Understanding Tower Controller Communication for Support in Air Traffic Control Displays

Oliver Ohneiser; Hartmut Helmke; Shruthi Shetty; Matthias Kleinert; Heiko Ehr
German Aerospace Center (DLR),
Institute of Flight Guidance,
Lilienthalplatz 7, 38108 Braunschweig,
Germany
Oliver.Ohneiser@DLR.de

György Balogh; Andreas Tønnesen
Indra Navia AS,
Hagaløkkveien 26,
1383 Asker, Norway
Gyorgyi.Balogh@INDRA.no

Walter Rinaldi; Sara Mansi; Giuseppe Piazzolla
LEONARDO S.p.A.
Via Tiburtina km 12,400,
00131 Roma, Italy,
Walter.Rinaldi@Leonardo.com

Šarūnas Murauskas, Tomas Pagirys
State Enterprise Oro Navigacija (ON),
Lithuanian Air Navigation Service Provider,
B. Karvelio str. 25,
02184 Vilnius, Lithuania

Gábor Kis-Pál; Roland Tichy; Viktor Horváth; Fanni Kling
HungaroControl
Igló str. 33-35,
1185 Budapest, Hungary

Haris Usanovic
Austro Control (ACG),
Schnirchgasse 17,
1030 Vienna, Austria

***Abstract*—Automatic speech recognition and understanding for air traffic control (ATC) communication has been extensively studied in the approach and en-route environment. *SESAR2020's Solution 97.2* is one of the first European attempts to analyze recognition rates and human performance of air traffic controllers (ATCos) in simulated tower and ground environments. Three validation exercises with 22 ATCos from four different European air navigation service providers were conducted in Germany, Norway, and Italy. The validated artificial intelligence-based prototypes of Assistant Based Speech Recognition systems (ABSR) supported ATCos fulfilling tasks in a ground and tower environment as well as multiple remote tower environment, respectively. Thus, in any relevant ATC display, (1) recognized callsigns of ATCo utterances have been highlighted, (2) fully recognized commands were shown, and (3) the ATCo was able to manually manipulate the ABSR output if needed or the output was automatically accepted by the ATC system otherwise. This paper evaluates callsign and command recognition rates as well as ATCo performance. It compares the results for the three validation exercises: a callsign recognition rate of 81-98%, a command recognition rate of 65-91%, and a slight reduction in ATCo workload on a low workload level.**

Keywords—Air Traffic Controller; Tower; Ground; Assistant Based Speech Recognition; Automatic Speech Recognition; Automatic Speech Understanding; Solution 97.2

I. INTRODUCTION

Air traffic controllers (ATCos) in the tower and ground environment still mainly issue time-critical instructions to pilots via radio telephony with voice utterances. The ATCos normally also note down the content of their verbal instructions. The paper flight strips for such notes have meanwhile been replaced by electronic flight strip systems [1] within the major air navigation service providers (ANSPs). However, the chain of actions and inputs has not been completely digitized yet. ATCos still need to manually enter their instruction content with mouse or keyboard into the electronic systems. This causes unnecessary workload for the ATCos as the instruction has already been expressed verbally [2]. Thus, ATCos could be supported by automatic speech recognition and understanding (ASRU) if their spoken words are recognized, if the resulting word sequences are understood, and if the semantics are entered into the right electronic air traffic control (ATC) systems, to feed a sequence of succeeding applications.

Similar ASRU system prototypes have already been developed for the following ATC domains: approach [3],[4], en-route [5], and apron [6]. Such systems consider relevant accessible context information from an assistant system, e.g., use the callsign information from surveillance data to improve callsign recognition performance or check plausibility of recognition results due to flight status information from electronic ATC systems. Those ASRU systems are, therefore, called assistant based speech recognition (ABSR) systems [7]. ABSR system prototypes in the above-mentioned ATC domains showed a reduction of ATCo workload [8] and an increase in ATM efficiency [7].

Hence, it makes sense to fill the gap and develop an ABSR system also for the tower and ground environment [9]. 'Tower environment' includes remote towers and even multiple remote towers, i.e., the ASRU system must recognize and understand ATCo utterances for up to three different airports at the same time (note: as there is only one ATC frequency in the multiple remote tower environment, the ATC utterances enter the ABSR system sequentially, i.e., there are not multiple utterances at the same time to be handled by the ABSR system). Instructions in the tower environment for example encompass runway clearances, instructions for the ground environment include taxi, startup, pushback and ATC clearances. These ABSR prototype developments happened in the course of the SESAR2020 wave 2 project "Digital Tower Technologies (DTT)" [10] and more specifically in its solution 97.2 on automatic speech recognition (ASR).

Three ABSR prototypes comprising of speech-to-text (STT) and text-to-concept (TTC) modules for three different validation platforms have been developed and tested in validation exercises, as shown in TABLE I. They aimed and claimed to reach Technology Readiness Level 4.

TABLE I. ABSR PROTOTYPES, PLATFORMS, AND VALIDATIONS

	LIT	*NOR*	*BUL*
Organizations	DLR	Indra Navia, HungaroControl	Leonardo
STT	based on KALDI: Idiap, Switzerland	HungaroControl (voice system from Indra)	based on KALDI: Leonardo
Audio Recording	16 kHz mono	8 kHz mono	8 kHz mono
TTC	DLR	HungaroControl	Leonardo
Human Machine Interface	DLR Electronic Flight Strip System Prototype	Indra Navia InNOVA System	Leonardo Lead In Sky Ground Working Position, A-SMGCS
Environment	Multiple Remote Tower and Ground (three airports for **Lit**huania)	Remote Tower (**Nor**wegian airports)	Tower and Ground (Sofia airport from **Bul**garia)
Validation Place and Date	Braunschweig, Germany, quarter 1, 2022	Asker, Norway, quarter 4, 2021	Rome, Italy, quarter 2, 2022
ATCos (Validation Participants)	10 ATCos: 5 from ON (Lithuania), 5 from ACG (Austria)	6 ATCos from HungaroControl (Hungary)	6 ATCos from ENAV (Italy)

This paper consists of related work with respect to speech understanding, i.e., the extraction of ATC concepts such as callsigns and command types plus values from ATC utterance transcriptions in Section II. Section III outlines the prototypes for tower/ground ABSR systems with relevant human machine interfaces. The three validation exercises with ATCos and the study setups for data recording are explained in Section IV. Section V compares results on speech understanding quality as well as ATCos' performance and feedback from the ATCos of the three validation exercises. Section VI presents the conclusions and discusses future work.

II. RELATED WORK ON SPEECH UNDERSTANDING

A. History of ASRU in ATC

At the beginning of this century, ASR systems in ATC targeted to reduce or replace simulation pilots and to advance simulation infrastructure [11],[12] Also, the estimation of ATCo workload has been derived from ASR data [13],[14]. Later, the project AcListant-Strips® explored ATCo support for the simulated Düsseldorf approach environment through automatically maintaining aircraft radar labels with a first ABSR prototype [2]. ATCos' workload for manually maintaining the labels could be reduced up to a factor of three. Subsequently, ATCos had more time for their primary task – controlling air traffic. As result, actual aircraft trajectories in the terminal maneuvering area were more efficient, yielding savings of more than 100 kg CO_2 per flight [7].

The SESAR exploratory research project MALORCA investigated machine learning algorithms to quickly adapt an ABSR system to a new ATC environment with the use cases of Vienna and Prague approach on operational ATCo speech data [4]. The prediction of callsigns and commands via machine learning for various ATC environments such as approach [15] and multiple remote tower [16] helped to reduce the recognition error rates on word level and especially on semantic level [17] through command extraction algorithms [18]. Applications such as callsign highlighting [19] and radar label maintenance have been studied with their effect on ATCo workload [8]. ABSR has also been advanced for apron/ground control at a major European airport to support ATCos and simulation pilots [6]. With more training data [20], the recognition rates further improved and the applications that use the ASRU output also advanced. Thus, more functions such as pilot weather report extraction [21], detection of runway incursions [22] and detection of readback errors on the ground [23] and in the air [5] have been developed.

The latter application has been implemented in the course of the SESAR exploratory research project HAAWAII [24], [25]. En-route utterances from both ATCo and pilots [26] in Icelandic en-route airspace were automatically recognized (STT), annotated (TTC), grouped, and analyzed for potential readback errors [5]. This ASRU prototype also ran in the operational ATC center of the Icelandic ANSP for demonstration purposes. As a benchmark, word error rates of below 4% for ATCos and 7-11% for pilots, callsign recognition error rates below 2%, and command recognition error rates below 7% were achieved, e.g., in HAAWAII on noisy data from the operational environment. These numbers were validated in an ASRU setup that goes far beyond competitive systems, i.e., covers a wider range of command types, recognizes command qualifiers and conditions [27], works well on unseen operational data in the given use cases, and offers well-appreciated ABSR applications as ATCo feedback shows.

B. Formats for Speech Understanding in Tower Environment

The first step in ASRU is STT, i.e., the **transcription**. A set of rules for the transcription of ATC utterances has been agreed [28] and was also used for the three prototypes of this paper. The second step in ASRU is TTC [29] also called **annotation** or "spoken instruction understanding" [30]. Almost two dozen stakeholders such as ANSPs and air traffic management (ATM) system providers agreed on an ontology, i.e., a set of rules of how to annotate transcribed ATC communication, in the SESAR industrial research project PJ.16-W1-04-ASR [31]. The rules were extended in this project (Solution 97.2) and especially in the HAAWAII project [5]. As an example, the transcription "air france one two victor now taxi via alfa hold short of runway one three" would be annotated as "AFR12V TAXI VIA A, AFR12V HOLD_SHORT RW13" given configuration files for airspace/airport topology entities. All three prototypes of this paper stuck to this ontology. They partly used machine learning techniques to, e.g., predict relevant ATC concepts that are expected in the following ATC utterances. Thus, the three ABSR prototypes of this paper used the same basics, but have very different background in the maturity of implementation, different amount of training data, and prototype dependent usage of ABSR output for ATC applications.

III. ASSISTANT BASED SPEECH RECOGNITION SYSTEMS FOR AERODROME ENVIRONMENTS

This section describes the ABSR prototypes LIT (DLR and Idiap), NOR (Indra Navia / HungaroControl), and BUL (Leonardo) with inputs and outputs.

A. LIT: Multiple Remote Tower and Ground ABSR Prototype

The ABSR system received a continuous live audio stream via voice over IP, also transmitted to the simulation pilots as soon as the ATCo pressed the push-to-talk button until releasing it again. For the ABSR system training phase, dozens of hours of publicly available ATC speech datasets were used to train acoustic model, language model, command prediction model, and command extraction model. These generic models were adapted by only 2,500 manually transcribed and annotated utterances, i.e., 3.6 h from ON and 0.9 h from ACG ATCos' speech from the specific multiple remote tower setup.

The STT engine of Idiap continuously sent the recognized words from the audio stream in a JavaScript Object Notation (JSON) format to the TTC module of DLR. The TTC module also received surveillance data from the three remote airports (called Vilnius, Kaunas, and Palanga) to know the currently available aircraft callsigns and their positions. The TTC module automatically extracted the callsign from the word sequence as soon as enough relevant words have been provided by STT. The callsign recognition process based on surveillance data benefited from callsign boosting on STT level and supported, therefore, deviations of ATCo verbalizations within and outside of the International Civil Aviation Organization (ICAO) phraseology on semantic TTC level.

If further recognized words were received by the TTC module, the command recognition algorithm automatically extracted up to 63 different implemented command types with their values, units, qualifiers, and conditions even in utterances with instructions to more than one aircraft according to the ATC annotation ontology. This process was predominantly based on keyword spotting ("taxi", "lineup"), but also worked robustly if just the name of the taxiway or runway has been uttered or recognized (just "alfa" or "one three right"). This functionality was working for three remote airports at the same time. The ABSR output was then sent to the prototypic electronic flight strip system for callsign highlighting and flight status changes (see Figure 1), command presentation with correction option, and displaying the complete STT and TTC output in the outside view.

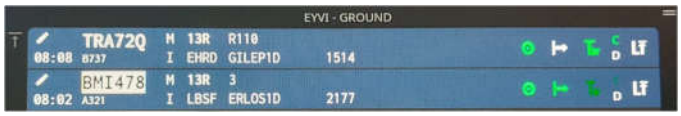

Figure 1. Callsign highlighting (white box around "BMI478") and automatically recognized flight status change from voice utterance (dark green "TAXI" via "C") on the electronic flight strip display's ground bay as well as earlier clearances (light green STARTUP, PUSHBACK).

B. NOR: Remote Tower ABSR Prototype

The ABSR system received a real-time transport protocol audio stream from the *Garex* voice communications control system. The same audio stream was also transmitted to the simulation pilots. The speech recognition module that based on a predefined US English language and acoustic model was responsible for processing the audio stream from the operator to identify any spoken words after the utterance ended. Relevant commands were then identified in the word sequences and were sent to the *InNOVA* system. The system's human machine interface (HMI) gave the operator a situational overview through the use of traffic situation displays and flight lists. Normally, the ATCo would have to manually enter any commands (such as clearances) into the system after communication with the pilot. When a command was recognized and received from the ASRU module, the relevant callsign was highlighted, and the command input was automatically applied in *InNova* if not rejected by the ATCo (see Figure 2). Up to 21 different command types have been covered for this environment.

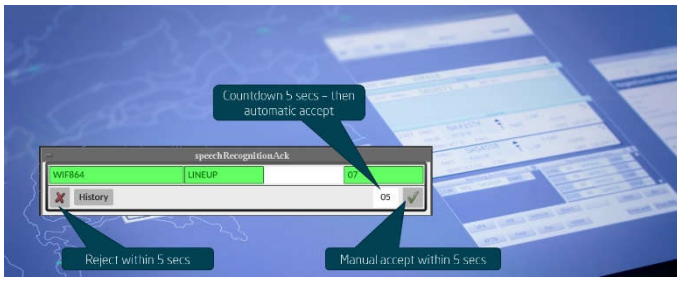

Figure 2. Callsign highlighting (green box around "WIF864") and recognized flight status change (green "LINEUP" for green runway "07") in a speech recognition acknowledgement window.

C. BUL: Tower and Ground ABSR Prototype

The ABSR system received the audio input via push-to-talk functionality and *Mumble* software with the ABSR component. The ABSR could also be activated without forwarding the utterance to the simulation pilots, e.g., if only certain callsign highlighting was intended. The ABSR prototype was integrated into the *Lead In Sky* product and infrastructure (see Figure 3). It was based on a Java widget for audio acquisition, coding and transmission, on widely available generic English language corpora, on KALDI for the STT conversion after an utterance was finally spoken, and on a custom-built set of python-coded blocks for the TTC extractor. The ATC concept extraction of the tested implementation covered six different command types due to limitations of the simulation platform. More types and qualifiers are already implemented, but were not part of this validation trials. The TTC conversion supported one command per utterance and relied on ATCos sticking to ICAO phraseology. The recognition phonetic model was trained offline and prior to simulations.

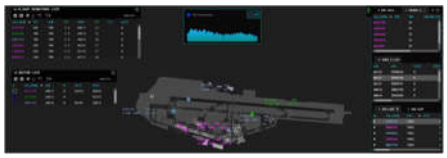

Figure 3. Speech recognition recording window (top middle) with airport surface map and flight information windows.

Model training was carried out via use of publicly available corpora and a number of sessions with ATCos based on a reduced version of the ATC annotation ontology for a total of about 400 utterances equating to 25 minutes of actual speech. Context-based data consisted of a list of active callsigns obtained from the *Lead In Sky* flight data processor and its communication infrastructure. The list was constantly updated in order to improve recognition performance. The ABSR output was presented to the ATCo in terms of highlighted callsigns and recognized commands in the HMI.

IV. Validation Exercises with Tower ATCos

This section explains the validation setup, simulation run conditions, and participating ATCos of LIT, NOR, and BUL. The envisaged questionnaires for all exercises covered the topics workload, situation awareness, acceptance/job satisfaction, usability, callsign and command recognition performance, interaction, and overall ASRU feedback.

A. LIT: Multiple Remote Tower and Ground Environment

The hardware setup of the LIT validation exercise is shown in Figure 4. It consists of three horizontal rows of monitors for the outside view, three radar screens, and the electronic flight strip system with one column for each of the three airports. The validation took a full day for each of the five Lithuanian and five Austrian tower ATCos, i.e., 10 days in total. The briefing was followed by a one-hour training run to familiarize the ATCos with both simulation conditions: baseline without any ABSR support and solution with ABSR support. 50% of the ATCos started with the baseline run and 50% started with the solution run to consider trainings effects. In the baseline condition, the ATCo needed to enter all commands with an electronic pen.

In the solution condition, the ABSR system automatically entered the recognized commands and the ATCo only had to correct, if something was wrong. After ten seconds without any corrections, the dark green command highlighting turned to light green as in the manual input condition and the command was accepted by the ATC system. The ATCos should use the same phraseology including deviations from ICAO phraseology as they use in their daily life. The two simulation runs with both simulation conditions took one hour each and were followed by various questionnaires and a debriefing. The validation setup is also visualized in [32].

Figure 4. Multiple remote tower and ground environment at DLR Braunschweig's Remote Tower Lab (LIT).

All ATCo utterances with their timings, automatic transcriptions, and automatic annotations have been recorded along with the questionnaire answers. All automatic transcriptions and automatic annotations have been manually checked and corrected by ATC-ASRU experts in order to calculate error rates and recognition rates on word level and on semantic level.

B. NOR: Remote Tower Environment

The NOR validation was performed in a remote tower simulator setting with six Hungarian ATCos (see Figure 5) as also visualized in [33]. Each ATCo performed three simulation runs in a single remote tower environment with a different remote airport in all three runs. However, the ABSR prototype was optimized for laboratory use for the technical validation, i.e., to be used with a predefined set of clearances. ATCos were requested to stick to the rather simple phraseology to make sure that the system can update the flight strips according to the recognized speech.

Figure 5. Remote tower environment of Indra (NOR).

The baseline runs without any ABSR system were conducted earlier in a different multiple remote tower environment. Measures of workload and situation awareness given a baseline-solution comparison were therefore limited. To be able to analyze the performance of the ABSR prototype after the validation, several key parameters were recorded: (1) Timing of ATCo voice transmission, ABSR processing time, HMI update and operator input, (2) recognized callsign and command, (3) the ABSR's confidence level of the recognition, and (4) whether the ATCo accepted or rejected the recognized callsign and command. In addition, the ATCo voice transmissions were recorded, and a manual check of the automatic word-by-word transcription was performed to support post-exercise analysis.

C. BUL: Tower and Ground Environment

The BUL validation in a simulated tower and ground environment as shown in Figure 6 was run with six ATCos from different parts of Italy [34]. Two ATCos were active at the same time: one as tower ATCo, the other as ground ATCo.

Figure 6. Simulated ground and tower environment at Leonardo Roma facilities (BUL).

A complete training day was held before the validations. In total, there were three validation days with four 45 minutes simulation runs per day: a baseline and three solution runs – with all simulation run conditions always remaining in the same order leading to limited interpretability.

Briefings, questionnaires, and debriefings were conducted accordingly. Captions were made of screens, audio, and instructed commands in order to generate annotated transcripts and command logs.

V. Results of Validation Exercises

A. Speech Recognition and Understanding Performance

The following tables summarize the quality of speech recognition and speech understanding performance. The reported numbers for LIT always contain baseline and solution runs with evaluation of recorded audio files, i.e., the numbers for offline recognition were evaluated on recorded files after the trials, the numbers for online recognition as shown in table footnotes were evaluated during the trials. NOR and BUL only report on utterances from the solution runs. The word error rate (WER) of the STT process, the number of words as well as number of utterances are shown in TABLE II.

TABLE II. Overview of Speech Recognition Performance

	LIT	*NOR*	*BUL*
Word Error Rate	5.1% [a]	-	16.9 %
# Words uttered	38,820	-	3,632
# Utterances	2,427	934	454

a. When only analyzing the 1,232 utterances of the ten solution runs, the WER was just 4.4% (offline recognition mode). In online recognition mode the WER was 13.6% (9.8% for solution runs).

WER is not calculated for NOR. These validation trials concentrated on more important metrics in the ATC context: callsign recognition performance, i.e., recognition, error, and rejection rates of the TTC process as shown in TABLE III.

TABLE III. Overview of Callsign Understanding Performance

Callsign Recognition Performance [b]	*LIT*	*NOR*	*BUL*
Callsign Recognition Rate	98.4% [c]	81.2%	89.8%
Callsign Error Rate	0.9%	7.8%	10.2%
Callsign Rejection Rate	0.7%	11.0%	

b. Restricted comparability of results due to different technical capabilities and ATCo briefings.

c. In online recognition mode, the recognition rate was just 91.7%, i.e., 89.1% in baseline runs without ASRU and 94.2% in solution runs with ASRU support.

The overall TTC rates in terms of the speech understanding process for complete ATC commands – core result for the ASRU performance – are shown in TABLE IV. This table lists the recognition/error/rejection rate on command level, as well as a recognition rate only for the command type, the number of commands, and the number of commands per utterance.

TABLE IV. Overview of Command Understanding Performance

Command Recognition Performance	*LIT*	*NOR*	*BUL*
Command Recognition Rate	91.4% [d]	76.0% [e]	64.6% [f]
Command Error Rate	4.5%	-	5.1%
Command Rejection Rate	5.0%	-	35.4%
Command Type Recognition Rate	94.0%	93.2%	75.9%
# Commands Uttered	7,560	993	454
# Commands per Utterance	3.11	1.06	1.00

d. For LIT, a command is only considered correct if callsign, command type and command second type, value(s), unit, qualifier, and conditions were correct. When considering only the relevant command types that appeared at least 25 times during the validation, the command recognition rate reached 94.6% with a command recognition error rate of 3.5%. In online recognition mode, the recognition rate for all utterances was just 82.9% in solution runs with ASRU. However, the online recognition rate for commands that have also been displayed in the electronic flight strips was 89.4%.

e. For NOR, a command was considered correct if callsign, command type and command second type, and value were correct, i.e., unit, qualifier, and condition were not considered.

f. For BUL, a command was considered correct if callsign, command type and qualifier were correct.

B. ATCos' Opinion on ABSR Performance

For the LIT prototype, ATCos rated the recognition of callsigns as almost perfect with a mean value of around 9 out of 10 (see Figure 7). Considering only male ATCos or only the Austrian (ACG) ATCos led to mean ratings of close to scale maximum value 10. The recognition performance of ATC commands was perceived as good with a mean value of around 7. The general quality of information presentation from ABSR was also 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 ABSR information displayed were rated much higher by the Lithuanian ATCos as compared to the Austrian ATCos.

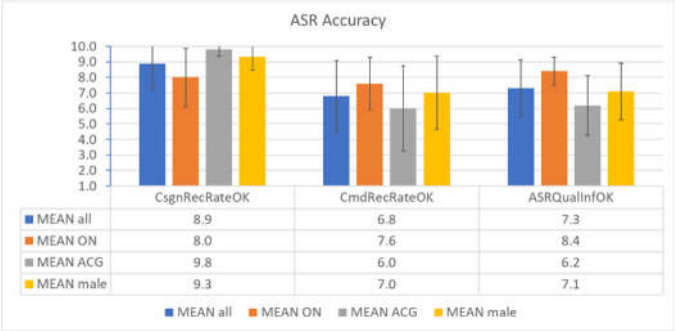

Figure 7. Feedback on the perceived ABSR performance for understanding callsigns, commands, and level of displayed information (LIT).

For the NOR prototype, Figure 8 shows ATCos' perception of successful 1) callsign, 2) clearances and 3) other parameter understanding of a reduced set of commands and qualifiers, in accordance. According to the results, the system seemed to perform best with the callsign recognition (67% positive feedback), and the other parameters (67% positive feedback), followed by the clearance recognition (56% positive feedback).

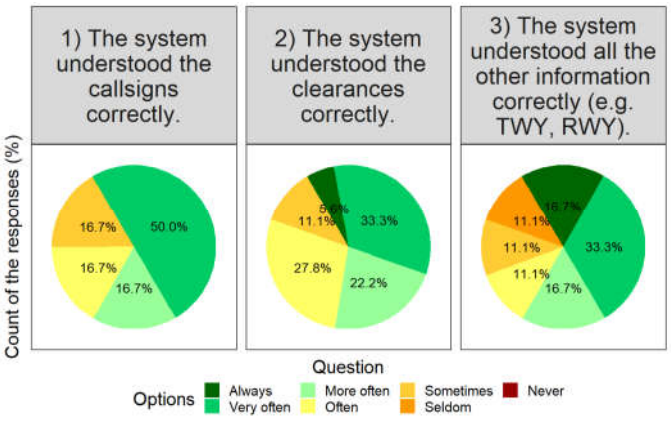

Figure 8. Feedback on the perceived ABSR performance for understanding callsigns, clearances, and other parameters (NOR).

It is important to highlight that practicing, experimenting and challenging the system by pronouncing a certain callsign differently or saying the wrong callsign intentionally happened often. These cases decrease the reliability of the results.

For the BUL prototype, the perceived performance of callsign and command recognition and rejection rates was at least in the neutral range, but predominantly in the positive scale range (see Figure 9).

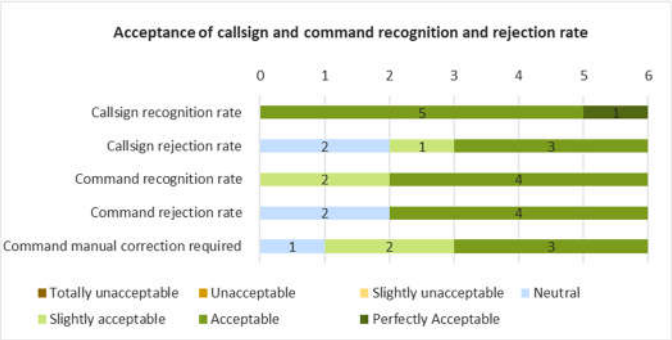

Figure 9. Feedback on the perceived ABSR performance with recognition and rejection rate of callsigns and commands (BUL).

C. Latency of ABSR Output

Basically, the latency of the ABSR prototypes was perceived as acceptable, but improvable. For the LIT prototype, ATCos rated the statement about timeliness of the ASR tool to be within acceptable limits with 7.5 on a scale from 1 to 10 (with 10=fully agree; 1=fully disagree). However, checking the ABSR output in the electronic flight strip display even with automatic acceptance after ten seconds slowed some ATCos down, because in the baseline runs, ATCos made entries simultaneously while speaking. Some ATCos judged the speed of ABSR output while speaking as sufficient, two ATCos wanted to have faster output.

For the NOR prototype, 50% of ATCos gave positive feedback on the latency of ABSR output. ATCos also emphasized that the system must become faster, because they do not want to continuously check the system during work. For the BUL prototype with limited ASRU functionality, all ATCos found the latency of the ABSR system to be acceptable, i.e., rated the statement within the positive answer scale half.

D. Differences in ATCo Utterances

The used phraseology in practice and pronunciation of words seem to be different if ATCos know about an ASRU component supporting them. For the LIT prototype, ATCos were asked if they spoke differently in baseline and solution condition (in the latter being supported by ABSR). Three ATCos stated that they spoke less carefully in baseline runs, because only pilots needed to understand them. Similarly, two other ATCos said, they spoke closer to ICAO phraseology in solution runs as they were better supported then. This corresponds to the measured TTC performance which was 5% better in solution runs compared to baseline runs as reported in the footnotes of TABLE III. One study participant stated that ATCos automatically become more phraseology conform and that this is one of the greatest advantages of such an ASRU technology for safety already. However, some other ATCos stated that there was no difference in their speaking behavior.

For the NOR prototype, five of six ATCos confirmed that they sometimes changed their pronunciation to make sure that the ABSR system will better understand uttered commands. For the BUL prototype, ATCos recommended not to use two keys for ABSR activation and for frequency communication, because this induced a further possible source of error, thus negatively affecting the recognition performance. ATCos also motivated to enhance the phraseology that is potentially being recognized by the ABSR system, which implicitly includes that ATCos also adapted their speaking behavior during the validation exercises. Hence, all ABSR prototypes showed the tendency to influence the style of ATCos' utterances in a positive way regardless of any recognition rate.

E. Feedback on Human Machine Interfaces

The system usability score (SUS[1]) for the LIT prototype was 75, thus slightly better than the rating for the baseline system. The ten seconds-highlighting with automatic acceptance and the highlighting colors were liked most. Also, all ATCos of the BUL exercise gave feedback in the positive range regarding the system usability. For the NOR prototype, HMI usability (e.g., click on accept/reject) was regarded as user-friendly by the majority of the ATCos.

F. ATCo Workload and Situation Awareness

For the LIT and BUL exercise, some evidences were collected that such an ABSR system can contribute to a reduction of ATCo workload especially in higher TRLs as also mentioned by ATCos of the NOR exercise. According to the Bedford workload scale[1] that ATCos rated once after each simulation run, there was no significant difference between the baseline and solution runs of LIT and BUL – for LIT, the solution rating was even slightly worse than for the baseline. However, for the instantaneous self-assessment of workload (ISA[1]) for LIT that was taken every five minutes during all simulation runs, the mental workload average slightly decreased in the solution runs compared to the baseline where there was no ABSR support. ABSR in the LIT prototype seemed 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 with 10 as the best value. Also, for BUL, all ATCOs rated situation awareness as either 'high' or 'perfect' during solution runs. However, all air traffic scenarios had a rather low to medium traffic density, i.e., the workload levels were at a low level anyway. In the NOR exercise ATCos felt that during the baseline run they had the opportunity for "self-checking" their own input into the system. However, in the solution run, the feeling of checking themselves was lost as the system took over the input after they provided the clearance or instruction. Also, in the LIT exercise, there was a tendency to over-rely on automation when the ABSR worked fine. Hence, just as with many other highly automated solutions there might be positive as well as negative effects on situation awareness to be analyzed deeper. The

[1] For further information on the human performance questionnaires, check https://ext.eurocontrol.int/ehp/

callsign highlighting function was much appreciated by ATCos of all three ABSR prototypes also to support situation awareness and reduce mental workload. The reduced search time for items in solution runs compared to the baseline was confirmed by the LIT ratings.

G. ATCos Ratings on Usefulness, Confidence & Acceptance

ATCos rated if the system was useful within the SATI[1] (SHAPE Automation Trust Index) questionnaire. For comparison reasons, the answers like "never", "sometimes", "often", "always" have been converted to a numbered interval scale. For LIT, the confirmation ratio increases by 12% absolute when comparing baseline (58%) to solution runs (70%). For BUL, the confirmation ratio increases by 5% absolute when comparing baseline (75%) to solution (80%). For NOR, there is just the value of 47% available for solution. Hence, it is neither reasonable to directly compare the three results for the solution runs with the prototypes, nor to compare the baselines. The confidence in and the hypothetical acceptance of such a system in their daily-life controller working positions was answered rather differently by the ATCos for the different ABSR prototypes. For the LIT prototype, 80% of ATCos stated with 8 or more points on the scale from 1 to 10 that they would appreciate such an ABSR system in their normal workplace. For the NOR prototype, only one third of the ATCos stated that they could confidently work with the ABSR prototype. For the basic BUL prototype, the overall acceptance was predominantly rated with the second-best option "high" on a seven-option scale.

VI. Summary and Outlook

A. Summary of SESAR2020 Solution 97.2 Validation Results

The technology of assistant based speech recognition and understanding has shown to be also feasible in an ATC environment for (multiple remote) tower and ground ATCo working positions in a laboratory environment. However, a list of recommendations on how to enhance aspects of the ABSR system have been identified. Very promising recognition rates for callsigns of 98% and for frequent commands of 94% with error rates for callsigns of below 1% and for commands below 5% are possible to achieve in case of mature implementation at least in the offline mode even with very low amount of tower specific training data (LIT). However, the tested alternative prototypes (NOR, BUL) achieved callsign recognition rates of 65% and command recognition rates of 81% as worst result. Hence, there is still improvement needed to achieve command recognition rates beyond 95% and command recognition error rates below 2.5% which have shown to significantly reduce ATCo workload in the en-route and approach environment. In general, ATCos were able to perform their ATC tasks when working with ABSR support. The positive results for system usability, acceptance and some workload measurements on a low workload level environment show the potential of ABSR in a (multiple remote) tower/ground environment – even if a row of other measurements do not show any significant differences between baseline and solution. The recorded data show that ATCos speak differently, i.e., closer to phraseology if being supported by ABSR (i.e., solution runs have higher command recognition rates than baseline runs; in the latter, the speech was analyzed as well, but the output was not shown to the ATCo). On the one hand, this might be because they get better support if recognition rates are higher, on the other hand, it might be due to the pure awareness of working with speech recognition in the background. If ATCos stick closer to the ICAO phraseology just due to the pure presence of an ABSR system, that could already be a safety feature. To summarize, the three validation exercises have shown potential of using ABSR system output. However, they also revealed relevant aspects to be considered for further enhancements. The quantitative and qualitative feedback of ATCos was motivating to go beyond technology readiness level 4 to offer the full potential of ABSR support to them.

B. Outlook on Future Work

The three described ABSR prototypes have identified different needs for improvements on different maturity levels. BUL needs to incorporate more context-based data, better train the phonetic models, and enlarge the coverage regarding different command types and elements of the ATC ontology. NOR needs to cover conditional and more non-standard clearances as well as improvement of callsign, command, and command type recognition error rates, because these indicators are amongst the most important ones from a safety point of view. LIT needs to improve the online recognition rates to close the gap to the better offline recognition performance. All prototypes should step closer to an operational environment to benefit from the potential of ASRU. Acceptable recognition and recognition error rates for the relevant ATC concepts need to be defined in risk assessment sessions. The amount of training data must be further increased in order to achieve better recognition rates given representative samples. Furthermore, a large amount of data must be recorded from operations rooms (not from labs), because ATCos speak differently in simulations. Finally, the recording configuration for training and validation or even operational use should be the same, e.g., same recording frequency, similar background noise, identical microphones, in both recordings with a face mask or in both without, etc. Further validations should also consider different workload levels of ATCos and the effect on ASRU performance. The European-wide agreed ontology for annotation of ATC utterances as used and enhanced in these three validation exercises should be further exploited and especially fully implemented by all ABSR prototypes to ease comparison of performance. The continuous mutual enhancements of the ontology in current European ASR projects build a basis for interoperability of systems. ASRU should be further developed as a potential on-the-job-training support, a help for incident analysis, implement callsign highlighting for initial pilot calls, and further safety net functions such as readback error detection (as already sketched for other ATC domains in section II.A) or plausibility checking of communication content.

Acknowledgment

PJ.05-W2-97.2 Automatic Speech Recognition solution has received funding from the SESAR Joint Undertaking under the European Union's grant agreement No. 874470. We like to thank the five tower ATCos from ON, five ATCos from ACG, six ATCos from HungaroControl, and six ATCos from ENAV for their participation in the three validation exercises.

REFERENCES

[1] C. Mertz, S. Chatty, and J.-L. Vinot, "Pushing the limits of ATC user interface design beyond S&M interaction: The DigiStrips experience," in 3rd USA/Europe Air Traffic Management Research and Development Seminar (ATM2000), Naples, Italy, 2000.

[2] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing controller workload with automatic speech recognition," in IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016.

[3] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, and M. Schulder, "Assistant-based speech recognition for ATM applications," in 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[4] M. Kleinert, H. Helmke, G. Siol, H. Ehr, M. Finke, A. Srinivasamurthy, and Y. Oualil, "Machine Learning of Controller Command Prediction Models from Recorded Radar Data and Controller Speech Utterances," in 7th SESAR Innovation Days, Belgrade, Serbia, 2017.

[5] H. Helmke, M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr, H. Arilíusson, T. Simiganoschi, A. Prasad, P. Motlicek, K. Veselý, K. Ondřej, P. Smrz, J. Harfmann, and C. Windisch, "Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety," in 14th USA/Europe ATM Research and Development Seminar, Virtual, 2021.

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

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

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

[9] O. Ohneiser, H. Helmke, S. Shetty, M. Kleinert, H. Ehr, S. Murauskas, and T. Pagirys, "Prediction and Extraction of Tower Controller Commands for Speech Recognition Applications," in Journal of Air Transport Management, 95, Elsevier, 2021.

[10] The Project PJ05-W2 DTT, https://www.remote-tower.eu/wp/project-pj05-w2/solution-97-2/, n.d.

[11] D. Schäfer, "Context-sensitive Speech Recognition in the Air Traffic Control Simulation," Eurocontrol EEC Note No. 02/2001, PhD Thesis of the University of Armed Forces, Munich, 2001.

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

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

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

[15] H. Helmke, M. Kleinert, J. Rataj, P. Motlicek, D. Klakow, C. Kern, and P. Hlousek, "Cost Reductions Enabled by Machine Learning in ATM - How can Automatic Speech Recognition enrich human operators' performance?," in 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria, 2019.

[16] O. Ohneiser, H. Helmke, M. Kleinert, G. Siol, H. Ehr, S. Hobein, A.-V. Predescu, and J. Bauer, "Tower Controller Command Prediction for Future Speech Recognition Applications," in 9th SESAR Innovation Days, Athens, Greece, 2019.

[17] O. Ohneiser, S. Sarfjoo, H. Helmke, S. Shetty, P. Motlicek, M. Kleinert, H. Ehr, and S. Murauskas, "Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances," in 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czechia, Hybrid, 1, pp. 266-270, 2021.

[18] H. Helmke, M. Kleinert, O. Ohneiser, H. Ehr, and S. Shetty, "Machine learning of air traffic controller command extraction models for speech recognition applications," in IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), Virtual, 2020.

[19] S. Shetty, H. Helmke, M. Kleinert, and O. Ohneiser, "Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload," in 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022), New York, NY, USA, Hybrid, 2022.

[20] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, "The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection," in 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019), Graz, Austria, 2019.

[21] S. Chen, H.D. Kopald, B. Avijan, and M. Fronzak, "Automatic Pilot Report Extraction from Radio Communications," in IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 2022.

[22] S. Chen, H.D. Kopald, A. Elessawy, Z. Levonian, and R.M. Tarakan, "Speech Inputs to Surface Safety Logic Systems," in IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czechia, 2015.

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

[24] The project HAAWAII® (Highly Automated Air Traffic Controller Workstations with Artificial Intelligence Integration), https://www.haawaii.de, n.d.

[25] H. Helmke, K. Ondřej, S. Shetty, H. Arilíusson, T.S. Simiganoschi, M. Kleinert, O. Ohneiser, H. Ehr, J.-P. Zuluaga, and 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.

[26] A. Prasad, J. Zuluaga-Gomez, P. Motlicek, S. Sarfjoo, I. Nigmatulina, O. Ohneiser, and H. Helmke, "Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition," in 12th SESAR Innovation Days, Budapest, Hungary, 2022.

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

[28] S. Shetty, O. Ohneiser, F. Grezl, H. Helmke, and P. Motlicek, "Transcription and Annotation Handbook, Excerpt (Transcription Rules)," HAAWAII D3.1, https://www.haawaii.de/wp/wp-content/uploads/2020/11/HAAWAII-D3-1-Transcription_And_Annotation_Handbook-v1-00_Excerpt.pdf, 2020.

[29] 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), Virtual, 2021.

[30] Y. Lin, "Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application," in Aerospace 8, No. 3: 65, 2021.

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

[32] DLR, "Speech recognition for air traffic control towers," https://www.youtube.com/watch?v=Y76kQmo_ANU, 2022.

[33] Indra, "Automatic Speech Recognition for air traffic control," https://www.youtube.com/watch?v=SbJEyaJj2UQ, 2022.

[34] Leonardo, "SESAR: Leonardo's contribution to the future of Air Traffic Management," https://www.linkedin.com/feed/update/urn:li:activity:6932276745387528192?updateEntityUrn=urn%3Ali%3Afs_feedUpdate%3A%28V2%2Curn%3Ali%3Aactivity%3A6932276745387528192%29, 2022.