASA-TLX (Weighted Workload Factors).

Appendix C.5. Questions from AIM-s Questionnaire

1. In the previous run, how much effort did it take to prioritize tasks? [PRIOT]
2. In the previous run, how much effort did it take to identify potential conflicts? [IDENT]
3. In the previous run, how much effort did it take to scan radar or any display? [SCRD]
4. In the previous run, how much effort did it take to evaluate conflict resolution options against the traffic situation and conditions? [EVAL]
5. In the previous run, how much effort did it take to anticipate the future traffic situation? [ANTIC]
6. In the previous run, how much effort did it take to recognize a mismatch of available data with the traffic picture? [RECOG]
7. In the previous run, how much effort did it take to issue timely commands? [TIMELY]
8. In the previous run, how much effort did it take to evaluate the consequences of a plan? [PLAN]
9. In the previous run, how much effort did it take to manage flight data information? [MANG]
10. In the previous run, how much effort did it take to share information with team members? [SHARE]
11. In the previous run, how much effort did it take to recall necessary information? [RECL]
12. In the previous run, how much effort did it take to anticipate team members' needs? [TMN]
13. In the previous run, how much effort did it take to prioritize requests? [PRIRQ]
14. In the previous run, how much effort did it take to scan flight progress data? [SCFP]
15. In the previous run, how much effort did it take to access relevant aircraft or flight information? [ACCD]
16. In the previous run, how much effort did it take to gather and interpret information? [GETI]

Appendix C.6. Statements from SUS Questionnaire

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.

4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Appendix D. Validation Setup Details

The left and right sides of the outside view areas presented current meteorological data as relevant for aircraft takeoff and landing (see Figure A2), i.e., wind speed in knots (here 10) and wind direction with an additional red arrow (here 070°) according to the runway orientation (grey rectangle), the active runway name (here 05), the airport International Civil Aviation Organization (ICAO) code (EYKA), the QNH (here 1001), the visibility conditions (here 9999, i.e., no visibility restrictions), and cloud information (in green circles).

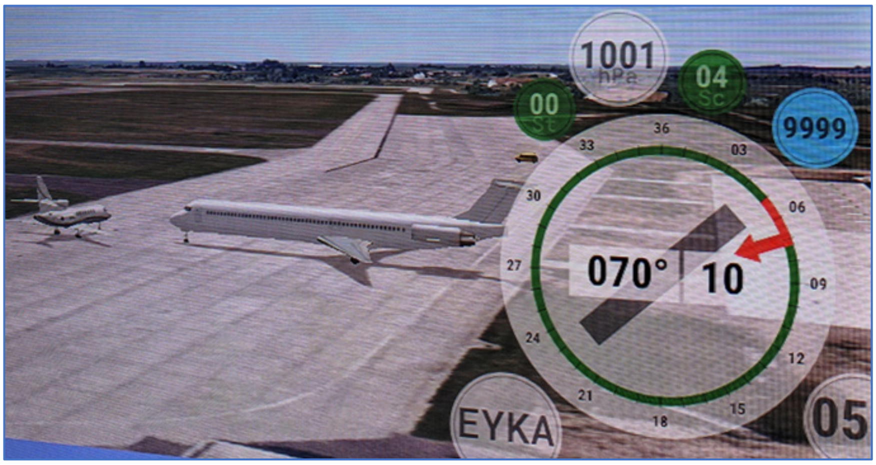


Figure A2. Remote tower outside view with a small aircraft passing a parking aircraft on the apron and meteorological information in and around the compass rose on the right.

An adjacent laboratory room accommodated three simulation pilot workstations. Each workstation consisted of a monitor to visualize the simulation pilot interface (see Figure A3) for one of the three simulated airports, a keyboard, and a mouse.

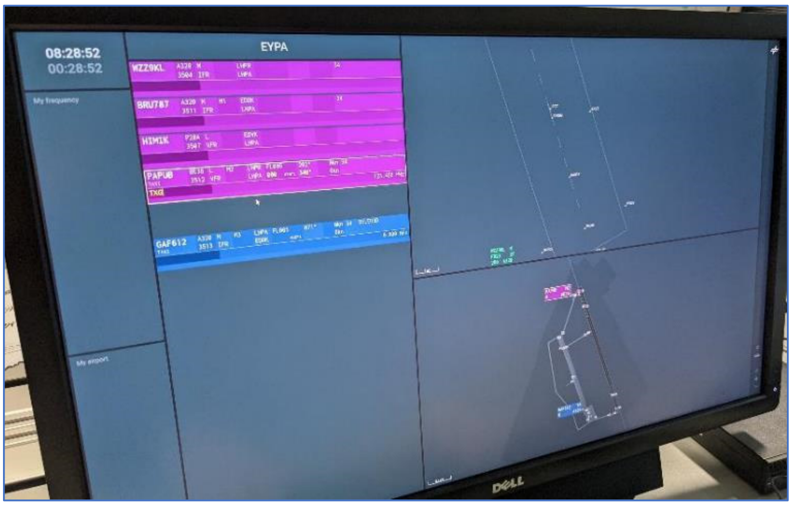


Figure A3. Simulation pilot interface for a simulated airport with time, pseudo flight strips for arrival and departure traffic, and radar views for airport surface and surrounding.

References

1. Lin, Y. Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application. *Aerospace* **2021**, *8*, 65. [CrossRef]
2. Schäfer, D. Context-Sensitive Speech Recognition in the Air Traffic Control Simulation. Ph.D. Thesis, The University of Armed Forces, Munich, Germany, 2001.
3. Updegrove, J.A.; Jafer, S. Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy. *Aerospace* **2017**, *4*, 50. [CrossRef]
4. Cordero, J.M.; Rodriguez, N.; de Pablo, J.M.; Dorado, M. Automated speech recognition in controller communications applied to workload measurement. In Proceedings of the 3rd SESAR Innovation Days, Stockholm, Sweden, 26–28 November 2013.
5. Cordero, J.M.; Dorado, M.; de Pablo, J.M. Automated speech recognition in ATC environment. In Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems, London, UK, 29–31 May 2012; pp. 46–53.
6. Kleinert, M.; Helmke, H.; Shetty, S.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Harfmann, J. Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In Proceedings of the IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), Virtual, 3–7 October 2021. [CrossRef]
7. Helmke, H.; Rataj, J.; Mühlhausen, T.; Ohneiser, O.; Ehr, H.; Kleinert, M.; Oualil, Y.; Schulder, M. Assistant-Based Speech Recognition for ATM Applications. In Proceedings of the 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 23–26 June 2015.
8. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.
9. Helmke, H.; Slotty, M.; Poiger, M.; Herrer, D.F.; Ohneiser, O.; Vink, N.; Cerna, A.; Hartikainen, P.; Josefsson, B.; Langr, D.; et al. Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04. In Proceedings of the IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, UK, 23–27 September 2018. [CrossRef]
10. Helmke, H.; Ondřej, K.; Shetty, S.; Arilíusson, H.; Simiganoschi, T.S.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Zuluaga-Gomez, J.-P.; Smrz, P. Readback Error Detection by Automatic Speech Recognition and Understanding—Results of HAAWAII project for Isavia's Enroute Airspace. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
11. Chen, S.; Kopald, H.D.; Elessawy, A.; Levonian, Z.; Tarakan, R.M. Speech inputs to surface safety logic systems. In Proceedings of the IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 13–17 September 2015. [CrossRef]
12. Kleinert, M.; Shetty, S.; Helmke, H.; Ohneiser, O.; Wiese, H.; Maier, M.; Schacht, S.; Nigmatulina, I.; Sarfjoo, S.S.; Motlicek, P. Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
13. Ohneiser, O.; Helmke, H.; Kleinert, M.; Siol, G.; Ehr, H.; Hobein, S.; Predescu, A.-V.; Bauer, J. Tower Controller Command Prediction for Future Speech Recognition Applications. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
14. Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Murauskas, Š.; Pagirys, T. Prediction and extraction of tower controller commands for speech recognition applications. *J. Air Transp. Manag.* **2021**, *95*, 102089. [CrossRef]
15. Badrinath, S.; Balakrishnan, H. Automatic Speech Recognition for Air Traffic Control Communications. *Transp. Res. Rec.* **2021**, *2676*, 798–810. [CrossRef]
16. Pellegrini, T.; Farinas, J.; Delpech, E.; Lancelot, F. The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019. [CrossRef]
17. García, R.; Albarrán, J.; Fabio, A.; Celorrio, F.; Pinto de Oliveira, C.; Bárcena, C. Automatic Flight Callsign Identification on a Controller Working Position: Real-Time Simulation and Analysis of Operational Recordings. *Aerospace* **2023**, *10*, 433. [CrossRef]
18. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
19. Kleinert, M.; Helmke, H.; Moos, S.; Hlousek, P.; Windisch, C.; Ohneiser, O.; Ehr, H.; Labreuil, A. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
20. Chen, S.; Kopald, H.D.; Chong, R.; Wei, Y.; Levonian, Z. Read back error detection using automatic speech recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
21. Fürstenau, N.; Jakobi, J.; Papenfuss, A. Introduction: Basics, History, and Overview. In *Virtual and Remote Control Tower Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 3–22. [CrossRef]
22. Möhlenbrink, C.; Papenfuß, A. Eye-data metrics to characterize tower controllers' visual attention in a multiple remote tower exercise. In Proceedings of the 6th International Conference on Research in Air Transportation (ICRAT2014), Istanbul, Turkey, 26–30 May 2014.
23. Papenfuss, A.; Friedrich, M. Head Up Only—A design concept to enable multiple remote tower operations. In Proceedings of the IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016.

24. Fürstenau, N.; Papenfuss, A. Model Based Analysis of Subjective Mental Workload During Multiple Remote Tower Human-In-The-Loop Simulations. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 293–342. [CrossRef]
25. Hamann, A.; Jakobi, J. Changing of the Guards: The Impact of Handover Procedures on Human Performance in Multiple Remote Tower Operations. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 343–363. [CrossRef]
26. Friedrich, M.; Timmermann, F.; Jakobi, J. Active supervision in a Remote Tower Center: Rethinking of a new position in the ATC Domain. In Proceedings of the 19th International Conference on Engineering Psychology and Cognitive Ergonomics, EPCE 2022 as part of the 24th HCI International Conference, HCII 2022, Virtual, 26 June—1 July 2022; Springer: Cham, Switzerland, 2022; pp. 265–278. [CrossRef]
27. Li, W.-C.; Kearney, P.; Braithwaite, G. The Certification Processes of Multiple Remote Tower Operations for Single European Sky. In *Virtual and Remote Control Tower. Research Topics in Aerospace*; Fürstenau, N., Ed.; Springer: Cham, Switzerland, 2022; pp. 511–541. [CrossRef]
28. Schier, S.; Rambau, T.; Timmermann, F.; Metz, I.; Stelkens-Kobsch, T.H. Designing the Tower Control Research Environment of the Future. Deutscher Luft- und Raumfahrtkongress, DLRK2013. In Proceedings of the English: German Aerospace Congress, Stuttgart, Germany, 10–12 September 2013.
29. Schier, S.; Manske, P. *VisiTop II—Briefing-Unterlagen*; Section 4.2. DLR-internal report; DLR Institute of Flight Guidance: Braunschweig, Germany, 2015.
30. Ohneiser, O.; Sarfjoo, S.; Helmke, H.; Shetty, S.; Motlicek, P.; Kleinert, M.; Ehr, H.; Murauskas, Š. Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In Proceedings of the InterSpeech, Brno, Czech Republic, 30 August–3 September 2021. [CrossRef]
31. Shetty, S.; Helmke, H.; Kleinert, M.; Ohneiser, O. Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload. In *Human Factors in Transportation, Proceedings of the International Conference on Applied Human Factors and Ergonomics (AHFE2022), New York, NY, USA, 24–28 July 2022*; Plant, K., Praetorius, G., Eds.; AHFE International: New York, NY, USA, 2022; Volume 60. [CrossRef]
32. Jordan, C.S.; Brennen, S.D. *Instantaneous Self-Assessment of Workload Technique (ISA)*; Defence Research Agency: Portsmouth, UK, 1992.
33. Bongo, M.F.; Seva, R.R. Evaluating the Performance-Shaping Factors of Air Traffic Controllers Using Fuzzy DEMATEL and Fuzzy BWM Approach. *Aerospace* **2023**, *10*, 252. [CrossRef]
34. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North Holland Press: Amsterdam, The Netherlands, 1988; p. 198. [CrossRef]
35. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA, USA, 16–20 October 2006; Volume 50, pp. 904–908. [CrossRef]
36. Roscoe, A.H. Assessing Pilot Workload in Flight. In Proceedings of the AGARD Conference Proceedings Flight Test Techniques, Lisbon, Portugal, 2–5 April 1984.
37. Dehn, D.M. Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. *Air Traffic Control Q.* **2008**, *16*, 127–146. [CrossRef]
38. Lee, K.K.; Kerns, K.; Bone, R.; Nickelson, M. The Development and Validation of the Controller Acceptance Rating Scale (CARS): Results of Empirical Research. In Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 3–7 December 2001.
39. Brooke, J. SUS—A quick and dirty usability scale. In *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.
40. Bangor, A.; Kortum, P.T.; Miller, J.T. An empirical evaluation of the system usability scale. *Intl. J. Hum.-Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]
41. Helmke, H.; Kleinert, M.; Ahrenhold, N.; Ehr, H.; Mühlhausen, T.; Ohneiser, O.; Motlicek, P.; Prasad, A.; Zuluaga-Gomez, J. Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload. In Proceedings of the 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023), Savannah, GA, USA, 5–9 June 2023.
42. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
43. Ohneiser, O.; Helmke, H.; Shetty, S.; Kleinert, M.; Ehr, H.; Murauskas, Š.; Pagirys, T.; Balogh, G.; Tønnesen, A.; Kis-Pál, G.; et al. Understanding Tower Controller Communication for Support in Air Traffic Control Displays. In Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5–8 December 2022.
44. ICAO. *ATM (Air Traffic Management): Procedures for Air Navigation Services*; DOC 4444 ATM/501; International Civil Aviation Organization (ICAO): Montréal, QC, Canada, 2007.
45. Helmke, H.; Shetty, S.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Cerna, A.; Windisch, C. Measuring Speech Recognition and Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates. In Proceedings of the 11th SESAR Innovation Days, Virtual, 7–9 December 2021.

46. Charness, G.; Gneezy, U.; Kuhn, M.A. Experimental methods: Between-subject and within-subject design. *J. Econ. Behav. Organ.* **2012**, *81*, 1–8. [CrossRef]
47. Ohneiser, O.; Adamala, J.; Salomea, I.-T. Integrating Eye- and Mouse-Tracking with Assistant Based Speech Recognition for Interaction at Controller Working Positions. *Aerospace* **2021**, *8*, 245. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.