Supporting Simulation Pilots by Automatic Speech Recognition and Understanding

Hartmut Helmke, Shruthi Shetty, Matthias Kleinert, Heiko Ehr
Institute of Flight Guidance, Controller Assistance
German Aerospace Center (DLR)
Braunschweig, Germany
Firstname.Lastname@dlr.de

Stephanie Hobein, Thiago Coelho Jordao
Institute of Flight Guidance, ATM Simulation
German Aerospace Center (DLR)
Braunschweig, Germany
stephanie.hobein@dlr.de; thiago.coelhojordao@dlr.de

***Abstract*—Systems like Alexa, Siri or the Google Assistant that recognize human speech have changed our daily lives during the last decade. Prototypic applications based around speech have since then also found their way into the air traffic management (ATM) domain. Recently pre-filling radar label entries by automatically understanding the air traffic controller to pilot communication has reached the technology readiness level before industrialization (TRL6). DLR is one of the main drivers of such speech-based technologies in the context of ATM. This report addresses an automatic speech recognition and understanding (ASRU) application to support simulation pilots during Human-in-the-Loop experiments. For this purpose, an ASRU system recognizes the verbal clearances of an air traffic controller and forwards the information to the visual interfaces of the human simulation pilots. The pilots confirm the information or make modifications in case of misrecognitions and send it to the simulator for execution. With this approach more than 75% of the commands from the air traffic controller, which the simulation pilot normally has to enter manually, are already recognized by ASRU and the simulation pilot just needs to confirm or modify the ASRU outputs. This dramatically reduces the simulation pilot workload. The remaining 25% of the commands are, however, a challenge. These often contain seldom spoken words related to the airspace, which are relevant in the ATM context. If those commands can also be recognized, more complex simulations are possible with less simulation pilots. This report therefore also presents first results on adjusting ASRU to these seldom spoken words, which are often waypoint names as a part of direct-to clearances, e.g, "mobsa", "ekern".**

***Keywords*—Speech Recognition; Speech Understanding; Human-in-the-Loop Simulation; Workload; Simulation Pilot**

I. Introduction

A. Problem

From August to September 2023, six different sector air traffic controllers (ATCos) performed human-in-the-loop trials in DLR's Air Traffic Management Operations Simulator (ATMOS) [1] within the DIAL project [2]. In March 2024, four ATCos from Vienna participated in the experiment and in April 2024 additional eight ATCos from Germany and France conducted the experiment as well. The experiment was to guide traffic in Maastricht upper airspace, especially in the Celle sector of Germany. In more than 60% of the experiments the ATCos were supported by a digital assistant [3] [4]. In these cases, ATCos and simulation pilots mostly communicated via data link (CPDLC) except in special cases like when there was a sick passenger on board or a CPDLC failure was simulated. So only a limited amount of voice communications happened. However, in the baseline experiments no support of the digital assistant was available. In these cases, the communication between ATCos and simulation pilots was conducted via voice communication only. These runs would therefore normally require more simulation pilots due to the amount of voice communication and the manual effort to control the simulated aircraft. During comparable experiments for the SESAR2020 project PJ.10-W2-96-ASR [5] with very heavy traffic, four simulation pilots were needed to handle the simulated aircraft and voice communication for just one ATCo. This is expensive and ties up a lot of resources. However, a lack of simulation pilots could lead to errors in the simulation due to too much workload and subsequently jeopardize the results. Therefore, the effort of four simulation pilots for just one ATCo was justified.

B. Solution

Now DLR has tried a different approach in the DIAL experiments. The research question was: Are two simulation pilots with only a few hours of training sufficient to make all the requested inputs of one air traffic controller, if the simulation pilot is supported by automatic speech recognition and understanding (ASRU)? In other words: if the given ATCo commands are recognized by an ASRU software and displayed on the simulation pilot interface in real-time, that he/she only needs to accept or make minor modifications to the ASRU output, can this reduce the workload of the simulation pilots to the extent that only two simulation pilots can handle a similar workload that otherwise needed four simulation pilots?

C. Paper Structure

Section II gives an overview of related work starting with ASRU applications and achievements in ATM and continuing ASRU support for simulation pilots during the last three decades. Section III describes the analyzed data. Section IV describes the achieved results with respect to ASRU performance. Section V describes results with respect to simulation pilot performance, i.e. how many of the ASRU errors were detected and corrected by simulation pilots. Section VI concludes the work.

II. Related Work

A. Speech Recognition and Understanding for ATM

Over the last 70 years, advances have led to dramatic improvements in the field of Automatic Speech Recognition (ASR). Juang and Rabiner [6] give an overview of the work until

Supported by DIAL project of DLR, DIAL = Individual and Automated Air Transport in German "**D**er **i**ndividuelle und **a**utomatisierte **L**uftverkehr".

2005. Connolly from FAA [7] was one of the first to describe the steps of using ASR in the ATM domain. In the late 1980s, a first approach to incorporate speech technologies in air traffic control (ATC) training was reported [8] to replace expensive simulation pilots.

Today ASR applications in ATC go beyond basic training scenarios. Modern ASR applications have to recognize experienced controllers with various accents, who more often deviate from standard phraseology. Nowadays, ASR is for example used to obtain more objective feedback concerning controllers' workload [9] or for readback error detection in the US [10] and Europe [11]. A good overview of the integration of ASR in ATC is provided in the paper of Nguyen and Holone [12]. A more technical overview is given by Lin [13]. Radar Label Maintenance supported by ASRU has recently achieved a Technological Readiness Level (TRL) of 6 being validated in DLR's ATMOS simulation environment [5].

Since speech recognition does not include speech understanding, European ATM partners agreed on a so-called ontology to ease understanding of approach controller utterances [14] being extended to apron controller utterances in the STARFiSH project [15] and even more important to pilot transmissions [16]. The ATCo transmissions "*speed bird eight five alfa descend flight level three two zero*" and "*eight five alfa down level three two zero*" mean the same on semantic level and, therefore, are mapped to the same ontology elements, i.e. "BAW85A DESCEND 320 FL". We use these ontology mappings throughout the rest of this paper. Ontologies for speech understanding were not only evaluated and implemented in Europe. Chen et al. compared the European and US ontologies in [17] and in the extended version in [18]. The term ABSR is often used in these publications and was since then extended to ASRU due to changes in the technology and to align with the already commonly used term of ASR.

B. *Simulation Pilot Replacement by ASRU*

ATCo trainings or ATC simulations with ATCos often involve simulation-pilot(s). A simulation-pilot responds to a clearance or issues a request to the ATCo to simulate ATC communication. They manually input the ATCo clearances in visual interfaces to control the behavior of the aircraft so that the ATCos can see the changes accordingly on the radar screen. It is a human-intensive task. Normally one or two simulation pilots are required for an ATCo. DLR reported of having used four simulation pilots for one ATCo during heavy traffic scenarios [5]. Therefore, the integration of ASR in ATC training started already in the late 80s [10]. Nowadays, enhanced ASR systems are used in ATC training simulators to replace expensive pseudo pilots (e.g., FAA [19], DLR [20], MITRE [21], DFS [22]). Most of the integration of ASR are commercial products of an ATC simulator, which is enhanced by ASR. For example, DFS relies on UFA. A newer publication describes the ESCAPE platform used by Eurocontol [23]. ATC simulators used for ATC training of young ATCos require – for good reasons -- that the standard ICAO phraseology [24] is strictly followed (ICAO = International Civil Aviation Organization). No simulation pilots are needed, because automatic readbacks are generated by text-to-speech output. These ATC simulators with ASR integration are, however, of limited value, when using them for simulations with experienced ATCos. They sometimes deviate from standard phraseology patterns, which leads to a dramatic decrease in ASR recognition performance. Recently Zuluaga-Gomez et al. have presented a virtual simulation pilot, which is fully based on public domain software. They integrated elements from Natural Language Understanding (NLU) so that deviations from standard phraseology are possible aiming for comparable results on semantic level as presented in [25].

C. *Simulation Pilot Support by ASRU*

This paper proposes a different approach: Do not try to replace the simulation pilot, but try to support them as it is reported in the STARFiSH project [15]. Are two or even one simulation pilot enough to run a full simulation when supported by ASRU, which otherwise requires up to four simulation pilots.

Figure 1 shows the interface of the simulation pilot. The recognized words of the last five ATCo transmissions are shown at the bottom. This avoids many "say again", even when ASRU has failed. As soon as a callsign is recognized, the flight strip is highlighted by a white frame. This avoids searching for the correct aircraft, because the simulation pilot often controls more than six aircraft at the same time.

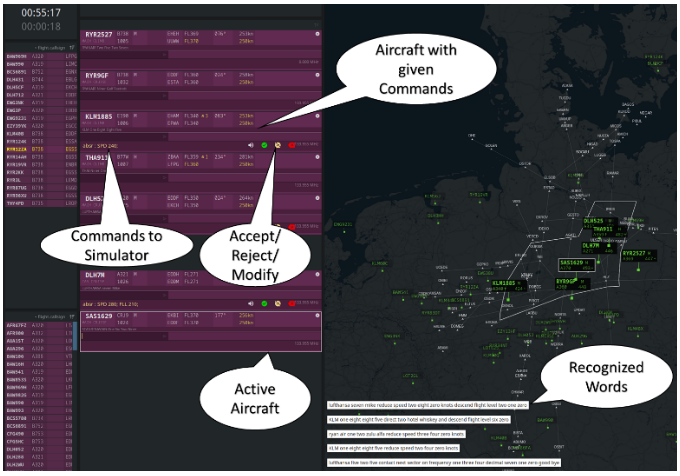

Figure 1.Integration of ASRU support into the simulation pilots interface

Figure 2 zooms into the flight strip of an aircraft. The simulator commands that the simulation pilot has to manually enter without ASRU support are automatically inserted into the flight strip – here "*QSY 134.710*" for a handover to the next frequency 134.710.

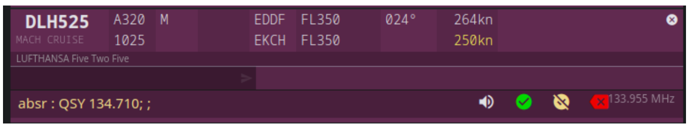

Figure 2. Simulation pilot support by ASRU integrated into the flight strip

The simulation pilot can now click on the green checkmark to accept all recognized commands for this aircraft, or on the red cross to reject everything, which clears the command line. The third option is to click on the yellow button, which enables the simulation pilot to modify the ASRU output, for example, to change one digit of the recognized frequency value.

D. Simulation Pilot Support within DIAL project

DLR and Idiap have developed different ASRU implementations. Already mentioned was the simulation for Vienna approach within SESAR PJ.10-W2-96-ASR project. An average word error rate (WER) of 3.1% were reported for 12 different ATCos with 0.7% as the best and 8.2% as the worst performance, see table 2 in [26]. In the context of multiple remote tower operations (project PJ.05-W2-97.2 [27]) only limited amount of voice recordings for training was available, i.e. 3.6 hours for Lithuanian controllers and only 54 minutes for controllers from Austro Control. DLR and Idiap achieved a WER of 13.6% and 9.8% for solution runs, when the ATCos benefit from ASRU support [28]. The project STARFiSH (Safety and Artificial Intelligence Speech Recognition) uses ASRU support for supporting both Frankfurt apron controllers and simulation pilots. An average WER of 3.1% was achieved: 3.3% for male speakers and 2.6% for female speakers (see table III in [29]). These three collaborations of DLR and Idiap used ASRU support in the lab environment. Within the HAAWAII project [30], voice recordings from the operational environment of the London TMA and from Isavia, the Icelandic air navigation service provider were used. For ATCo transmissions, a WER of 2.8% for recordings from London and 2.9% for recordings from Iceland were achieved. Even for noisy pilot transmissions, WERs of 7.1% and 10.4% were achieved, respectively (see table II in [11]).

All these speech-to-text engines needed training data from the target area. The objective derived from the research question of subsection I.B was to reuse the speech-to-text engine improved in SESAR PJ.10-W2-96 project for Vienna approach without modifications in the DIAL project. The simulation environment, i.e. the recording environment and microphones, is in both cases the same. The acoustic and language model of SESAR PJ.10-W2-96 is based on Vienna approach, whereas DIAL is based on Celle enroute traffic. The difference between approach and enroute also provides the first challenge. While ILS clearances or QNH information are not used in the enroute airspace, speed clearance with mach units are used. The main challenges for speech recognition are, however, waypoints and station names, which were never seen in the training data as they are specific to a sector or area. More than 580 different waypoints like *rakit* or *koduk*, and 17 different frequency station names like *ostsee*, *rhine* or *holstein* were modelled. It was not expected that these words will be recognized, i.e. it was assumed that the simulation pilots will not get any support from ASRU for DIRECT_TO or CONTACT clearances. The hope was, however, that misrecognitions of these words will not affect the recognition performance of other clearances like altitude, speed or heading.

Early on, it was clear that the word *maastricht* will be a problem. It occurred very often, because the word *maastricht* was often contained in the initial calls of the sector controller, e.g. "*lufthansa four uniform echo maastricht hello identified*". The word *maastricht* is not so important in this context, but it was often mixed with other words like *mach*. A combination of G-boosting and Lattice-Rescoring was used to detect new seldom occurring airport specific word entities during the first trials of DIAL in August and September 2023 [31]. The recognition performance of the word *maastricht* increased from 0% to 95%. Nevertheless, a bad recognition performance on sematic level of 11.5% was achieved for the DIRECT_TO command [31]. More waypoints were modelled for the trials in March 2024 and the word list used for boosting was reduced from 3-grams to 2-grams.

III. Performed Experiments and Used Data

107 different simulation runs were evaluated as shown in table I. 52 were performed in April 2024, 27 in March 2024 (with bad voice recording conditions) and 28 in August/September 2023. 3698 ATCo voice utterances were recorded, manually transcribed (word by word transcription of spoken content) and annotated (semantic meaning of spoken content).

TABLE I. DATA STATISTICS OF THE PERFORMED RUNS

	Aug/Sep 23	**March 24**	**April 24**	**All runs**
# runs	28	27	52	107
# Utterances	1315	1068	2754	3698
# Commands	2754	1728	4645	9127
# Relevant	2635	1605	4523	8763
# For Sim Pilots	1356	665	2175	4196
# ATCos	6	4	8	18
Av WER	8.1%	21.6%	9.7%	11.5%

All runs combined contained an overall of 9127 ATCo clearances (*# Commands*). The row headed "*# Relevant*" considers only those clearance types that occurred more frequently. Hence, clearance types, which occurred less than 20 times in the entire 107 runs were not considered as "Relevant" as they seldom occurred. Examples of "Relevant" types in shown in table IV in the following subsection IV.A. The type "*MAINTAIN SPEED*" occurred only four times and is, therefore, not counted as a relevant type. The row "*# For Sim Pilot*" measures the total number of ATCo commands which are relevant for the simulation pilot interface because they influence the aircraft behavior and which normally have to be entered manually by the simulation pilots. "*# ATCos*" shows the number of ATCos who took part in the various trials. "Av WER" shows the average word error rate, i.e., how good the speech recognition part of ASRU performed for the word by word recognition. During the March 2024 trials with four ATCos, a wrong setting was used for the voice recordings, leading to noisy voice recordings. This is reflected in the bad word error rate performance of 21.6% for March 2024 data. Here, the simulation pilots were not really supported by ASRU. The trials in April 2024 with eight different ATCos could on the other hand benefit from the improved recognition of seldom used airport specific words due to the used boosting technique mentioned in the previous section.

Table II shows the distribution of the different runs. 20 runs were considered as training runs in total. 36 runs were baseline. In these runs the ATCos were not supported by data link and planning support. Many voice commands were given. During the remaining 51 solution runs the ATCos were heavily supported by assistant systems and data link. They used voice commands only in emergency situations. Therefore, the number of spoken commands was quite small in those runs. The lower half of table II shows the scenario distribution with respect to the traffic amount. Not all 107 runs are included in this distribution since it excludes the 20 training runs. In Aug/Sep 23 only low and high traffic scenarios were performed.

TABLE II. DISTRIBUTION OF BASELINE, SOLUTION AND TRAINING RUNS AND OF LOW, HIGH AND VERY HIGH TRAFFIC SCENARIOS

	Aug/Sep 23	**March 24**	**April 24**	**All runs**
# Baseline	10	7	19	36
# Solution	13	16	22	51
# Training	5	4	11	20
# Low Traffic	11	6	14	31
# High Traffic	12	10	13	35
# Very High	0	7	14	21

IV. RESULTS WITH RESPECT TO ASRU PERFORMANCE

A. *Command and Callsign Recognition Rates*

Table III shows the performance of the ASRU system with respect to recognizing the semantics of ATCo clearances (commands). *Rec Rates* presents the command recognition rate. A command is considered to be correctly recognized, only if [32]

- the callsign is correct, e.g., *DLH3ER* even if only “three echo romeo” was spoken, but it is clear from context that it must be *DLH3ER*
- the type is correct, e.g., *REDUCE, DESCEND*, etc.
- the value is correct, e.g., 300 for a heading or *MOBSA* as a waypoint
- the unit is correct, e.g., flight level, feet, none, etc.
- the qualifier is correct, e.g., or greater, or less, left, etc.
- and the condition is correct.

TABLE III. ASRU PERFORMANCE FOR THE SIMULATION PERIODS

	Rec Rates	**Err Rates**	**Csgn R-Rates**	**Csgn Err Rates**	**Sim Rec Rates**
Aug/Sep 23	68.0%	6.4%	95.8%	2.4%	65.8%
March 24	51.8%	3.2%	80.7%	10.0%	50.5%
Apr 24	73.6%	3.7%	94.9%	2.3%	79.6%
All runs	67.8%	4.4%	92.2%	3.9%	70.5%

If “*DLH123 DESCEND 100 FL*” is recognized, but “*DLH123 DESCEND 100 none*” is the correct ATCo command because the ATCo did not mention a unit, this would be counted as an error and not as a recognition. Column “*Err Rates*” presents the command recognition error rate. Recognition and error rates do not sum up to 100% because we have another metric called the command rejection rate which is not shown in table III. Command rejection rate is the percentage of ATCo commands which were not recognized and for which no output was provided by ASRU. For example, if a *CLIMB* command was given by the ATCo but ASRU does not output anything, such commands are said to be rejected. “*Csgn R-Rates*” and “*Csgn Err Rates*” consider the same rates, but on callsign level, i.e. “only” the callsign needs to be correct. In some cases, a callsign is said, but no callsign is recognized. Therefore, callsign recognition and error rates also do not sum up to 100%, because this is considered a rejection (at least a wrong callsign is not provided). Column “*Sim Rec Rates*” shows the command recognition rates, considering only those command types, which are relevant for the simulation pilot interface. For example, a recognized *GREETING* command is not relevant for the simulation pilot and is hence not shown.

Table IV presents the recognition performance for the command types, which occurred at least 10 times during the 107 runs. We present the command recognition rates of the three validation campaigns and the total number of occurrences of each command type.

TABLE IV. ASRU PERFORMANCE PER COMMAND TYPE

Type	Aug/Sep Rec Rate	March 24 Rec Rate	Apr 24 Rec Rate	All Trials #Total
AFFIRM	88.2%	66.7%	90.0%	183
ALTITUDE	92.3%	41.7%	93.5%	69
CALL_YOU_BACK	76.2%	40.0%	94.7%	64
CLEARED TO	0.0%	7.7%	57.1%	34
CLIMB	90.0%	70.7%	86.3%	1050
CONTACT	35.6%	1.7%	67.3%	926
CONTACT_FREQUENCY	86.7%	62.4%	90.8%	1181
CONTINUE PRESENT_HEADING	62.5%	71.4%	88.2%	32
CORRECTION	83.3%	66.7%	71.0%	105
DESCEND	89.7%	59.5%	92.7%	356
DIRECT_TO	18.0%	3.8%	40.9%	258
DISREGARD	100.0%		50.0%	13
FAREWELL	83.7%	61.8%	55.0%	874
GREETING	55.4%	20.5%	46.4%	980
HEADING	33.3%	75.0%	92.3%	28
INFORMATION MISCELLANEOUS	0.0%	0.0%	0.0%	10
INIT_RESPONSE	82.6%	46.6%	76.6%	1177
MAINTAIN ALTITUDE	100.0%	33.3%	77.8%	15
NEGATIVE	75.0%	66.7%	80.8%	37
NO_CONCEPT	92.6%	74.8%	90.2%	786
RATE_OF_CLIMB	8.1%	42.9%	80.0%	146
RATE_OF_DESCENT	11.9%	50.0%	81.4%	114
REPORT_MISCELLANEOUS	50.0%	45.8%	64.9%	123
SAY_AGAIN	75.0%	50.0%	94.7%	25
SPEED	66.7%		33.3%	12
STATION	43.4%	40.0%	43.9%	353
STOP_CLIMB	41.7%	33.3%	88.2%	35
TURN_BY	100.0%	66.7%	88.9%	66
VERTICAL_RATE	76.5%	33.3%	100.0%	22

In the above table the command recognition rates of rarely occurring command types (said < 10 times) like *CONTINUE APPROACH, INCREASE, INFORMATION QNH, INFORMATION TRAFFIC, LEAVE_FREQUENCY, MAINTAIN SPEED, NAVIGATION_OWN, NO_SPEED_RESTRICTIONS, RATE_OF_CLIMB OWN, RATE_OF_-DESCENT EXPEDITE, RATE_OF_DESCENT OWN, REDUCE*, several *REPORT* commands, *RESUME_NORMAL_SPEED, STOP_DESCEND* and *VERTICAL_RATE OWN* are not shown. Command types shaded with grey are relevant for the simulation pilots. If a cell is empty the corresponding command type was not observed during the trials of the corresponding campaign, e.g. *SPEED* in March 24 trials. Cell values are shaded with blue, if the command type contains airspace dependent words like waypoint names, frequency station names or words like *maastricht* in the *STATION* command type.

The improvable recognition of these airspace dependent words also has a negative impact on the command recognition performance of the *CLIMB* or *DESCEND* command types, whose extraction performance was better than 95% in previous experiments, see table 8 in [25]. Some reasons are misrecognition of the word *"two"* and *"to"* like in the following example, in which recognition of "*climb flight level* ***two*** *three eight zero own rate of climb*" instead of "*...level* ***to*** *three eight zero ...*" resulted in a rejected *CLIMB* command. More often, the generic command type *ALTITUDE* was extracted instead of *CLIMB* or *DESCEND*, e.g. when ASR recognizes "*roger flight level three two zero*", but "*roger* ***climb*** *flight level three two zero*" was said. This is an example for extraction of the wrong command type, but fortunately this does not matter for the simulation pilot interface because just the actual flight level or altitude value is relevant here.

The bad performance of the vertical clearance rates in Aug/Sep 2023 was a problem in the implementation of the command extraction, e.g. "*descend flight level three two zero one thousand four hundred*" was not correctly recognized as two values, 320 for the flight level and 1400 for the vertical rate. The bad performance of extraction of *GREETING* and *FAREWELL* itself is not a problem, but wrongly recognizing words for these concepts could result in incorrectly extracting important commands, which, however, was very seldom the case. A typical example was the recognition of *"****commuter*** *nine one for* ***now transition low*** *identified*" instead of the correct transmission "***pobeda*** *nine one four* ***maastricht hello*** *identified*". Very often the word *"hello"* was not recognized.

B. *ASRU Performance for simulation pilots*

The following table V shows the ASRU recognition performance for the simulation pilot interface. In other words, how many of the commands relevant for the simulation pilot interface could be correctly provided by ASRU, as compared to a theoretically perfect system that does not make any errors?

TABLE V. CORRECT, WRONG AND MISSING SIMULATOR COMMANDS DUE TO ASRU PERFORMANCE

	total	correct	subs	ins	del	Rec Rate	Err Rate	Sim Rec Rate
Aug/Sep 23	1079	902	14	14	163	83.6%	2.6%	65.8%
March 24	566	355	4	12	207	62.7%	2.8%	50.5%
Apr 24	1721	1465	55	26	202	85.1%	4.7%	79.6%
All runs	3366	2722	73	52	572	80.9%	3.7%	68.8%

In table V, we see that 1079 commands should have been given by the simulation pilot during the trials in August and September 2023. Note that the column "total" in table V is different from the row "For Sim Pilot" in table I. Not every command counted in table I is mapped to one simulator pilot command, e.g. *CONTACT* (nearly 300 in 2023) and *CONTACT_FREQUENCY* (nearly 400 in 2023) result in one simulator command. 902 of the 1079 commands were correctly provided by the ASRU software to the simulation pilot interface, i.e. 83.6% of the commands were correct. "*subs*" counts the substitutions, i.e., a wrong simulator command was provided by ASRU instead of the correct one, e.g., a descend to fight level 320 instead of 330. *"ins"* counts the insertions, i.e., a simulator command was provided by ASRU, but no such command was given by the ATCo, e.g., a given frequency value was interpreted as a rate of descent. "*del*" counts the deletions, i.e., the ATCo gave a command relevant for the simulator, but ASRU does not provide anything.

"*Err Rate*" in table V is the percentage of wrong inputs to the simulation pilot from ASRU, i.e., (subs + ins) / total. "*Sim Rec Rate*" has been already described in the previous section. It measures the ASRU performance when calculating the command recognition rate, considering a smaller set of commands, which are relevant for the simulation pilot interface. For the simulation pilot interface, it is enough if it receives, for example "*flight level 320*" for the correct callsign. It is irrelevant if a *CLIMB* is recognized as a *DESCEND*, or when the qualifier *OR_GREATER* or the unit is not correctly recognized. For example, the value of a flight level itself shows if the aircraft has to go up or down and the value for a vertical rate can be assumed to be feet per minute even if no unit is recognized.

The performance has improved in terms of the recognition rate from Aug/Sep 23 to April 2024 from 83.6% to 85.1%. Nevertheless, each sixth command given by the ATCo is not correctly outputted by the ASRU software.

The following table VI shows the wrong inputs of the ASRU software from table V with respect to command classes relevant for the simulation pilot interface.

TABLE VI. WRONG AND MISSING SIMULATOR COMMANDS DUE TO ASRU PERFORMANCE PER COMMAND TYPE CLASS

	Alt	Waypoint	Vertical Rates	Hand over	Speed	Heading
Aug/Sep 23	47	53	29	54	4	4
	24.6%	27.7%	15.2%	28.3%	2.1%	2.1%
March 24	95	27	9	85	1	6
	42.6%	12.1%	4.0%	38.1%	0.4%	2.7%
Apr 24	89	106	17	58	3	10
	31.4%	37.5%	6.0%	20.5%	1.1%	3.5%
All runs	231	186	55	197	8	20
	33.1%	26.7%	7.9%	28.3%	1.1%	2.9%

- "*Alt*" subsumes *CLIMB*, *DESCEND, ALTITUDE, STOP_CLIMB* and *STOP_DESCEND* commands.
- "*Waypoint*" subsumes *NAVIGATION_OWN* and *DIRECT_TO* commands.
- "Vertical *Rates*" subsumes all commands which provide a vertical rate, independent of if it is for a climb or descend.
- "*Handover*" subsumes *CONTACT_FREQUENCY*, *CONTACT* and *LEAVE_FREQUENCY*.
- "*Speed*" subsumes all commands providing a speed value directly or indirectly. This also includes *NO_SPEED_RESTRICTIONS*.
- "*Heading*" subsumes all commands containing an absolute or relative heading value.

From table VI, we see that the majority of the problems result from wrong waypoint recognitions (37.5% of all problems in

April 24). Nevertheless, we also have problems with the recognition of altitude commands and handovers (31.4% and 20.5% of all problems in April 24, respectively). The following table VII presents a deeper analysis.

TABLE VII. ANALYSIS OF COMMAND TYPE RESULTS FOR THE MAIN SUB TYPES ALT, WAYPOINT AND HANDOVER

	Alt			Waypoint			Handover		
	del	subs	ins	del	subs	ins	del	subs	ins
Aug/Sep 23	42	2	3	48	3	2	41	7	6
	27.3%	1.3%	1.9%	31.2%	1.9%	1.3%	26.6%	4.5%	3.9%
March 24	89	0	6	26	1	0	78	1	6
	43.0%	0.0%	2.9%	12.6%	0.5%	0.0%	37.7%	0.5%	2.9%
Apr 24	75	3	11	72	28	6	32	22	4
	29.6%	1.2%	4.3%	28.5%	11.1%	2.4%	12.6%	8.7%	1.6%
All runs	206	5	20	146	32	8	151	30	16
	33.6%	0.8%	3.3%	23.8%	5.2%	1.3%	24.6%	4.9%	2.6%

Considering the 186 problems with waypoint commands, we have 146 deletions (del), where a waypoint was said by the ATCo, but no waypoint (*DIRECT_TO*) was recognized. In 32 cases we have substitutions (subs), where a wrong waypoint was recognized and in 8 cases we have insertions (ins), where a *direct-to* was recognized, but was not given by the ATCo.

The number of substitutions marked in red in table VII has heavily increased from the Aug/Sep 2023 trials to the trials in April 2024. This is due to the improved version of the ASRU software, which recognizes many more waypoints, but at the cost of increasing the number of wrong recognitions.

TABLE VIII. ASRU PERFORMANCE FOR DIRECT_TO COMMAND

	Rec Rates	Err Rates
Aug/Sep 23	18.0%	8.2%
March 24	3.8%	7.7%
Apr 24	40.9%	19.3%
All runs	31.8%	15.5%

Table VIII shows that in April 2024, the recognition rate for *DIRECT_TO* command was 40.9% with an error rate of 19.3%. This is a significant improvement, compared to the recognition rate of 18% in Aug/Sept 23, which is of course not sufficient, but a first step.

V. PERFORMANCE OF THE SIMULATION PILOT TO CORRECT ASRU PROBLEMS

The question remains - which errors from wrong ASRU outputs are compensated by the simulation pilot? Are almost all errors recognized and corrected or are there wrong or missing inputs? The question is analyzed in this section.

A. Performance of Simulation Pilot Entries to Wrong ASRU Outputs

Table IX shows the results of how the simulation pilots corrected/accepted the ASRU outputs, which either contained one or more errors or rejections. "corr" is the number of cases in which the simulation pilot corrects wrong outputs from ASRU. This also includes cases were ASRU provides no output but the simulation pilot makes the correct entry on the interface. "*ign*" is the number of cases in which the ASRU software suggests/invents a command, which was never given and which the simulation pilot correctly ignores. The column "*% corr*" is the percentage of corrections/inputs made by the simulation pilot on wrong or missing ASRU outputs, *i.e.* where the simulation pilot has corrected within 30 seconds if an output was sent or within 60 seconds if no output was sent by ASRU. "*ins*" are the insertions which counts the number of cases, in which the simulation pilot inputs a command, which was not given by the ATCo. "*del*" are the deletions which counts the number of cases, in which the simulation pilot ignores or does not enter an ATCo command which should have be inputted, for example, a rate of descent. "*subs*" are the substitutions or cases where the command type is correct, but the simulation pilot enters the wrong value, for example *FLL 320* instead of *FLL330* for a flight level clearance. *"syn"* counts the syntax errors of the simulation pilot in the interface, for example inputting F instead of *FLL* for altitude commands. *"sum wrong"* is the number of total errors, which is the sum of *"ins"*, *"del"*, *"subs"* and *"syn"*.

TABLE IX. SIMULATION PILOT CORRECTIONS OF ASRU OUTPUTS

	corr	ign	% corr	ins	del	subs	syn	sum wrong
Aug/Sept 23	151	11	75.4%	5	44	9	0	58
March 24	176	8	82.9%	17	27	10	1	55
Apr 24	225	18	79.7%	15	48	13	1	77
All runs	552	37	79.4%	37	119	32	2	190

B. Analysis of Uncorrected ASRU Problems

TABLE X. TOTAL ASRU ERRORS VERUS REMAINING SIMULATION PILOT ERRORS

	total	ASRU Errors	%ASRU Errors	Rem Sim P Errors	%SP Errors
Aug/Sep 23	1079	191	17.7%	69	6.4%
March 24	566	223	39.4%	55	9.7%
Apr 24	1721	283	16.4%	77	4.5%
All runs	**3366**	**697**	**20.7%**	**201**	**6.0%**

Table X shows the total number of relevant simulation pilot commands in column *"total"*. *"ASRU Errors"* and *"%ASRU Errors"* denotes the total number errors in the ASRU output and their corresponding percentages with respect to the total number of relevant commands, respectively. *"Rem Sim P Errors"* denotes the number of ATCO commands, which were not corrected by the simulation pilots within the 30 (or 60) seconds and column *"%SP Errors"* is the percentage of the simulation pilot errors with respect to all relevant commands (*"total"*). 4.5% of uncorrected commands in the final runs of April 2024 is still a high percentage of problems, which might have a major influence on the simulation results. Therefore, we analyzed the problems of the April 2024 runs in more detail. 77 uncorrected ASRU errors is a high number, but distributed over 52 simulation runs, which makes this number less dramatic.

Figure 3 summarizes the contents of table X in graphical form. During the first trials in 2023 using the existing version of speech recognition, 82% of the commands sent to the simulation pilot interface were correct. Out of the 18% incorrectly sent commands, 7% were remained uncorrected. These uncorrected commands are the ones which may disturb the simulation results.

Figure 3.Correct ASRU recognitions, corrections of simulation pilot and uncorrected ATCo commands

The number of ATCo commands, which were not correctly entered by simulation pilots, are worse during the March 2024 runs, with bad voice recording conditions resulting in word error rates of 21.6% (see table I). Each tenth command was not corrected. With better waypoint and altitude command recognition as well as adequate voice recordings in April 2024, the ASRU performance was better and the percentage of uncorrected commands decreased from 7% in 2023 to 4% in the final runs in April 2024. Currently we are analyzing the performance of simulation pilots without ASRU support and when supporting highly trained simulation pilots with ASRU.

Table XI analyses the ARSU errors which were not corrected by the simulation pilots per command type category. Similar to the previous sub-section, *"ins"*, *"del"* and *"subs"* denote the number of insertions, deletions and substitutions, respectively. The second row *"#errors"* for each type and trial shows the sum of insertions, deletions and substitutions. The third row *"%errWrtAllErrors"* shows the percentage of errors for each command type with respect to all errors, i.e., the 7 altitude errors make up 12.5% of all the 55 errors, which are not corrected by the simulation pilots during the Aug/Sep 23 runs. The fourth row "*%errWrtTotal*" shows the percentage of errors for each command type with respect to the total number of commands sent to the simulation pilot interface. The errors for some command types like "*RESUME_OWN_NAVIGATION*" and "*SPEED*" are not shown in the table, because they are insignificantly small. That is why the error percentages do not always sum up to 100%.

In the April 24 runs, the 23 cases of missing flight levels ("*ALT del*") are the most serious problems. Most of them occurred when ASRU did not output the flight level of the clearance from the ATCO, i.e., when the simulation pilot interface did not receive a flight level from ASRU, many times the simulation pilot also misses to enter them manually. In some cases, the simulation pilot ignored the flight level input because the aircraft had already reached that level, but these were not the majority of cases. The five flight level substitutions in table XI can also not be neglected. We had cases where the simulation pilot entered flight level of 330 instead of 320, 290 instead of 390, 300 instead of 330.

TABLE XI. ERRORS IN SIMULATION PILOT CORRECTIONS ON WRONG ASRU OUTPUTS PER COMMAND TYPE

		ALT			DIRECT_TO / HEADING			VERTICAL RATES			HANDOVER		
		ins	del	subs	ins	del	subs	ins	del	subs	ins	del	subs
Aug/Sept 23		0	7	0	1	12	4	1	14	2	1	11	2
	#errors	7			17			17			14		
	%errWrt AllErrors	12.5%			30.4%			30.4%			25.0%		
	%errWrt Total	9.6%			32.1%			56.7%			29.8%		
March 24		4	11	1	1	8	1	10	0	2	0	8	5
	#errors	16			10			12			13		
	%errWrt AllErrors	30.2%			18.9%			22.6%			24.5%		
	%errWrt Total	16.8%			37.9%			120.0%			16.5%		
Apr 24		1	23	5	1	14	3	9	9	0	2	9	2
	#errors	29			18			18			13		
	%errWrt AllErrors	37.2%			23.1%			23.1%			16.7%		
	%errWrt Total	27.1%			19.4%			94.7%			21.0%		
All runs		5	41	6	3	34	8	20	23	4	3	28	9
	#errors	52			45			47			40		
	%errWrt AllErrors	27.8%			24.1%			25.1%			21.4%		
	%errWrt Total	18.9%			26.1%			79.7%			21.3%		

The biggest problem for the simulation pilot is the wrong or missing recognition of waypoints. The simulation pilots often did not understand the waypoints, because they are not real pilots flying in that area. Therefore, they are often just ignored. The support for the simulation pilot has increased from 2023 runs to April 2024, when the command recognition rate of the *DIRECT_TO* command increased from 18.0% to 40.9% (table VIII). For vertical rates, the insertions mostly occurred, when the simulation pilot enters a rate based on his/her own judgement, because inserting a new flight level resets a previously given vertical rate. The deletions are, however, a problem of ASRU. The missing handover actions of the simulation pilots complicate their work, because more flights are shown in their interface than necessary. Last but not the least, we need to mention that 77 uncorrected commands mean that at least 206 of the 283 ASRU errors were corrected by the simulation pilots.

VI. CONCLUSIONS

We've shown that Automatic Speech Recognition and Understanding (ASRU) eases the task of the simulation pilots. Three validation campaigns were performed between August

2023 and April 2024 in DLR's Air Traffic Management Simulation Environment in Braunschweig. Each time the interface of the simulation pilot was improved. Just displaying the word transcriptions of the transmissions is already enough to avoid some "say again" of the simulation pilots or even avoids wrong simulator inputs. In addition, using speech understanding, i.e. transforming the recognized sequence of words into corresponding simulator inputs, which the simulation pilot can just accept, manipulate or reject, enables the biggest reduction in workload.

The reduction in workload with ASRU support could, however, also result in some errors, where wrong or missing inputs are sent to the simulator, because ASRU outputs wrong or no simulator commands. In the third campaign in April 2024, 4% of the air traffic controller (ATCo) clearances were still not correctly entered by the simulation pilots after 30 (or 60) seconds. Although a baseline run without ASRU support was not performed yet, there are strong suggestions that the wrong or missing inputs are not due to over trusting the ASRU system. In March trials, i.e. the second campaign, with bad configuration of the speech recognition system, the simulation pilots were aware of the fact that most of the ASRU outputs are not reliable. The word error rate was above 20% compared to 10% in the previous campaign, resulting in a command recognition rate of only 50% instead of 80% in the previous campaign. Nevertheless, 9% of the simulation pilot commands were not correct in March as compared to only 4% in the next April 24 campaign. The biggest drawback currently is the moderate performance of recognizing airspace specific words, e.g., the five letter codes of waypoints. Simulation pilots familiar with the airspace understand words like ekern, mobsa or sirlu, but new simulation pilots are lost in 50% of the cases with ASRU support, but in nearly all cases without ASRU support. The research question, whether it is possible to use an ASRU system which is trained for one airspace for another airspace without re-training the acoustic and language models is now answered with "Re-Training is still needed".

The glass, however, is not half empty, but more than half full, when we do not try to completely replace the simulation pilots, but just support them. 80% of ATCo commands can be automatically transformed into correct simulator inputs. Two instead of four simulation pilots were sufficient to conduct the 107 simulations runs.

REFERENCES

[1] German Aerospace Center (DLR), "ATMOS," [Online]. Available: https://www.dlr.de/de/fl/forschung-transfer/validierungszentrum-luftverkehr/echtzeitsimulation/atmos. [Accessed 16. August 2024].

[2] I. Gerdes, M. Jameel, R. Hunger, L. Christoffels and H. Gürlük, "The Automation Evolves: Concept for a Highly Automated Controller Working Position," *33rd International Congress of the Aeronautical Sciences, ICAS,* 2022.

[3] M. Jameel, L. Tyburzy, I. Gerdes, A. Pick, R. Hunger and L. Christoffels, "Enabling Digital Air Traffic Controller Assistant through Human-Autonomy Teaming Design," *IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), Barcelona, Spain, doi: 10.1109/DASC58513.2023.10311220,* 2023.

[4] R. Hunger, L. Christoffels, M. Friedrich, M. Jameel, A. Pick, I. Gerdes, P. von der Nahmer and F. Sobotzki, "Lesson Learned: Design and Perception of Single Controller Operations Support Tools," *21st International Conference, EPCE 2024, held as part of 26th HCI International Conference (HCII 2024), Proceedings, Part I; Washington, DC, USA, June 29 – July 4,* 2024.

[5] N. Ahrenhold, H. Helmke, T. Mühlhausen, O. Ohneiser, M. Kleinert, H. Ehr, L. Klamert and J. Zuluaga-Gómez, "Validating Automatic Speech Recognition and Understanding for Pre-Filling Radar Labels—Increasing Safety While Reducing Air Traffic Controllers' Workload," *Aerospace 10, 538.,* 2023.

[6] B. Juang and L. Rabiner, "Automatic speech recognition -- a brief history of the technology development," in *Ga. Inst. Technol. Atlanta Rutgers*, Univ. Univ. California St. Barbara, 2005.

[7] D. Connolly, "Voice Data Entry in Air Traffic Control," in *Report N93-72621; National Aviation Facilities Experimental Center*, Atlantic City, NJ, USA, 1977.

[8] C. Hamel, D. Kotick and M. Layton, "Microcomputer System Integration for Air Control Training," in *Special Report SR89-01; Naval Training Systems Center*, Orlando, FL, USA, 1989.

[9] J. Cordero, N. Rodríguez, J. de Pablo and M. Dorado, "Automated Speech Recognition in Controller Communications applied to Workload Measurement," in *3rd SESAR Innovation Days*, Stockholm, Sweden, 26–28 November, 2013.

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

[11] H. Helmke; K. Ondřej; S. Shetty; H. Arilíusson; T. S. Simiganoschi; M. Kleinert; O. Ohneiser; H. Ehr; J.-P. Zuluaga; P. Smrz, "Readback error detection by automatic speech recognition and understanding: results of HAAWAII project for Isavia's enroute airspace," in *12th SESAR Innovation Days*, Budapest, Hungary, 2022.

[12] V. Nguyen and H. Holone, "N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in Air Traffic Control," in *16th International Conference on Control, Automation and Systems (ICCAS), Gyeong*, Gyeongju, Republic of Korea, 16–19 October, 2016.

[13] Y. Lin, "Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application," *Aerospace 8(3),* March 2021.

[14] H. Helmke, M. Slotty, M. Poiger, D. Ferrer Herrer, O. Ohneiser, N. Vink, A. Cerna, P. Hartikainen, B. Josefsson, D. Langr, R. García Lasheras, G. Marin, O.-G. Mevatne, S. Moos, M. N. Nilsson, M. Boyero Pérez, "Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04," in *IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, London, United Kingdom, 2018.

[15] M. Kleinert, O. Ohneiser, H. Helmke, S. Shetty, H. Ehr, M. Maier, S. Schacht and H. Wiese, "Safety Aspects of Supporting Apron Controllers with Automatic Speech Recognition and Understanding Integrated into an Advanced Surface Movement Guidance and Control System," *Aerospace 2023, 10, 596,* 2023.

[16] H. Helmke, M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr, H. Arilíusson, T. S. Simiganoschi, A. Prasad, P. Motlicek, K. Veselý, K. Ondřej, P. Smrz, "Readback error detection by automatic speech recognition to increase ATM safety," in *14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)*, Virtual Conference, 2021.

[17] H. Helmke, O. Ohneiser, M. Kleinert, S. Chen, H. D. Kopald and R. M. Tarakan, "Transatlantic Approaches for Automatic Speech Understanding in Air Traffic Management," in *submitted to 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)*, Savannah, GA, USA, 2023.

[18] S. Chen, H. Helmke, R. Tarakan, O. Ohneiser, H. Kopald and M. Kleinert, "Effects of Language Ontology on Transatlantic Automatic Speech Understanding Research Collaboration in the Air Traffic Management Domain," *Aerospace 2023, 10, 526.,* 2023.

[19] FAA, "National Aviation Research Plan (NARP)," March 2012.

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

[21] R. Tarakan, K. Baldwin and N. Rozen, "An automated simulation pilot capability to support advanced air traffic controller training," in *The 26th Congress of ICAS and 8th AIAA ATIO*, Anchorage, AK, USA, 2008.

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

[23] A. Bouchal, P. Had and P. Bouchaudon, "The Design and Implementation of Upgraded ESCAPE Light ATC Simulator Platform at the CTU in Prague," in *Proceedings of the 2022 New Trends in Civil Aviation (NTCA)*, Prague, Czech Republic, 26–27 October 2022.

[24] International Civil Aviation Organization (ICAO), "Doc 4444 ATM/501; ATM (Air Traffic Management): Procedures for Air Navigation Services," Montréal, QC, Canada, 2007.

[25] H. Helmke, M. Kleinert, N. Ahrenhold, H. Ehr, T. Mühlhausen, O. Ohneiser, L. Klamert, P. Motlicek, A. Prasad, J. Zuluaga Gomez, J. Dokic and E. Pinska Chauvin, "Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload," in *15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)*, Savannah, GA, USA, 2023.

[26] H. Helmke, M, M. Kleinert, O. Ohneiser, N. Ahrenhold, L. Klamert and P. Motlicek, "Safety and Workload Benefits of Automatic Speech Understanding for Radar Label Updates," *AIAA Journal of Air Transportation; https://arc.aiaa.org/doi/abs/10.2514/1.D0419,* 2024.

[27] J. Jakobi, "The Project PJ05-W2 DTT," [Online]. Available: https://www.remote-tower.eu/wp/project-pj05-w2/solution-97-2. [Accessed 13. September 2024].

[28] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, Š. Murauskas, T. Pagirys, G. Balogh, A. Tønnesen, G. Kis-Pál, R. Tichy, V. Horváth, F. Kling, W. Rinaldi, S. Mansi and H. Usanovic, "Understanding Tower Controller Communication for Support in Air Traffic Control Displays," in *12th SESAR Innovation Days*, Budapest, Hungary, 2022.

[29] M. Kleinert, S. Shetty, H. Helmke, O. Ohneiser, H. Wiese, M. Maier, S. Schacht, I. Nigmatulina, S. S. Sarfjoo and P. Motlicek, "Apron Controller Support by Integration of Automatic Speech Recognition with an Advanced Surface Movement Guidance and Control System," in *12th SESAR Innovation Days (SID 2022)*, Budapest, Hungary, 2022.

[30] Deutsches Zentrum für Luft- und Raumfahrt (DLR), "HAAWAII: highly automated air traffic controller workstations with artificial intelligence integration," [Online]. Available: https://www.haawaii.de/wp/.

[31] M. Bhattacharjee, P. Motlicek, I. Nigmatulina, H. Helmke, O. Ohneiser, M. Kleinert and H. Ehr, "Customization of Automatic Speech Recognition Engines for Rare Word Detection Without Costly Model Re-Training," in *13th SESAR Innovation Days*, Seville, Spain, Nov. 2023.

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