Tower Controller Command Prediction for Future Speech Recognition Applications

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

Andrei-Vlad Predescu
Politehnica University of Bucharest,
„Elie Carafoli" Department of
Aerospace Science,
Splaiul Independenței 313,
București 060042, Romania

Jakob Bauer
FH Joanneum –
University of Applied Sciences,
Department Engineering,
Aviation,
Alte Poststraße 149,
8020 Graz, Austria

***Abstract*—Air traffic controllers' (ATCos) workload often is a limiting factor for air traffic capacity. Thus, electronic support systems intend to reduce ATCos' workload. Automatic Speech Recognition (ASR) can extract controller command elements from verbal clearances to deliver automatic air traffic control (ATC) system input and avoiding manual input. Assistant Based Speech Recognition (ABSR) systems with high command recognition rates and low error rates have proven to dramatically reduce ATCos' workload and increase capacity as an effect. However, those ABSR systems need accurate hypotheses about expected commands to achieve the necessary performance. Based on the experience with an ATC approach hypotheses generator, a prototypic tower command hypotheses generator (TCHG) was developed to face current and future challenges in the aerodrome environment. Two human-in-the-loop multiple remote tower simulation studies were performed with 13 ATCos from Hungary and Lithuania at DLR Braunschweig. Almost 40 hours of speech with corresponding radar data were recorded for training of the TCHG prediction models in 2017/2018. More than 45 hours of speech and radar data comprising roughly 4,600 voice utterances were recorded in the second simulation campaign for the TCHG evaluation test end of 2018. The TCHG showed operational feasibility with a sufficiently low command prediction error rate of down to 7.3% and low context portion predicted having a sufficiently fast command prediction frequency of once per 120ms to timely deliver the hypotheses to a speech recognition engine. Thus, the next step is to build an integrated ABSR system for the tower environment.**

Keywords—Air Traffic Controller; Tower Command Hypotheses Generator; Assistant Based Speech Recognition; Automatic Speech Recognition; PJ.16-04; Multiple Remote Tower

I. INTRODUCTION

Air traffic control (ATC) systems are commonly limited by their capacity, as their goal is to balance demand and capacity for an optimized overall performance. Air traffic controllers' (ATCo) workload is a limiting factor to increase this overall system capacity.

Nowadays, ATCos' workload is increasing due to the growing number of worldwide flights every year. This can limit the number of aircraft handled per sector respectively per ATCo in the future. ATCos issue clearances via voice and radio communication to aircraft pilots for controlling all relevant flights under responsibility.

The flight crew is expected to confirm the clearance by a readback or acknowledge the information – this means instant feedback to the ATCo. For their effective operation, ATC systems need accurate data in timely manner as well. The issued clearances are one of the relevant necessary input data for ATC systems. This input is done manually by the ATCo using the mouse or another input device through the interaction with an electronic flight strip. However, these input devices generate high workload for the ATCo. Thus, the necessary information for ATC system input is doubled as it already exists within the voice clearance in analogue format. Automatic speech recognition (ASR) can support ATCos by extracting spoken concepts of issued clearances and automatically feeding the digital ATC system with them.

The digitized concepts, i.e. effective meaning of commands, can be used as input for further ATC support functionalities and may lead to reduced workload and increased safety. However, this requires high reliability in the automatic speech recognition and extraction of clearance elements. Hypotheses about the content of the controller utterance tremendously help the speech recognition engine to choose from a reduced set of possible contents. Such controller command hypotheses integrated into an assistant based speech recognition system (ABSR) already proved to dramatically decrease the command recognition error rate and increase the command recognition rate for the approach area [1].

One chain of effects succeeding high controller command recognition rates starts with a reduction of ATCo workload to enter clearances, resulting in more timely and accurate commands. This, in turn, can already be a safety gain and can further lead to shorter flight routes and shorter flight times that go along with reduced fuel consumption and carbon dioxide emissions. But already the visualization of clearance elements for better awareness and further tracking of conform aircraft trajectory changes can overcome controller-pilot communication problems and increase safety.

To achieve these possible benefits also in the aerodrome ATC environment, a tower command hypotheses generator (TCHG) is developed to predict commands for usage in a later ABSR system. More precisely, a multiple remote tower simulation is used to compare automatically generated tower command hypotheses with actual given controller commands.

This paper consists of related work with respect to speech recognition and command hypotheses in chapter II. Chapter III outlines the concept for a tower command hypotheses generator. The human-in-the-loop study setup for data recording and some implications for machine learning are explained in Chapter IV. Chapter V presents the results regarding quality of command hypotheses. Chapter VI summarizes, concludes, and gives an outlook on future work.

II. Related Work on Speech Recognition and Controller Command Hypotheses

A. History of Speech Recognition in ATC

ASR systems convert spoken words into machine-usable digital data and thus serve as an alternative input modality. Today, voice recognition is used in various areas of human life such as navigation systems or smartphone applications. Already three decades ago first ASR systems were developed [2], [3] and integrated for ATC training [4]. Years later, this led to replacing simulation pilots and to enhanced simulator infrastructure (e.g., DLR [5], MITRE [6], FAA [7], and DFS [8]). ASR also supports to improve safety, e.g. for closed runway incursions [9] or pilot readback errors [10], and to perform ATCo workload assessment [11], [12].

However, an ABSR system [13] can also significantly reduce ATCos' workload as shown in the projects AcListant® and AcListant®-Strips [14]. In addition, air traffic management efficiency can be increased with fuel savings of 50 to 65 liters per flight [15]. DLR and its speech recognition partner Saarland University used KALDI as the ASR platform. DLR developed a hypotheses generator making assumptions about the next possible controller commands [1]. If e.g. an arriving aircraft is in FL100, it is more likely that the ATCo will issue a descent to FL80 than a climb to FL140 or even a non-reasonable descent to FL140. The most probable hypotheses are sent to the speech recognition engine to reduce its search space and improve recognition quality. Command recognition error rates (CRER) below 1.7% were achieved [1].

One main issue to transfer ABSR from the laboratory to the operational systems is the costs of deployment, because modern speech recognition models require manual adaptation to local requirements and environments (language accents, phraseology deviations, environmental constraints etc.) [16]. AcListant® needed more than 1 Mio € for development and validation for Düsseldorf approach area.

The SESAR exploratory research project MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance) proposes a general, cheap and effective solution to automate this re-learning, adaptation and customization process by automatically learning local speech recognition and controller models from radar and speech data recordings [17].

Command prediction error rates (CPER) of 3.2% for Vienna and 2.3% for Prague approach were achieved in MALORCA with corresponding CRERs of 3.2% respectively 0.6% [18]. Those low CRERs were reached, because of using command hypotheses and plausibility checking components as they reduce the CRER by roughly 12% for Vienna respectively 6% for Prague [17].

In another setup, CPERs of 4.8% for Vienna approach respectively 0.3% for Prague approach were measured [19]. Again, the use of an assistant system for ASR dramatically decreases the CRER while only slightly decreasing the command recognition rate [19].

B. Controller Command Hypotheses Formats

A necessary step for the evaluation of command hypotheses is to extract the concepts of actual given commands. Thus, ATCo utterances need to be manually transcribed and annotated. Transcription is defined as the word-by-word equivalent of the verbal utterance, e.g. "good morning lufthansa two alpha altitude four thousand feet reduce one eight zero knots or less turn left heading two six zero".

The annotation is defined as the meaning of utterances. Therefore, a set of rules – an ontology – has been developed. In the AcListant® project [20] a first version of an ontology was created, which consists of four elements: 1) callsign, 2) command type, 3) commanded value, and 4) unit [21], [22] with mandatory and optional elements. The example above is annotated as "DLH2A ALTITUDE 4000 ALT DLH2A REDUCE_OR_BELOW 180 DLH2A TURN_LEFT_ HEADING 260". More than 30 command types were supported. The approach reaches its limits in the MALORCA project [23], [24] when it was extended for command annotation for live traffic for Vienna and Prague approach [25] also including departure and overflight traffic. More command types were needed (e.g. QNH, INFORMATION, REPORT_SPEED, EXPECT_RUNWAY) and the necessity to handle conditional clearances occurred [26]. For the tower environment, even controller command types used only on ground such as PUSHBACK, TAXI, and LINEUP need to be considered.

Therefore, all major European air traffic management (ATM) system providers and European air navigation service providers defined a common enhanced ontology suggested and coordinated by DLR [27]. This ontology encompassed approach, en-route as well as tower clearances. It can be used for annotation of ATCo and pilot utterances as well as for command predictions. In the above example, the new annotation would be "DLH2A ALTITUDE 4000 ft DLH2A REDUCE 180 kt OR_LESS DLH2A HEADING 260 LEFT".

The next section describes the concept, evaluation metrics, and implementation of command predictions for the tower environment that was implemented for the first time.

III. Tower Command Hypotheses Generator Concept, Implementation and Validation Goals

The TCHG is a new system developed by DLR using experience of former projects regarding hypotheses information generation in the approach area. However, the command types used in tower environment are different from those of approach controllers of former projects. Additionally, the tower area comprises much more command types than the implementation for approach. The TCHG predicts possible tower controller commands for the near future taking into account surveillance data (e.g. radar data, flight plans, and meteorological data). This prediction is not a single forecasted command, but a set of possible commands (context).

Examples for single predictions in different situations according to the defined ontology are: “AEE2019 STARTUP”, “BAW123 PUSHBACK”, “AFR456 TAXI VIA A”, or “DLH789 CLEARED TAKEOFF RW13R”. Furthermore, this is done for three remote airports at the same time. Therefore, there were three geographical regions defined to forecast commands with respect to aircraft within those regions. One global geographic area covers the airspace between and around the airports, e.g. to predict commands for flights that fly from one to another of those three airports.

The technical validation plan for the TCHG evaluation foresaw two objectives and three criteria goals. The first objective was to assess the stability of the (ASR) system performance. The second objective was to assess the operational feasibility of the integration of the (ASR) system and its sub-systems into operations. Furthermore, three target numbers regarding the prediction quality should be reached. The relevant numbers are the command prediction error rate (CPER) with its standard deviation (SD) and the context prediction time (CPT).

CPER is defined as the number of given controller commands that were not forecasted divided by the number of all given controller commands per run. Or in other words: Number of actually given commands by the controller that are not part of the set of predicted command hypotheses divided by the total number of commands actually given by the controller. The lower the CPER the better, as an ABSR system can rely on accurate forecasts to only falsely reject as few commands as possible due to stated non-conformity to the context. The average CPER should be below 10% with a standard deviation of less than 2.5%. It was assumed that the CPER for a first tower command prediction – particularly due to a greater variety of commands than in the approach environment – will be slightly higher than CPERs of already advanced approach command predictions. The CPT should be below five seconds to enable a prediction at least for each radar data update cycle.

Besides, another metric will be assessed – the context portion predicted (CPP). CPP is defined as the number of forecasted commands divided by the total number of commands, which an ATCo theoretically could give. Multiple hundreds of commands are possible per aircraft (commands for speed, altitude, direction, ground clearances, etc. with reasonable values). Or in other words: Total number of predicted commands divided by the number of commands which are theoretically modelled and possible, e.g. the number of predicted headings commands is normally in the range of 10-40 per callsign, the total number in our model is 144 (005, 010, 015, … 355, 360 multiplied by two because the qualifiers LEFT and RIGHT are possible). The total number of heading commands is even higher, i.e. 720, if heading commands of step size one are considered. The lower the CPP, the better, because a lower number of command hypotheses for the ABSR system helps to faster choose the best fitting command hypotheses for a given utterance and to increase the command recognition rate in case of correct forecasts. After implementation and integration of the TCHG in a multiple remote tower environment, the prediction quality was assessed using data from a series of four successive human-in-the-loop studies.

IV. Human-in-the-Loop Study with Tower Command Hypotheses Generator

A. Validation Setup and Simulation Run Conditions

The SESAR2020 industrial research project PJ.16-04 CWP HMI (Controller Working Position Human Machine Interface) contained – amongst others – a validation exercise for the TCHG. The PJ.16-04 ASR exercise 240 “Controller Command Prediction for Remote Tower Environment” was hosted at DLR’s Multiple Remote Tower Experimental Setup in Braunschweig, Germany. The series of four human-in-the-loop studies to evaluate the TCHG prototype took place in 2017 and 2018. The ATCos – as study subjects – had three rows of monitors presenting the camera image of the respective three airports and a head-down ATM system unit to monitor and safely influence the given traffic (see figure 1).

Figure 1. Multiple Remote Tower Environment at DLR Braunschweig.

However, the command hypotheses did not influence the ATCo’s or controller support system’s active work. The simulated remote airports were run in parallel with different traffic. In the study with ATCos from HungaroControl (HC) airports were located in Hungary: Budapest (LHBP), Debrecen (LHDC), Papá (LHPA), for Oro Navigacija (ON) in Lithuania: Vilnius (EYVI), Kaunas (EYKA), Palanga (EYPA). Five respectively four different traffic scenarios have been used for Hungary and Lithuania. They comprised Instrument and Visual Flight Rules (IFR/VFR) traffic, but VFR traffic was never more than 20%. All simulation scenarios lasted 50 minutes and took place at day light conditions. The majority of traffic had to be controlled from the first listed tower (LHBP/EYVI). There were a few special situations that ATCos were faced with, e.g.

- four simultaneous movements, i.e. departure or arrival (2 at the main airport, 1 at the smaller airports each),
- VFR arrival and departure crossing,
- Remotely Piloted Aerial System (RPAS) in airspace,
- responsibility for ground movements.

The exercise was conducted jointly by DLR and HC respectively ON under the umbrella of solution PJ.05-02. This solution was responsible for the validation platform itself – without the TCHG – and the remote tower concept validation. The communication between ATCos and simulation pilots was done via radio telephony (Yada console) on three different frequencies. The resulting wav files with controller utterances and radar data were captured on a Linux laptop.

B. Data Recordings

Pre-trials with seven HC ATCos running four different air traffic scenarios took place from November 13 to 21, 2017. Pre-trials with six ON ATCos running also four different scenarios took place from March 19 to 27, 2018. Pre-trials were used to collect data to develop the models for command prediction of the TCHG. Final trials with seven HC ATCos running five different scenarios took place from November 12 to 22, 2018. Final trials with six ON ATCos running four different scenarios each took place from December 3 to 11, 2018. The data recordings of those trials were used for evaluation of TCHG prediction accuracy. With pre-trial data (before summer 2018), machine learning algorithms were implemented to improve the accuracy of command hypotheses.

The complete data set comprised 52 simulation runs with a duration of roughly 142'000 seconds (39.4h). This included roughly 4'700 voice utterances (wav files). 100% of the Hungarian and 30% of the Lithuanian tower utterances have manually been transcribed (speech to text), annotated (text, i.e. word sequences to ATC concepts and commands), and checked for this learning approach. This sums up to more than 3,400 transcription files and the same number of annotation files. When ignoring the "silence" between different wav-file occurrences, there were roughly 7 hours (26 h "with silence") of annotated tower commands available for learning.

The data resulting from the final trials (after summer 2018) was used to test and perform the evaluation of prediction accuracy. The reported portion of actually given annotated controller commands was compared to the tower command hypotheses from the trials. The complete data set comprised 59 simulation runs with a duration of roughly 164'000 seconds (45.6h). This included roughly 4'600 voice utterances. 25% of the Hungarian and 35% of the Lithuanian tower utterances have manually been transcribed, annotated, and checked for the testing and evaluation (9 simulation runs each). These data sums up to more than 1'000 transcription and annotation files each consisting of more than 2 hours (12 hours "with silence") of annotated tower commands available for testing.

C. Determining Parameters for Machine Learning

The determination of parameters for the machine learning algorithms is a pre-result that needs to be found first. Therefore, the methodology of how to find this result is shortly outlined in the following. The proportion of 80% training data and 20% test data that can roughly be used here is a very typical one in the machine learning community to proof the accuracy of the learned model.

As a first step, an appropriate window size for the machine learning approach needed to be found (for more background of the "window" use and the machine learning algorithms, refer to [24]). The "window" is a raster size (a certain rectangle in terms of latitude and longitude) and is used to cluster airspace areas where certain controller commands are given respectively expected (hypotheses). If the window size is huge, command types are predicted everywhere in the airspace. However, a lineup far away from an airport does not make sense. Furthermore, a speech recognition engine would receive too many hypotheses to choose from.

If the window size is small, valid command types might not be forecasted, e.g. because the aircraft position was just a few meters next to the forecast region being too small. Besides, a speech recognition engine would not receive an accurate set of hypotheses (context) including the actually given ones of the ATCo. Thus, a trade-off needs to be found for the window size. The window size is completely different to the "context size". The context size comprises all command predictions at a given time. The absolute context size indicating the number of predicted commands per controller utterance occasion is a very important parameter to the CPP and helps to find reasonable values for machine learning.

For determining the best window size, the Hungary-2017-11 data was used to train the command prediction model. This model was then used to test the Hungary-2017-11 data that were split into two halves (A/B). As this was done with all available data end of 2017 and is only a pre-result for applying the machine learning algorithms on later data, only Hungary-2017-11 data was used for determining parameters. Different window sizes from 1 to 14 were used for this test. The CPER and the context size (number of forecasted commands) should be as low as possible. However, big context size normally results in low CPERs and vice versa. Hence, it was decided to choose the window size parameter that does not show great differences in the results of the two aforementioned values compared to the parameter step before. The analysis result is shown for CPER in figure 2 and for context size in figure 3.

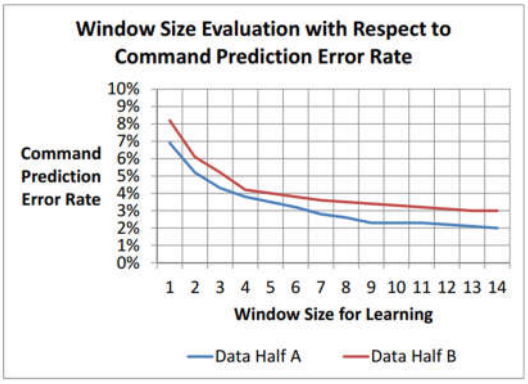

Figure 2. Comparison of command prediction error rates for different window sizes.

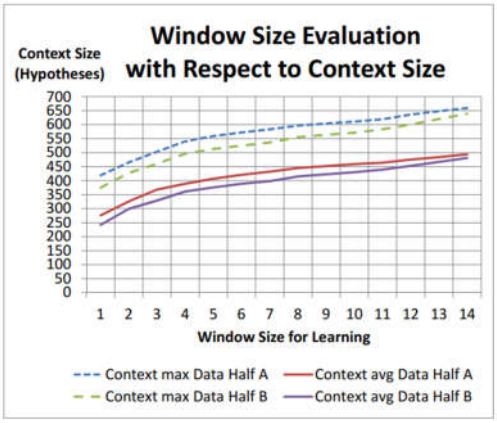

Figure 3. Comparison of context size for different window sizes.

ISSN 0770-1268

The CPER of window size 11 was 97% of the error rate of window size 10 for data half B. For data half A, the error rate from window size 10 to 11 only changed in the second value after the decimal. The context size of window size 11 was already at 99% respectively 98% of the context size at window size 12 and thus reached a sufficiently high number. Thus a window size of 11x11 was chosen as the best compromise between low error rates and low context size. Further increasing the window sizes only very slightly improves the CPER, but further increases the context size. The above reported window size is valid when analyzing all controller command types together.

However, there might be better fitting window sizes for single command types that have other characteristics with respect to airspace regions that are usually instructed by an ATCo. The top twenty commands actually given by ATCos are as follows in descending order (the command most often used (rank 1) appeared 1631 times, rank 20 appeared only 37 times in all Hungary-2017-11 data): INFORMATION (WINDDIRECTION, WINDSPEED, ATIS, TRAFFIC, ...), CLEARED (LANDING, TAKEOFF, TOUCH_GO, TO, VIA, etc.), INIT_RESPONSE, TAXI, CLIMB, SQUAWK, REPORT (FINAL, BASE, etc.), CONTACT_FREQUENCY, CONTACT, STARTUP, LINEUP, INFORMATION QNH, REPORT_MISCELLANEOUS, CONTINUE, VACATE (also with TO, VIA), CALL_YOU_BACK, PUSHBACK, VFR_CLEARANCE, DIRECT_TO, ENTER_CTR. Therefore, an analysis of selected commands on one half of the data shows which window size could be chosen best for them individually as shown in TABLE I. This analysis serves as an input for optimization of the TCHG towards future technology readiness level (TRL) 6. For the further analysis, the above reported determined window size of 11x11 is used.

TABLE I. VARIATION OF BEST WINDOW SIZES FOR MACHINE LEARNING OF DIFFERENT TOWER COMMAND TYPES

Controller Command Type	*Best Window Size*
INFORMATION	9
CLEARED	8
INIT_RESPONSE	11
REPORT	12
TAXI	11
CLIMB	2
CONTACT_FREQUENCY	7
VACATE	9
PUSHBACK	1

D. *Transcription and Annotation of Controller Utternaces*

Manual transcription and annotation of controller utterances is a very time consuming process. The new software tool CoCoLoToCoCo (Controller Command Logging Tool for Context Comparison) concentrating on efficient usage has been developed to accelerate this process (see figure 4). This tool also performs a context check to evaluate whether the given commands were forecasted or not. It also performs automatic plausibility checks for transcriptions and annotations with respect to ontology format, air traffic rules, common typing errors, etc. The graphical user interface of CoCoLoToCoCo

- lists all audio wav-files (“wave”) with timestamps and different colors for transcription and annotation progress levels in the upper left,
- allows to generate and edit word-by-word transcriptions in cor-files (“correct”) at the bottom,
- shows the list of cmd-files (“command”) for the current annotations above, e.g. “DLH2A LINEUP RW05R” – with context check (green is in context; red is not) – being editable via the six column menus, and
- maintains nfo-files (“inform”) for comments on the upper right side.

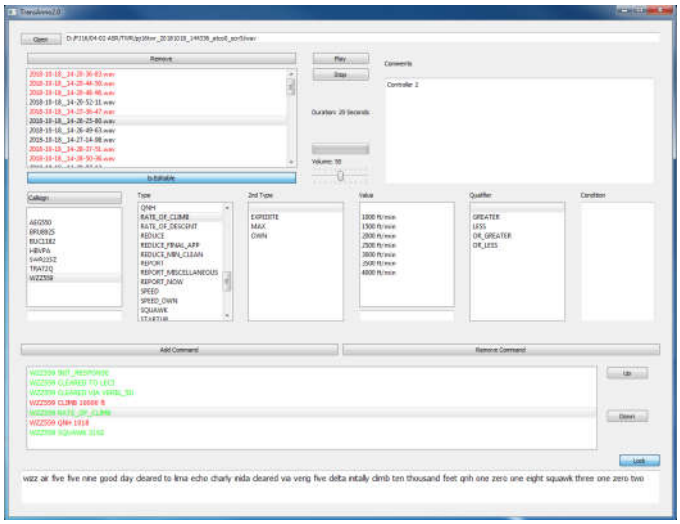

Figure 4. Software tool for efficient transcription and annotation of controller utterances using standardized ontology terms and performing integrated hypotheses and plausibility checks.

V. RESULTS OF TCHG VALIDATION EXERCISE

A. *Applying Machine Learning Techniques and Evaluation of Command Prediction Quality*

There are four different data sets called Hungary-2017-11, Lithuania-2018-03, Hungary-2018-11, and Lithuania-2018-12. As described above, an evaluation analysis consists of training the prediction models via machine learning and testing afterwards. Four different training and test set combinations have been used according to history of data. Those training/test sets are called as listed in TABLE II.

TABLE II. OVERVIEW ON TRAINING AND TEST DATA SETS WITH TIME OF RECORDING AND AIR NAVIGATION SERVICE PROVIDER INFORMATION

Name of Evaluation	*Training with Dataset*	*Test with Dataset*
HUNGARY	Hungary-2017-11	Hungary-2018-11
LITHUANIA	Lithuania-2018-03	Lithuania-2018-12
BOTH_COUNTRIES	Hungary-2017-11, Lithuania-2018-03	Hungary-2018-11, Lithuania-2018-12
COMPLETE	Hungary-2017-11, Lithuania-2018-03, Hungary-2018-11	Lithuania-2018-12

For each of the annotated runs there were context files (ctx) every time the ATCo uttered something. These context files contain all predicted commands of a certain timetick. The results of the command hypotheses evaluation are reported in historical order in the following sub-sections after the overview in TABLE III.

ISSN 0770-1268

TABLE III. OVERVIEW OF CPER (UNDERLINED IF SIGNIFICATNLY BELOW 10%), STANDARD DEVIATION OF CPER (SD) AS WELL AS AVERAGE AND MAXIMUM NUMBER OF HYPOTHESES RELEVANT FOR CONTEXT SIZE IN CTX-FILES (CTX_AVG, CTX_MAX) FOR THE FOUR DIFFERENT EVALUATIONS

Name of Evaluation	*CPER*	*SD*	*ctx_avg*	*ctx_max*
HUNGARY	7.8%	2.57%	450	593
LITHUANIA	12.5%	3.6%	340	498
BOTH_COUNTRIES	7.3%	2.46%	548	760
COMPLETE	7.5%	3.7%	629	900

1) HUNGARY

For the six annotated Hungarian simulation runs with a scenario without runway configuration change, the average CPER per run is 7.8% (SD: 2.57%). As there was no runway configuration change in the training data (Hungary-2017-11 runs), this of course could not be learned. Hence, such a new aspect negatively influences the command prediction quality i.e. for the runway configuration change this affected TAXI, VACATE, CLEARED LANDING/TAKEOFF, etc. commands.

Taking three further simulation runs with runway configuration changes also into account, the CPER is 9.1% (SD: 2.85%). 450 commands have been predicted on average (ctx_avg). When analyzing the biggest set of predicted commands per run (ctx_max), this averages to 593 for the Hungary-2018-11 runs.

All of the top 14 commands (that were at least used in 2.2% of all commands from the Hungary-2018-11 scenarios without runway change) showed CPERs below 8.3%. However, for the "TAXI TO"-command this is only true if a generalization of stands is made (so specific stands such as "R107" were not forecasted, but only "TAXI TO STAND").

2) LITHUANIA

For the nine annotated Lithuanian simulation runs, the average CPER per run is 12.5% (SD: 3.6%). Many scenarios had aircraft repeating touch-and-gos respectively go-arounds which are more difficult to predict. When ignoring such command predictions, the CPER would be around 7%. 340 commands have been predicted per context file in average (ctx_avg). When analyzing the ctx_max, this averages to 498 for the Lithuania-2018-12 runs.

3) BOTH_COUNTRIES

For the six annotated Hungarian simulation runs – however, machine learning performed on Hungary-2017-11 and Lithuania-2018-03 – the average CPER per run is 7.3% (SD: 2.46%). These numbers are reported as the main result as the technical validation plan foresaw validation trials with Hungarian ATCos and a command prediction model that learned from all available data before.

Taking also the nine annotated Lithuanian simulation runs into account, the average CPER per run is 7.9% (SD: 3.2%). Taking all 18 annotated simulation runs (with runway configuration changes) into account, the CPER is still below 10% – in a range between 8 and 9%. For the Autumn-2018 runs the ctx_avg and ctx_max were 548 respectively 760.

4) COMPLETE

For the nine annotated Lithuanian simulation runs – however, machine learning performed on Hungary-2017-11, Lithuania-2018-03, and Hungary-2018-11 – the average CPER per run is 7.5% (SD: 3.7%). The ctx_avg was 629, the ctx_max was 900 for the Lithuania-2018-12 runs.

5) Significance of Results

The confidence in the results of the exercise is high due to the number of simulation runs, i.e. the CPERs have high statistical significance. The performed t-test tested against the required average value of 10%. The obtained p-value is 1.18% for the "BOTH_COUNTRIES-data" respectively 3.29% for the core "HUNGARY-data". Normally, a statistical significance of below 5% is required to significantly support the underlying assumption. Hence, we can conclude that the average CPER per run is very surely below 10%. For the reported "LITHUANIA-data" with a p-value of 2.46%, we assume that the CPER is above 10% due to the t-test. However, for the "COMPLETE"-data, testing the same Lithuanian files with the reported model learned on more data, we can assume that the CPER is below 10% with a p-value of 2.91%.

6) General Notes

The CPP was below 10% all the times. However, the ctx_avg shows that more training data lead to more command predictions. This results in less command prediction errors, but is increasing the search space for a speech recognition engine. Furthermore, it has to be noted, that there are still unintended human-made transcription and annotation errors in the data. The CoCoLoToCoCo tool notifies the human annotator about possible errors. This tool continuously improves; however, it is not able to detect all errors. Besides, it might have a small influence which simulation runs have been chosen to be annotated and thus analyzed. In general, it can be stated, that a CPER below 10% was achieved and can be further enhanced.

B. Real-Time Aspects of Command Prediction with Respect to Given Commands

In average, software-based generation of command hypotheses took 119 milliseconds (analyzed from log files of 59 simulation runs of final trials in autumn 2018) which is a factor of 40 better than required.

Context (set of command hypotheses) has been generated more than 21'000 times. As traffic density during tower trials was rather medium to low (compared to former ATC approach trials, in which context was used), context was generated only every 10 seconds. However, the measurements show that it is possible to do it 80 times more often if needed. The context generation frequency was also reasonable, because the average length of radio telephony (RT) calls from tower controllers is shorter than 10 seconds as shown in figure 5. These five to seven seconds of simple ATCo command communication have also been reported for en-route sectors [28].

The portion of time used for commands and the number of commands is visualized in figure 6. This emphasizes the potential for workload reduction by usage of ASR, because each command nowadays means that ATCos also need to perform manual input into the ATC system.

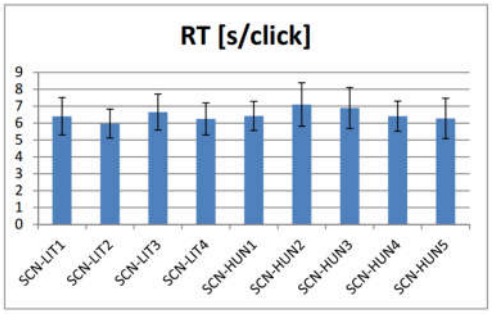

Figure 5. Average duration of radio telephony transmissions in seconds from tower controller to pilot (with positive and negative standard deviation).

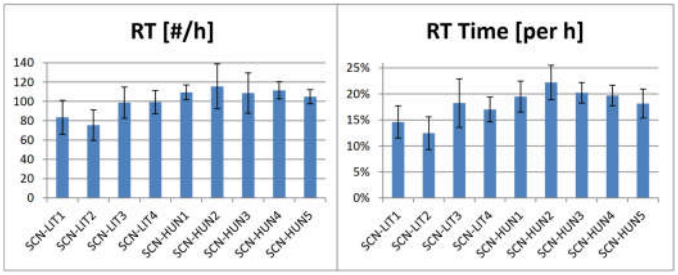

Figure 6. Percentage of time and number of tower controller utterances per hour using radio telephony (with positive and negative standard deviation).

Within the number of roughly 21'000 command prediction attempts, only four of them lasted longer than one second. However, it was always faster than two seconds. Further analyses showed, that not command prediction itself was slow, but storing the data into the data bases respectively getting input data from data base. All other attempts succeeded in less than one second. These numbers are highly reliable regarding the number and duration of runs.

C. *Debriefing Comments*

Each ATCo also joined a debriefing session. This included a discussion about strengths and weaknesses of transcription and annotation focusing on the appropriateness of the ontology and the CoCoLoToCoCo support tool. Furthermore, the ATCos were asked for their opinion about an ABSR system in the daily life CWP. Paraphrased answers are reported in the following.

ABSR do not seem to be of huge interest for one of the controllers, who does not need to enter many things in the CWP HMI today. A list of last clearances in written form would be good for another ATCo as he also uses the hearback replay button in his CWP. ABSR would be a good support for improvement of the HMI, for departure clearances, as well as flight levels and squawks. Also regarding future ATC systems for the ground – that will be introduced in the next two years – ABSR support would be helpful as ATCos need to enter all waypoints, taxi points, routes, etc. for applications such as "follow-the-greens". Other ATCos also mentioned safety critical aspects that could be supported by ABSR. ABSR should recognize "RWY blocked" and show it on the HMI if ground personnel enters a runway. Furthermore, a reliable readback failure presentation would be great, especially for digits in clearances, frequencies, and for the attributes "left" respectively "right".

VI. Summary and Outlook

A. *Summary of Command Hypotheses Validation Results*

The complete trials generated 107 recorded simulation runs for data analysis. There was no need to evaluate also the data of 26 ATCo training runs. The results of the simulation runs with respect to the TCHG are positive and encouraging. The TCHG fulfilled both validation objectives, i.e. operational feasibility as well as performance stability was validated. One aim was to have a command prediction time at least as fast as the update rate of radar data, i.e. the prediction time should be below 5 seconds. Command predictions were forecasted timely and could be generated on average even every 120ms and was always below 2 seconds. The CPER achieved its targets to stay below 10% with a standard deviation below 2.5%. Furthermore, the CPP was below 10% for all simulation runs, however, having limited expressiveness as a valuable result. Nevertheless, the CPP was in a comparable dimension as for the approach hypotheses generator.

AcListant® project showed that ATC command hypotheses improve the recognition quality of an ABSR system in the approach environment. Hence, the positive evaluation of the forecast quality of the TCHG is a central factor for a later tower ABSR system. This future system could present the recognized commands to the controller and might be used similar to the actions in the other four PJ.16-04 ASR exercises.

B. *Outlook on Future Work*

There will be further industrial research on ABSR in SESAR's Wave 2 project PJ.05 "Digital Technologies for Tower" more precisely solution 97 "HMI Interaction modes for Airport Tower". It is foreseen to integrate an enhanced TCHG with a speech recognizer to generate the first ABSR system for towers transitioning from TRL2 to TRL4. Two real-time simulation exercises are planned that consist of a generic or Vienna tower environment.

Validation EXE-003 will be performed at EUROCONTROL tower platform in Brétigny with partners DLR, COOPANS (ACG and CCL), B4 (ON and ANS-CR), and ENAV. Validation EXE-006 will be conducted in Remote TowerLab at DLR Braunschweig with the partners ACG, CCL, and ANS-CR. As the aerodrome ATC includes more types of clearances, e.g. for VFR flights than the approach ATC, it is of interest if the results from AcListant® can be extrapolated, i.e. if tower command hypotheses improve the command recognition error rate (CRER) of a not yet known speech recognition engine as much as in the approach environment. Positive effects on tower controllers' workload in the future can be assumed, but – conform to planning – this has not yet been proven.

Regarding the implementation of the enhanced TCHG there are different aspects of improvement for the context quality. The set of hypotheses should be minimized and fulfil even more requirements valid on ground. A state machine approach – complementing the machine learning approach – could deliver even more background knowledge for the TCHG. There are more single actions and command types that succeed each other in a certain order in the tower than in the approach environment.

If e.g. a taxi clearance followed a pushback clearance, the likelihood of a startup or landing clearance afterwards is almost zero. However, the likelihood of a lineup clearance is very high. This of course depends on the accuracy of the data quality of former clearances. If they are derived from the ABSR system output, the follow-up states of the aircraft and thus of clearances is connected to a certain probability. Individual window sizes per controller command type prediction as described above can further improve machine learning results and thus command hypotheses accuracy. Additionally, the growing amount of data, i.e. (1) available radar and speech data for training of all scenarios and environments that are tested and (2) annotated speech data to then optimize CPER and CPP helps to build a TCHG on higher TRLs.

Moreover, it will be investigated if the developed ontology for annotation of controller commands and the local accent of the English radio communication language is feasible for other tower controllers. This evaluation of speech data will be done on already performed multiple remote tower trials in May 2019 with ATCos from Finland and will be done with real life data from Vienna, Vilnius, and Prague tower. Furthermore, the ABSR systems with TCHG should be brought closer to the operation's room, i.e. real towers or remote towers to support various applications of using speech information.

ACKNOWLEDGMENT

The PJ.16-04 CWP HMI project also comprising the Automatic Speech Recognition activity (PJ.16-04-02 ASR) has received funding from the SESAR Joint Undertaking under the European Union's grant agreement No. 734141.

REFERENCES

[1] 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.

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

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

[4] C. Hamel, D. Kotick, and M. Layton, "Microcomputer system integration for Air Control Training," Special Report SR89-01, Naval Training Systems Center, Orlando, Florida, USA, 1989.

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

[6] R. Tarakan, K. Baldwin, and R. Rozen, "An automated simulation pilot capability to support advanced air traffic controller training," 26th Congress of the International Council of the Aeronautical Sciences, Anchorage, Alaska, USA, 2008.

[7] FAA, "2012 National aviation research plan (NARP)," March 2012.

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

[9] S. Chen, H.D. Kopald, A. Elessawy, Z. Levonian, and R.M. Tarakan, "Speech inputs to surface safety logic systems," IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), Prague, Czech Republic, 2015.

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

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

[12] J.M. Cordero, N. Rodríguez, J.M. de Pablo, and M. Dorado, "Automated speech recognition in controller communications applied to workload measurement," 3rd SESAR Innovation Days, Stockholm, Sweden, 2013.

[13] T. Shore, F. Faubel, H. Helmke, and D. Klakow, "Knowledge-based word lattice rescoring in a dynamic context," Interspeech 2012, Portland, Oregon, USA, Sep. 2012.

[14] 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, California, USA, 2016.

[15] 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, Washington, USA, 2017.

[16] J. Rataj, H. Helmke, and O. Ohneiser, "AcListant with Continuous Learning: Speech Recognition in Air Traffic Control," ENRI Int. Workshop on ATM/CNS, Tokyo, Japan, EIWAC, 2019.

[17] M. Kleinert, H. Helmke, H. Ehr, C. Kern, D. Klakow, P. Motlicek, M. Singh, and G. Siol, "Building Blocks of Assistant Based Speech Recognition for Air Traffic Management Applications," 8th SESAR Innovation Days, Salzburg, Austria, 2018.

[18] 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?," 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria, 2019.

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

[20] The project AcListant® (Active Listening Assistant) http://www.aclistant.de/wp, n.d.

[21] A. Schmidt, "Integrating Situational Context Information into an Online ASR System for Air Traffic Control," Master Thesis, Saarland University (UdS), 2014.

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

[23] The project MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance) http://www.malorca-project.de, n.d.

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

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

[26] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszák, Y. Oualil, and H. Helmke, "Semisupervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, Aug. 2017.

[27] 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.

[28] N. Rodríguez and J.M. Cordero, "Relationship between workload and duration of ATC voice communications," 6th International Conference on Research in Air Transportation, Istanbul, Turkey, 2014.