Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety

Hartmut Helmke, Matthias Kleinert, Shruthi Shetty,
Oliver Ohneiser, Heiko Ehr
German Aerospace Center (DLR),
Braunschweig, Germany
Firstname.Lastname@dlr.de

Hörður Arilíusson, Teodor S. Simiganoschi
Isavia ANS ehf., 102 Reykjavik, Iceland,
Hordur.Ariliusson@isavia.is, Teodor.Simiganoschi@isavia.is

Amrutha Prasad, Petr Motlicek
Idiap Research Institute, Martigny, Switzerland,
Amrutha.Prasad@idiap.ch, Petr.Motlicek@idiap.ch

Karel Veselý, Karel Ondřej, Pavel Smrz
Brno University of Technology (BUT),
Brno, Czech Republic, iveselyk@fit.vutbr.cz,
ondrej@fit.vutbr.cz, smrz@fit.vutbr.cz

Julia Harfmann
NATS (Enroute) PLC, Whitely, Fareham, Hampshire,
United Kingdom, Julia.Harfmann@nats.co.uk

Christian Windisch
Austro Control, Vienna, Austria,
Christian.Windisch@austrocontrol.at

***Abstract*—One of the crucial tasks of an air traffic controller (ATCo) is to evaluate pilot readbacks and to react in case of errors. Undetected readback errors, when not corrected by ATCos, can have a dramatic impact on air traffic management (ATM) safety. Although they seldomly occur, the benefits of even one prevented incident due to automatic readback error detection justifies the efforts. This, however, requires highly reliable detections, which is beyond the performance of currently available automatic speech recognition implementations. The HAAWAII project aims to achieve false alarm rates below 10% and readback error detection rates better than 50%. After performing a preliminary analysis by comparing ATCo utterances with pilot readbacks on word level, this approach proves to be very ineffective. Callsigns are abbreviated or not even pronounced, altitude and speed units are often not used, for example "nineteen eight" is the same as "one one nine decimal eight". Therefore, the presented approach transforms recognized word sequences into so-called ATC concepts, as agreed with the ontology of the SESAR project PJ.16-04. Detecting readback errors on concept level is more reliable and robust as it also considers different forms of conveying the same semantic messages and is also more tolerant to partially misrecognized words. Nevertheless, a good recognition rate on word level is essential to correctly transform words into concepts, which will be achieved by integrating voice data from ATCo utterances and pilot readbacks with context information such as data concerning radar, flight plans, and weather. This paper presents relevant use cases, the ontology-based algorithm, and initial results regarding callsign recognition accuracy for automatic readback error detection purposes.**

Keywords—Automatic Speech Recognition (ASR), Readback Error Detection, Air Traffic Control (ATC)

I. INTRODUCTION

Voice communication between air traffic controller (ATCo) and pilot using radio equipment is still widely used. The ATCo gives verbal commands to an aircraft. The pilot has to repeat all the commands that influence the motion of the aircraft, e.g., altitude, speed or direction commands. This repetition of the ATCo clearances by the pilot is called readback. Beside other tasks, the ATCo is also responsible for the hearback, i.e., monitoring that all pilot readbacks are correct. During high traffic periods, the ATCo simultaneously communicates with many aircraft pilots, which can lead to a lower situation awareness level due to high workload. Readback errors which are not corrected in time can cause incidents and can even (very seldom) result in accidents. In order to reduce the workload and increase the awareness level of the ATCo, Automatic Speech Recognition (ASR) could be a solution to support readback error detection (RED). ASR based RED, however, requires a good accuracy, a low false alarm rate and a close-to real time availability. Accuracy translates to high detection rates of pilot readback errors. A low false alarm rate means an ASR-supported RED assistant should not falsely trigger the ATCo' s attention too often in case of false detections. Otherwise, the ATCo will most likely start to ignore the readback error alarms. A readback error detection rate of 50% seems to be acceptable if the false alarm rate is in the order of 10%. In other words, from 100 readback errors, at least 50 should be detected and from 100 readback error alarms at least 90 of them should be correct. The ATCo's user interface will have to be integrated and designed in such a way that it provides close-to real time detection, i.e., the classification whether a pilot readback is an error or not must be available immediately after the pilot's verbal readback.

This paper provides some related work to readback errors and their detection in the next section. Section III details relevant use cases, section IV describes the RED approach of the HAAWAII project, which transforms the ASR output into air traffic control (ATC) concept elements. Section V analyzes the performance requirements for the used ASR system to achieve the required false alarm rates. Sections VI and VII present initial results, before the last section concludes.

II. RELATED WORK

The content of communication between ATCos and pilots is of utmost importance for the safety of air traffic. Roughly 80% of incidents or accidents involve miscommunication between ATCos and pilots at least as a circumstance based on NASA aviation safety reporting system reports [1]. Based on EUROCONTROL data, miscommunication is the reason for roughly 30% of incidents [2]. Miscommunication can comprise of different aspects such as not responding at all or mishearing that may lead to partial or full misunderstanding by either ATCo or pilot [3]. The communication feedback loop between ATCos and pilots shall ensure a low level of communication errors using information redundancy. ATCos transmit verbal ATC instructions via radiotelephony whose safety-related parts need to be read back by pilots according to International Civil Aviation

HAWWAII project and PJ.10-96 (Wave-2) solution are partly funded by SESAR Joint Undertaking (Grant Numbers 874464 resp. 884287).

Organization (ICAO) Annex 11. ATCos need to hear back pilot readbacks and correct the readback in case of errors [4]. Furthermore, the ICAO phraseology defines clear structure and vocabulary to be used in aviation radiotelephony for avoiding misunderstandings [5]. It has to be noted that readback errors are a subset of communication errors/miscommunication in general. A hearback error is a readback error, which is undetected by the ATCo and is left uncorrected.

Hence, communication errors occur very seldomly, i.e., depending on the definition of an error in less than every hundredth [6],[7] or even up to every sixteenth ATC communication with some transmissions even containing multiple errors [8]. When analyzing only specific ATC command types or communication error effects, "one runway incursion for every 163,000 hearback errors, one runway incursion for every 407,000 readback errors, or one runway incursion for every 40,700,000 commands" has been calculated [9]. For this calculation, the assumption is that 2% of the given ATCo commands result in a readback error with 40% of those 2% are not corrected by the ATCo, i.e., 0.8% of all commands are hearback errors. From another European ATC radiotelephony sample, 1.4 communication errors per 10,000 flights or 2.4 communication errors per million instructions actually result in an incident whereof 40% occur during cruise-flight [10]. Half of ATC miscommunication in US en-route sectors is attributed to pilots mishearing or not responding, a quarter due to the same errors by ATCos [3]. En-route ATCos are capable of detecting 90% of pilot readback errors by themselves [6], but a small amount of readback errors with unclear effects remain undetected. The hearback error rate seems to increase with more transmissions per time slot, because tower/local ATCos detect only 63% and radar approach ATCos detect only 50% of all readback errors [7], [11]. Further factors increase the likelihood of readback errors and clarification requests such as long utterances [12], more complex instructions [13], non-native English speakers [14], deviations from the ICAO phraseology [11],[15], or the current flight phase, i.e., pilots in approach produce more readback errors than in departure phase [8]. It was found that 85% of ATCos claim following ICAO phraseology, but in reality, not even 20% of their utterances follow them in detail [16]. Also, roughly one third of ATC utterances contain a greeting that is not recommended by ICAO, 3% contain hesitations, and 1% include corrections [16]. Some labels to annotate content errors of ATC utterances have been proposed, namely *grouped, sequential, omission, substitution, transposition, excessive verbiage, partial readback* [17].

Miscommunication affects different aircraft states, i.e., almost 40% of miscommunications result in altitude deviations [10], more than one-third of readback errors in en-route deal with frequency changes [6], and 10% of communication errors result from speed being mixed up with headings [18]. Moreover, about 20% of communication errors are also caused due to the presence of similar callsigns on the same frequency [10], [19]. This has unintended effects on safety such as runway incursions [20], e.g., in 0.2%-0.8% of the cases, pilots responded to a transmission, which was not intended for them [6], [11] and only 39% of them have been detected by the humans involved [6]. It is noteworthy to mention that roughly one-quarter of real-life data readbacks and 42% of erroneous readbacks either did not contain a callsign or they were uttered incompletely [6]. In simulations, however, only 5% of the utterances did not contain a callsign [16]. These numbers are confirmed in this paper.

Given the above findings on miscommunication and human error detection, a reliable system for automatic readback error detection without bothering ATCos with too many false alarms seems reasonable, but extremely challenging. Such a system requires ASR to initially convert spoken ATC utterances into written text to enable further analysis of ATCo and pilot utterance semantics, i.e., a "language technology system" like proposed for Icelandic oceanic environment [21]. It is especially challenging ("twice as hard") to correctly recognize pilot speech with their tendency to shorten utterances as compared to ATCo speech [22]. The word error rates currently achieved for real ATC recordings (low quality data) are 8% or worse, and that of simulated data (clean data) are slightly better [23]. The second important step for readback error detection is language understanding, also called as spoken instruction understanding in ATC [24]. With this understanding, semantics of utterances can be converted into a standardized form. An ontology for annotating ATCos' and pilots' utterances as agreed between European ATC stakeholders [25] helps to compare the semantic contents of ATCo transmissions and pilot readbacks as they often use different words and readback order [9]. Automatic extraction algorithms for ATC concepts have already been developed for the tower [9], [26] and approach domains [27]. In addition, automatic pairing of utterance semantics from ATCos and pilots belonging together is part of further research [9]. This helps in avoiding readback error alarms, if the error has already been detected and corrected in the ATCo's hearback. In addition, it needs to be defined, which differences in readbacks are tolerable to all actors, which are alert worthy, and which are clearly unsafe [9]. As a final step, a graphical user interface for handling readback errors as sketched in [9] is also needed. However, the biggest challenge remains in having a low false detection rate without decreasing the true positive rate significantly. Chen et al. [9] provided an example that in the tower area 79% of callsigns spoken by ATCos and 63% spoken by pilots have been recognized correctly by their first implementation as well as over 90% (ATCos) and 80-90% (pilots) of specific tower commands such as *lineup* and *hold short* [9]. However, these numbers result to the fact that each tenth readback would be highlighted as readback error, although the authors observe readback errors only in 2% of the utterances, i.e., the false detection rate would be above 80%, which will not be operationally acceptable [9].

In addition to traditional rule/pattern-based approaches to readback error detection and classification, modern methods employ the concept of machine-learning (ML) that trains specific models on available data and devises statistical data-dependent classifiers. As in other fields of computer science, recent ML approaches take advantage of deep neural networks. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models are applied in [28], [29], [30] to compute contextual representation of transcribed pairs of ATCo-pilot conversation. The data for training is collected from manually transcribed communication and books for civil aviation radiotelephony training in Chinese. A simple, one-layer convolutional neural network for readback error classification is introduced in [31]. This model classifies pairs of ATCo command and the corresponding pilot readback into six classes: *correct readback, partial information loss, call sign readback error, altitude readback error, runway readback error,* and *heading readback error*. 2,500 pairs containing a readback error were collected.

The above literature review makes obvious that a big variation in communication error rates exists. The HAAWAII project

(**H**ighly **A**utomated **A**ir Traffic Controller **W**orkstation with **A**rtificial **I**ntelligence **I**ntegration) will provide a better data base at least for NATS Terminal Manoeuvring Area (TMA) airspace and Icelandic enroute airspace. HAAWAII is led by DLR with the partners Idiap, BUT, NATS, Isavia ANS ehf., Austro Control, and Croatia Control. HAAWAII will develop a reliable, error resilient and adaptable solution to automatically transcribe voice commands from ATCos and pilots with the objective to develop a readback error detection assistant for the ATCo [32].

III. Use Cases of Readback Error Detection

A readback (RB) error can occur at different stages in the communication. Fig. 1 describes some of the cases using ATCo-pilot communication timeline view.

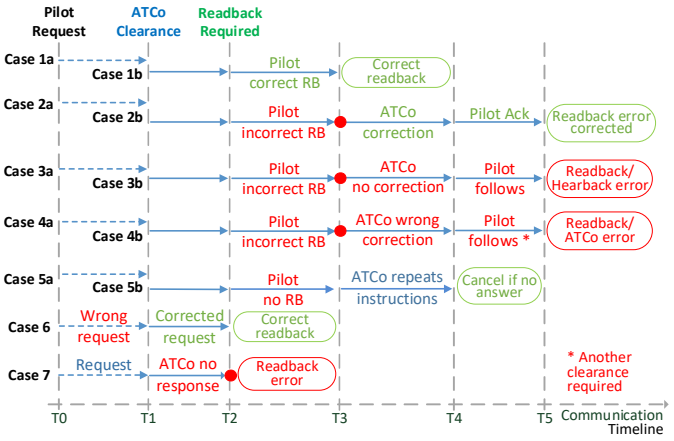

Figure 1. ATCo-pilot communication displayed in timeline view.

The use cases 1 to 5 are split into two sub use cases depending on whether the pilot starts communication (first vertical dotted line or the ATCo, which is the usual case. In the following examples which are related to Fig. 1, we show the content of ATCo-pilot conversations and also what the expected reaction of the readback error detection system should be. We mark potential readback errors in yellow and correct readbacks in green. TABLE I shows a common ATCo-pilot communication with a pilot request, an ATCo clearance, and a correct pilot readback.

TABLE I. Use Case 1A

	Sequence of spoken Words / Transcription
Pilot	reykjavik control faeroline five five requesting lower
ATCo	faeroline five five descend out of controlled airspace
Pilot	descend below controlled airspace faeroline five five
System	No Readback error

Pilot requests, ATCo issues clearance, Pilot correct Readback → No Readback Error

The use case in TABLE II shows an example, where the ATCo corrects a wrong pilot readback. The concept of readback error detection assumes that a potential readback error is immediately indicated. If the ATCo, however, detects the readback error after a couple of seconds, the system should recognize this and not demand further action. The idea is that the ATCo immediately gets a *yellow light*, which only turns into *red*, if the wrong readback is undetected for some seconds, see the attention guidance guidelines created by SESAR project PJ.16-04-AG in wave-1 [33]. The example also shows that a unit, e.g., "flight level", is not always repeated. Although this could technically be considered as a readback error, the ATCo should not be warned, because these cases occur in real world operations.

TABLE II. Use Case 2A

	Sequence of spoken Words / Transcription
Pilot	reykjavik control lufthansa four zero zero requesting flight level three eight zero
ATCo	lufthansa four hundred climb flight level three seven zero
Pilot	climb three eight zero lufthansa four zero zero
System	Potential readback error is indicated
ATCo	correction lufthansa four zero zero climb flight level three seven zero
Pilot	three seven zero lufthansa four hundred
System	Readback error indication is cancelled

Pilot requests, ATCo issues clearance, Pilot incorrect Readback, ATCo correction, Pilot ACK→ No Readback Error

Use Case 3b in TABLE III is quite straight forward. The ATCo gives a clearance to flight level 100. The pilot reads back 110. Here, the system has to detect this as a readback error.

TABLE III. Use Case 3B

	Sequence of spoken Words / Transcription
ATCo	scandinavian nine two four descend flight level one hundred
Pilot	descend flight level one one zero scandinavian nine two four
System	Readback error

ATCo issues clearance, Pilot incorrect Readback, ATCo no correction → Readback Error

In the use case in TABLE IV, the pilot requests a waypoint. The ATCo, however, issues a clearance to another waypoint which sounds similar (by accident or on purpose). The pilot in turn repeats the initially requested waypoint. Therefore, a readback error will be indicated. The ATCo now confirms the wrong readback to the first clearance. This could be done by intention or by accident. It is assumed that this could be done by accident, because the phraseology *recleared* should be used. Therefore, the readback error indicated is not released.

TABLE IV. Use Case 4A

	Sequence of spoken Words / Transcription
Pilot	reykjavík control arctic eagle one one six requesting direct alfa kilo india
ATCo	arctic eagle one one six cleared direct alfa kilo charlie
Pilot	direct alfa kilo india arctic eagle one one six
System	Potential readback error is indicated
ATCo	roger alfa kilo india
System	Readback error indicator is retained

Pilot requests, ATCo issues clearance, Pilot incorrect Readback, ATCo wrong/no correction → Readback Error

TABLE V. Use Case 5B

	Sequence of spoken Words / Transcription
ATCo	foxtrot lima india descend flight level one zero zero
Pilot	"no reply"
System	Missing readback indicator, which will be deleted when ATCo repeats the clearance
ATCo	foxtrot lima india descend flight level one zero zero
Pilot	"no reply"
System	Missing indicator is cancelled, i.e., previous clearance is kept

ATCo issues clearance, Pilot "no reply", ATCo repeats instructions, Pilot "no reply"→ No Readback Error

TABLE V illustrates use case 5b, where the ATCo issues a clearance to which the pilot does not react. The system should indicate that there is a missing readback. The ATCo then repeats the instruction, to which the pilot again does not read back. The RED system then removes the missing indicator and retains the previously issued clearance.

In TABLE VI the pilot asks for a strange speed respectively the speech recognizer understands mach 0.2. Mach numbers for aircraft are usually above 0.5. The ATCo corrects the requested value and the pilot accepts with a simple "that is correct". The system should not indicate this as a readback error. Without the correction of the ATCo, it would be use case 7.

TABLE VI. USE CASE 6

	Sequence of spoken Words / Transcription
Pilot	reykjavik control delta six zero five requesting mach decimal two
ATCo	delta six zero five confirm requesting mach decimal eight two
Pilot	that is correct
System	No Readback error

Strange pilot request, ATCo corrects request, Pilot confirms correction → No Readback Error

The following use case in TABLE VII is not shown in Fig. 1 and is related to a case, in which more than one aircraft/pilot is involved.

TABLE VII. USE CASE WITH MORE THAN ONE PILOT INVOLVED

	Sequence of spoken Words / Transcription
ATCo	lufthansa two alfa four turn left heading three two zero
Pilot1	two alfa four turning right three two zero
ATCo	speed bird one one descend flight level one two zero
System	Readback error for DLH2A4
Pilot2	descending level one two zero speed bird one one
System	No readback error for BAW11
ATCo	lufthansa two alfa four negative turn left heading three two zero turn left
System	Readback error indicator now disappears for DLH2A4

Incorrect readback without ATCo correction, ATCo gives instructions to another aircraft not correcting the readback error with the first aircraft

After detecting a possible readback error the next step would be to integrate it to the ATCo's radar display. The visualization on the display is not in the scope of the HAAWAII project, but nevertheless there are high level guidelines defined for readback error integration within the project. The readback errors could be integrated into the call sign label as a readback error icon. The ATCo has the possibility to click on the icon and extend the notification message showing the exact text, where the error occurred. The indication is not intrusive and the ATCo can decide to acknowledge the readback error indication or to ignore the indication. The warning will timeout after a predefined time and will be acknowledged automatically.

IV. ONTOLOGY BASED READBACK ERROR DETECTION

In principle, the readback cycle is very easy. The ATCo issues one or more commands in an utterance and the pilot repeats them. If they are not correctly repeated, the ATCo repeats the commands again, until the pilot correctly reads back the given command(s). However, an example as simple as - ATCo: "*saudia one zero seven hello descend flight level one five zero*" and the pilot's readback "*descending one five zero saudia one zero seven*" already indicates the complexity of the task. The ATCo says the callsign in the beginning. (S)he provides a greeting "*hello*" and also mentions the unit ("*flight level*"), which is missing in the pilot's readback. The callsign can be in the beginning (as for most ATCos), towards the end (as for most pilots) or in the middle (occurs seldom) of an utterance. This simple example may have already convinced some of the readers that RED on word level is hopeless. The following example in TABLE VIII should convince most of the readers.

TABLE VIII. EXAMPLE OF ATCO-PILOT COMMUNICATION THAT SHOWS READBACK ERROR DETECTION ON WORD LEVEL VS. CONCEPT LEVEL

	Spoken Words / Transcription	Ontology Instructions / Annotation
ATCo	speed bird two zero zero zero alfa reduce one eight zero knots until DME four miles contact tower on frequency one one eight decimal seven zero zero	BAW2000A REDUCE 180 kt UNTIL 4 NM DME BAW2000A CONTACT TOWER BAW2000A CONTACT_FREQUENCY 118.700
Pilot	one eighty to DME four tower one eighteen seven speed bird two thousand alfa	BAW2000A PILOT SPEED 180 none UNTIL 4 none DME BAW2000A PILOT CONTACT TOWER BAW2000A PILOT CONTACT_FREQUENCY 118.700

Transcription refers to the word-by-word representation of the speech data. **Annotation** refers to the semantic interpretation of the transcription, consisting of a sequence of ATC concepts. Or in other words, annotation refers to the transformation of a sequence of words to a sequence of ATC concepts. The transformation rules (so called ontology) were first defined by fourteen European partners from ATM industry and research as well as by air navigation service providers (ANSPs) funded by SESAR 2020 [25]. The ontology is applied and further improved by different projects, such as STARFiSH [34] and "HMI Interaction Modes for Airport Tower" [35] in the tower environment, by "HMI Interaction modes for approach control" [36], and HAAWAII [32],which also include pilot utterances.

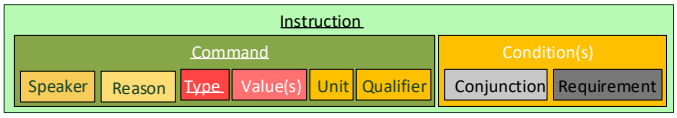

Figure 2. Instruction consisting of a callsign, a command, and condition(s).

Fig. 2 summarizes that an utterance consists of one or more instructions and each instruction starts with the callsign, even if the callsign is only said once. The full intended callsign (from the flight plan or surveillance data) is provided, i.e., BAW2000A is used even if only "speed bird alfa" is said or recognized. If no callsign is said or the callsign could not be uniquely determined, "*NO_CALLSIGN*" is used in the annotation. An instruction consists of a callsign, a command, and optional conditions. A command consists of various fields such as type, value, unit, etc. The command type is always mandatory. It determines how many values are expected. The remaining fields such as the unit (e.g., FL, ft, kt), qualifier (e.g., LESS, OR_BELOW, LEFT), speaker (PILOT or empty), and reason (REQUEST, REPORTING or empty) are optional. We will not be describing the ontology rules here, but concentrate in TABLE IX to TABLE XI on examples showing the advantages of the transformation for readback error detection.

A one-to-one comparison of the extracted (and not extracted) concepts is still not possible as shown by the ATCo-pilot communications in TABLE IX. In the first example, the pilot reports the current descending altitude, which does not require an action by the ATCo. Here, the ATCo allows the aircraft to further descend to flight level 150, but under the condition that it must be reached before reaching the waypoint *TIGER*. The pilot reads back the descend command, the value, and the condition, but leaves out the unit "flight level", which happens quite commonly but is considered bad practice.

TABLE IX. EXAMPLE USE CASE WITH A PILOT REPORT, FOLLOWED BY AN ATCO COMMAND, AND A PILOT READBACK

	Spoken Words / Transcription	Ontology Instructions / Annotation
Pilot	saudia one zero seven descending one eight zero	SVA107 PILOT REPORTING DESCEND 180 none
ATCo	saudia one zero seven hello descend level one five zero be level by tiger	SVA107 GREETING SVA107 DESCEND 150 FL UNTIL REACHING TIGER
Pilot	descending one five zero to be level by tiger saudia one zero seven	SVA107 PILOT DESCEND 150 none UNTIL REACHING TIGER

In the second example (TABLE X), the pilot again initiates the conversation. (S)he requests a direct to the waypoint *ULKIM*.

TABLE X. EXAMPLE USE CASE WITH A PILOT REQUEST, FOLLOWED BY ATCO COMMANDS, AND A PILOT READBACK WITH ONE MISSING COMMAND

	Spoken Words / Transcription	Ontology Instructions / Annotation
Pilot	mike tango yankee request present position direct to position ulkim for the RNAV runway two two and will be ready for descent very shortly	MINTY PILOT REQUEST DIRECT_TO ULKIM none
ATCo	mike tango yankee cleared direct ulkim and when ready descend out of controlled airspace QNH egilsstadir nine nine four	MINTY DIRECT_TO ULKIM none MINTY DESCEND CA none WHEN READY MINTY INFORMATION QNH 994
Pilot	okay when ready descend out of controlled airspace egilsstadir QNH nine nine four mike india november tango yankee	MINTY PILOT DESCEND CA none WHEN READY MINTY PILOT INFORMATION QNH 994
ATCo	mike tango yankee just to confirm you are cleared direct ulkim	MINTY DIRECT_TO ULKIM none
Pilot	sorry cleared direct ulkim	NO_CALLSIGN DIRECT_TO ULKIM none

The request for the RNAV approach to runway 22 is not modeled by (current version of) the ontology rules. The ATCo gives a clearance for *ULKIM* and in addition, also provides a conditional descent clearance and the QNH value. All three command types, i.e., DIRECT_TO, DESCEND, and INFORMATION QNH require a readback. Therefore, the ATCo repeats the *direct to* clearance again due to the missing pilot readback for the DIRECT_TO command. The pilot then reads back the DIRECT_TO command, but without the callsign.

The last example starts with a *climb* command for callsign *MILAN*, followed by a *handover* command for a second aircraft *AUA3BA*. However, the pilot of *AUA3BA* reads back before *MILAN,* and both readbacks are correct. Later, the ATCo gives a *REDUCE* clearance to MILAN, which is followed by a usual pilot readback, although the word *reduce* and the unit are not repeated by the pilot. The pilot immediately asks for a confirmation of the already cleared flight level in a separate utterance. Maybe (s)he is confused by the unusual speed clearance for a departure. The ATCo acknowledges the already cleared altitude to which the pilot reads back, but without mentioning the altitude unit.

In principle, the example utterances and their transformation to annotations show that RED on ontology/annotation level is much simpler than a hopeless comparison on word level. This assumption, of course heavily depends on a reliable algorithm that transforms a sequence of recognized words with a word error rate (WER) above 0% to the corresponding annotations. Indeed, such an algorithm was developed by DLR. Command recognition rates of 99% for Prague and 95% for Vienna Approach were reported [25]. First results from the HAAWAII project are presented in section VII.

TABLE XI. COMPLEX CONVERSATION WITH TWO PILOTS INVOLVED

	Spoken Words / Transcription	Ontology Instructions / Annotation
ATCo	mike alfa november climb flight level two four zero	MILAN CLIMB 240 FL
ATCo	austrian three bravo alfa contact one three four decimal three five zero ciao	AUA3BA CONTACT_FREQUENCY 134.350 AUA3BA FAREWELL
Pilot1	one three four three five zero austrian three bravo alfa servus	AUA3BA CONTACT_FREQUENCY 134.350 AUA3BA PILOT FAREWELL
Pilot2	mike alfa november climb flight level two four zero	MILAN PILOT CLIMB 240 FL
ATCo	mike alfa november reduce two two zero knots or less	MILAN REDUCE 220 kt OR_LESS
Pilot2	speed two two zero or less mike alfa november	MILAN PILOT SPEED 220 none OR_LESS
Pilot2	just to confirm continue climb level two four zero	NO_CALLSIGN PILOT CLIMB 240 FL
ATCo	mike alfa november correct two four zero	MILAN AFFIRM MILAN ALTITUDE 240 none
Pilot2	two four zero mike alfa november	MILAN PILOT ALTITUDE 240 none

We define that a readback is correct, if the same callsign, type, value(s), unit, qualifier, **and** condition(s) are extracted from both ATCo and pilot utterance. Nevertheless, it is not as simple and straightforward as it might seem:

- Not all command types defined in the ontology require a readback, e.g., a GREETING or an INFORMATION TRAFFIC do not require a readback.
- The sequence of ATCo instructions does not necessarily have to be read back in the same order by the pilot.
- A DESCEND command is annotated as ALTITUDE, if the value is read back without the command type. The same applies for other command types, e.g., REDUCE and SPEED command types. Vienna ATCos used, e.g., 52% REDUCE and 48% SPEED command types for reducing the speed.
- The unit is not always repeated by the pilot (see results in next section). Although this is not recommended as per ICAO phraseology standards, reporting this as a readback error may result in heavy workload for the ATCo.
- The callsign is not always provided, especially when the ATCo and pilot continuously communicate with each other.

V. WHICH ACCURACY ON COMMAND LEVEL IS NEEDED?

For simplicity, in this section we do not consider if a readback error is corrected or not. We also ignore if the error is caused by the pilot or ATCo, i.e., the sub case a and b in Fig. 1 are treated as one use case. Without loss of generality, for the rest of the paper we assume that the ATCo gives a command, and the readback (correct or wrong) comes from the pilot. Assuming that we have a human readback error rate (RE) of 2%,

this means that 200 out of 10,000 given commands are not correctly repeated by the pilot. The HAAWAII objective of a readback error detection rate (RD) of 50% requires that at least 100 of these 200 readback errors are detected. In the best case, these readback errors include all hearback errors. The second objective of the HAAWAII project, a false detection rate of 10%, requires that if 111 readback errors are reported, at least 100 of them must be actual readback errors.

The RED rate of course heavily depends on how good the automatic transformation of words into ATC concepts works. The quality of this transformation is measured using command recognition rate (R), command recognition error rate (E), and command rejection rate (RR). The command recognition rate refers to the percentage of correctly recognized commands. The command recognition error rate, on the other hand refers to the percentage of wrongly extracted commands. A command is said to be rejected, if it is not extracted. Therefore, the command rejection rate is the percentage of commands, which are not extracted [37]. One of the best reported results with respect to quality of the transformation for data from outside the laboratory environment is from the MALORCA project with command recognition rates and command recognition error rates of 92% (R_{ATCo}) and 0.6% (E_{ATCo}) for ATCos, respectively.

We assume that a readback error for a command is correctly detected, if all ATC concept elements (callsign, type, value, etc.) from both the ATCo utterance and the pilot's readback are correctly recognized and neither the ATCo's recognition nor the pilot's recognition is rejected. Assuming that the command recognition rates for the ATCo (R_{ATCo}) and pilot utterances (R_{Pilot}) are independent of each other (which is not fully true), the recognition rate for the combined commands R_{both} can be written as:

$$R_{Both} = R_{ATCo} * R_{Pilot} \quad (1)$$

TABLE XII. TERMINOLOGIES USED IN READBACK ERROR DETECTION EVALUATION

True Positive	TP: Readback error is present which is **correctly** detected, i.e., recognition correct and not rejected
False Positive	FP: No readback error is present, but is **falsely** classified as a readback error, i.e., recognition wrong and no rejection for none or for both
True Negative	TN: No readback error present which is **correctly** classified as no readback error, i.e., recognition correct or no recognition
False Negative	FN: Readback error is present, but is **falsely** classified as no readback error, i.e., one recognition is wrong or one of the two recognitions is rejected
Precision	True Positives (TP) divided by the sum of True Positives and False Positives (TP+FP) [38]
Recall	True Positives (TP) divided by the sum of True Positives and False Negatives (TP+FN) [38]

The alarm is the *positive event*, comparable to Covid-19 positive definition in medical domain

The above ATCo command recognition rate of 92% with an assumed R_{Pilot} of 85% for the pilot would according to Eq. (1) result in 78%, which means that 78% of the readback errors would be detected.

A command involving a readback is said to be wrongly recognized as a readback error, if there is an error in the command recognition either for the ATCo (E_{ATCo}) or for the pilot (E_{Pilot}). Assuming again that the recognitions are independent of each other, we write:

$$E_{both} = E_{ATCo} + E_{Pilot} - E_{ATCo} * E_{Pilot} \quad (2)$$

TABLE XII, defines two commonly used metrics *precision* and *recall* in the context of readback error detection. Precision is the percentage of correctly classified readback errors among all which are classified as readback errors. Recall is the percentage of actual readback errors which are correctly detected.

In this work, we further define and use two metrics: readback error detection rate (RD) and false alarm rate (FA). The readback error detection rate is the number of correctly detected readback errors divided by the total number of readback errors. It is equal to the recall (TP/[TP+FN]). The ATC community might be more familiar with *readback error detection rate*, whereas machine learning community might prefer *recall*.

$$RD = \frac{TP}{TP + FN} = \frac{RE * R_{both}}{RE * R_{both} + RE * (1 - R_{both})} = R_{both} \quad (3)$$

The detection rate is independent of the error rate. The detection rate, however, depends on the rejection rate, which in turn is influenced by the error rate. This is because a high error rate can be reduced by rejecting some of the errors, which would also lead to a lower RED rate, which is, therefore, indirectly influenced by the command recognition error rates E_{ATCo} and E_{Pilot}. The false alarm rate (FA), also known as false detection rate can now be defined as the number of wrongly detected readback errors divided by the sum of the wrongly and the correctly detected readback errors. In other words, it is the number of False Positives (FP) divided by the sum of False Positives plus True Positives (TP):

$$FA = \frac{FP}{TP + FP} = \frac{(1 - RE) * E_{both}}{RE * R_{both} + (1 - RE) * E_{both}} \quad (4)$$

The false alarm rate is equal to one minus the precision (TP/[TP+FP]). The above example from the MALORCA project would result in a false alarm rate of 62%, which is far beyond the desired 10% false alarm rate:

$$62\% = \frac{(1 - 2\%) * (0.6\% + 2\% + 0.6\% * 2\%)}{2\% * 78\% + (1 - 2\%)(0.6\% + 2\% + 0.6\% * 2\%)} \quad (5)$$

TABLE XIII. DEPENDENCY OF FALSE ALARM RATE FROM COMBINED COMMAND RECOGNITION ERROR RATE AND RECOGNITION RATE

R_{both} / E_{both}	0.1%	0.2%	0.3%	0.4%	0.5%	0.6%
98%	4.8%	9.1%	13.0%	16.7%	20.0%	23.1%
95%	4.9%	9.4%	13.4%	17.1%	20.5%	23.6%
90%	5.2%	9.8%	14.0%	17.9%	21.4%	24.6%
85%	5.5%	10.3%	14.7%	18.7%	22.4%	25.7%
80%	5.8%	10.9%	15.5%	19.7%	23.4%	26.9%
75%	6.1%	11.6%	16.4%	20.7%	24.6%	28.2%
70%	6.5%	12.3%	17.4%	21.9%	25.9%	29.6%
60%	7.6%	14.0%	19.7%	24.6%	29.0%	32.9%
50%	8.9%	16.4%	22.7%	28.2%	32.9%	37.0%
40%	10.9%	19.7%	26.9%	32.9%	38.0%	42.4%
20%	19.7%	32.9%	42.4%	49.5%	55.1%	59.5%
10%	32.9%	49.5%	59.5%	66.2%	71.0%	74.6%

TABLE XIII shows the dependency between the combined recognition rate R_{both} and the combined recognition error rate E_{both}. In dark green we mark the pairs, which result in a false

alarm rate below 10% and in lighter green the pairs, which still have a false alarm rate at least below 20%. Although we have made some simplifications with respect to independence of the different pilot and ATCo rates, the results are nevertheless very clear: The combined command recognition rate is not so important. Even a command recognition rate of only 50.0% would achieve the required precision. However, a combined command recognition error rate E_{both} of maximum 0.5% is needed for a false alarm rate below 20%. For a false alarm rate of 10% or less, we need a combined command recognition error rate of less than 0.2%. If we compare this with the best available results only on ATCo data from the MALORCA project, i.e., with command recognition rates of 92% and command recognition error rates of 0.6%, the challenge is obvious.

We assumed independence of pilot and ATCo recognition rates, which is not fully given. In a real implementation we can assume that both recognitions will rely on each other, so that the combined recognition rate is higher than the product of the individual recognition rates. The same applies for the recognition error rates. They will not be independent of each other, so in Eq. (2) the term $E_{ATCo} * E_{Pilot}$ will be greater and, therefore, E_{both} is also greater.

Another heuristic to reduce E_{both} is using more than one speech recognition engine and only create a readback alert, if all engines have extracted the same concepts from the ATCo's and the pilot's utterance. However, in case of doubt, it still holds true that no alarm is preferred, i.e., the recognition of the ATCo or the pilot should be rejected, which could be solved by plausibility values on word and on semantic level and also by comparing the extracted commands with the corresponding surveillance data.

VI. Experimental Setup

Surveillance data and the corresponding voice utterances of both pilots and ATCos were recorded in 2020 from Isavia's en-route airspace, London TMA of NATS, and from Vienna TMA and adjacent sectors of Austro Control. The voice utterances containing both pilot and ATCo speech in a continuous audio stream were automatically split, so that each voice recording file contains only one utterance. The automatic splittings of the utterances were manually corrected. Each file was then automatically transcribed, and finally manually checked. The transcriptions were automatically annotated, and parts of these annotations were manually corrected. TABLE XIV shows the amount of data being available from this analysis from the three ANSPs.

TABLE XIV. Amount of Data for Readback Analysis

ANSP	#Utterances	# Commands	Gold Transcriptions [h]	Annotations [h]	Annotations Gold [h]
ACG	2694	6635	3.5	3.5	0.5
Isavia	6744	12,265	7.6	7.6	2.0
NATS	7656	13,165	7.2	7.2	2.5

The term "Gold" always refers to **manually** checked and corrected transcriptions and annotations.

The RED algorithms highly rely on correct callsign detection. Otherwise the readback may come from the wrong pilot, see use case in TABLE VII. Therefore, we have also analyzed, how good the callsign recognition rate is. The callsign recognition rate is improved by using information from the surveillance data specifying which callsigns are currently in the air. Furthermore, the callsign recognition in the readback is improved, if the callsign information from the previous utterance is also used. We also search for the best matching callsign, if surveillance data is provided: if "speed bird two alfa four" is said and only a BAW3A4 (speed bird three alfa four) is in the surveillance data, then the BAW34A is assumed to be the correct one. The plausibilities of the extracted concepts are, however, decreased, which helps to reduce the recognition error rate E_{both} of Eq. (2).

The manual transcriptions were used to improve the recognition models. TABLE XV shows the achieved recognition performance for different models. The basic model was developed without using any data from the three ANSPs, i.e., the initial model, which supported the ANSPs in performing the first manual transcriptions. The first number shows the achieved word error rate for the ATCo utterances and the second one for the pilot utterances. The second model uses roughly two hours of data from manual transcriptions to update the language models. The third model uses six hours of data from manual transcriptions.

TABLE XV. WER for ATCo and Pilot for different ASR Models

	ANSP 1	ANSP 2	ANSP 3
Basic Model	20.4%, 30.3%	13.9%, 31.6%	16.9%, 26.5%
2 hours of domain data	Not available	Not available	15.8%, 26.5%
6 hours of domain data	9.3% / 17.3%	8.0%/23.3%	11.3%, 22.7%

Due to data privacy reasons the clear names of the ANSPs are not provided

In our experiments, conventional bi-phone Convolutional Neural Network (CNN) [39] + Factorized Time Delay Neural Network (TDNN-F) [40] based acoustic models (AM) trained with Kaldi [41] toolkit (i.e., nnet3 model architecture) are used. AMs are trained with the LF-MMI [42] training framework considered to produce state-of-the-art performance for hybrid ASR systems. 3-fold speed perturbation [43] and i-vectors are the basis for all experiments. The 3-gram language model (LM) is used. The baseline AM and LM are trained with a combination of air-traffic command-related databases: NNMATC, HIWIRE, ATCOSIM, AIRBUS, and MALORCA [44] to [49].

In order to improve the baseline system, we perform LM adaptation [50] in two iterations. In the first iteration, roughly two hours of manual transcriptions available were split into development set (1h 20min) and test set (40 min). The development set was then used to build a 3-gram LM and interpolated with LM used in the baseline model. The results are shown in TABLE XV in row "2 hours of domain data". In the next iteration, six hours of manual transcriptions were available which were used to train a 3-gram LM, which was then interpolated with the baseline model, see last row in TABLE XV.

The gold transcriptions were used to analyze the ATCo-pilot conversation with respect to readback relevant information. We analyzed from the automatic annotations, how often the first, second, third, etc. utterance of an ATCo-pilot conversation contains the callsign. When ATCo or pilot start a communication, they nearly always provide their callsign. Otherwise ATCo or pilot would not know who is speaking or who is being addressed.

TABLE XVI. Callsign and Unit provided in [%]

A = ANSP	A 1	A 2	A 3
ATCo utterance without callsign	15%	12%	8%
Pilot utterance without callsign	19%	10%	6%
ATCo utterance without unit	3%	20%	5%
Pilot utterance without unit	26%	42%	26%

The results in TABLE XVI show that the callsign is not always said by both the ATCo and the pilot. This is especially true when they answer immediately. Therefore, a good callsign recognition is very important. They need to be extracted in the first utterance. These numbers are confirmed in TABLE XVII by the data recordings used during the MALORCA and SESAR project 16-04-ASR for Prague and Vienna airports, respectively. MALORCA data results from recordings from the operational environment whereas 16-04-ASR recordings result from the laboratory environment for both Vienna and Prague, respectively. In the operational environment, Prague ATCos omit the callsign in about 5.2% of 3,400 utterances, related to issuing significant commands. In 4.4% of the ops room utterances the callsign was used without a following command. In lab environment Prague ATCos used no callsign in 0.9% of the utterances with at least one valid command. Similar rates were observed for Vienna. The percentage of words used in an utterance, which are not used for command extraction or could not be classified are provided in the last row. These results show that the used phraseology in the laboratory, when ATCOs are aware of the speech recognizer, is less complex, than in the operational environment.

TABLE XVII. NO CALLSIGN SAID IN OPS-ROOM AND IN LAB-ENVIRONMENT

	Prague		**Vienna**	
	Ops	**Lab**	**Ops**	**Lab**
NO_CALLSIGN with commands	5.2%	0.9%	2.0%	0.9%
NO_CALLSIGN NO_CONCEPT	4.4%	2.2%	3.4%	0.3%
Words unused for semantic extraction	10%	1.2%	12%	4.3%

The HAAWAII project has decided in its Operational Concept Document [51], [52] in cooperation with the International Federation of Air Traffic Controllers' Associations (IFATCA) that a missing unit (e.g., feet, flight level, knots) in a readback should be marked as a readback error. The numbers highlighted in yellow in TABLE XVI, however, show that this would result in at least one reported readback error per minute. The results also indicate the more the ATCo deviates from the ICAO rules, the more the pilot is also ignoring units.

VII. FIRST RESULTS

The previous sections have shown that a moderate (combined) command recognition rate R_{both} is necessary, but a very low (combined) command recognition error rate E_{both} is of decisive importance for having low false alarms rates.

Although the callsign is not provided in each utterance, callsign recognition quality is of decisive importance for readback error detection. Therefore, TABLE XVIII provides first results for different callsign recognition heuristics. The callsign recognition rate (CsR) is the number of correctly recognized callsigns including "NO_CALLSIGN" divided by the number of all used callsigns in all utterances. The callsign recognition error rate (CsE) is the number of wrongly recognized callsigns divided by all callsigns. A given callsign which is recognized as *NO_CALLSIGN* is not counted as an error, but as a callsign rejection.

The manually transcribed utterances were automatically annotated using the callsigns from the corresponding surveillance data. Parts of them were manually checked. The rates CsR and CsE, which are provided in rows "*gold callsign info*", are counting only the utterances, which result from manually checked files. It should be clear that the real rates CsR and CsE are lower than the results in row "gold callsign info", when manually checking is performed for more extraction. Nevertheless, already these automatic extractions enable to evaluate the benefits of additional training data and of using callsign information. It shows that the current implementation of callsign extraction from manual transcriptions correctly extracts most of the callsigns. The rows "*gold no callsign info*" show that the recognition performance degrades, if no callsign information from the surveillance data is provided. The callsign BAWA3CC as an example is not recognized any more from the utterance "three charlie charlie continue present …".

TABLE XVIII. CALLSIGN RECOGNITION PERFORMANCE IN [%]

A = ANSP	**A 1**	**A 2**	**A 3**
CsR/E gold callsign info	99 / 0.5	99 / 0.5	99 / 0.5
CsR/E gold no callsign info	84 / 12	75 / 15	84 / 6
CsR/E Sp2T callsign info, 0 hours	78 / 16	65 / 28	67 / 17
CsR/E Sp2T no callsign info, 0 hours	53 / 27	38 / 39	37 / 30
CsR/E Sp2T callsign info, 2 hours	No data available		73 / 16
CsR/E Sp2T no callsign info, 2 hours			44 / 27
CsR/E Sp2T callsign info, 6 hours	95 / 3	No data	85 / 10
CsR/E Sp2T no callsign info, 6 hours	79 / 13		60 / 19

The following rows with "Sp2T" show the performance of the command extraction of the different Speech-To-Text engines from TABLE XV for the cases, when callsign information is used ("Sp2T callsign info") or not used ("Sp2T no callsign info") for the command extraction. With increasing the amount of ANSP dependent domain data the extraction performance increases, i.e., as expected the command extraction performance correlates with the performance of the acoustic and language model of the ASR system. At the end of the HAAWAII project, HAAWAII intends to benefit from roughly 1,000 hours of mostly non-transcribed training data.

The following TABLE XIX shows the results, when the full command and not just the callsign is considered.

TABLE XIX. COMMAND RECOGNITION PERFORMANCE IN [%]

A = ANSP	**A 1**	**A 2**	**A 3**
CmdR/E gold callsign info	99 / 0.6	99 / 1.2	100 / 0
CmdR/E, gold no callsign info	83 / 5	79 / 9	84 / 10
CmdR/E Sp2T callsign info	59 / 9	40 / 15	51 / 18
CmdR/E Sp2T no callsign info	41 / 16	23 / 24	29 / 29
CmdR/E Sp2T callsign info, 2 hours	No data		54 / 17
CmdR/E Sp2T no callsign info, 2 hours	available		33 / 27
CmdR/E Sp2T callsign info, 6 hours	86 / 5	No data	69 / 12
CmdR/E Sp2T no callsign info, 6 hours	72 / 10		49 / 23

Although no results with respect to readback error detection rate and false alarm rate are available yet, the results again show that using an Assistant Based Speech Recognition (ABSR) [37], dramatically improves both – command recognition rate and command recognition error rate. ABSR uses an assistant system to provide system dependent callsign information and predict possible commands that the ATCo may give in the near future. The ABSR improvements are observed for both callsign recognition and also for recognizing the whole command. Using additional six hours of transcribed recordings from the corresponding airport increase recognition rates for the whole command and for the callsign by only 20% absolute. Currently, the HAAWAII project has recorded more than 500 hours of silence reduced voice data. Further improvements with respect to word error rates and command recognition rates can be expected.

VIII. Conclusions and Next Steps

The presented readback error use cases clearly demonstrate that readback error detection on word level cannot be successful. More than 95% of the analyzed ATCo utterances contain a unit, (e.g., feet or knots), whereas less than 75% of the pilot utterances contain them. Instead, an abstraction of the recognized words to ATC concepts consisting of callsigns, command types, command values, etc. is necessary.

User acceptance of a readback error assistant system requires a low false alarm rate. Assuming a readback error rate of 2%, we show that a moderate command recognition rate would be sufficient. The error rate on command level, however, must be very low, i.e., an error rate of less than 0.5%, is required, if a readback error false alarm rate below 10% is intended. This either requires redundant, but at least partly independent readback error detection algorithms, or plausibility values or using context information from the corresponding surveillance data. Our current implementation based on only six hours of voice utterances from pilots and ATCos enables word error rates of 20% and 10%, respectively, which will be improved by integration of Assistant Based Speech Recognition. Nevertheless, using callsign information reduces the command recognition error rate by more than 5% absolute for each of the investigated airspaces from Isavia, NATS, and Austro Control.

Although the current implementation resulting from the first nine months of the HAAWAII project does not yet achieve the required false alarm rates, the reported results already show the direction of future work, and may help other research teams to benefit from. A joint approach will be needed so that ASR performance can increase ATM safety resulting from automatic readback and hearback error detection.

Acknowledgment

The authors want to thank all the ATCos from Austro Control, Isavia ANS, and NATS who supported during the tedious transcription process especially of the pilot utterances.

References

[1] Airbus, “Flight Operations Briefing Notes, human Performance, Effective Pilot / Controller Communications,” 2004.

[2] A. Isaac, “Effective Communication in the Aviation Environment: Work in Progress,” *Hindsight*, 5, pp. 31–34, 2007.

[3] G. Skaltsas, J. Rakas, M.G. Karlaftis, „An analysis of air traffic controller-pilot miscommunication in the NextGen environment,“ *Journal of Air Transport Management*, 27, Elsevier, pp.46–51, 2013.

[4] Flight Safety Foundation, FSF ALAR Briefing Note, 2.3 – Pilot-Controller Communication,” 2000.

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

[6] K. Cardosi, “An analysis of En Route Controller-Pilot Voice Communications,”, Tech. Rep. DOT/FAA/RD-93-11, 1993.

[7] K. Cardosi, “An Analysis of Tower (Local) Controller-Pilot Voice Communications,“ Tech. Rep. DOT/FAA/RD-94/15, 1994.

[8] O.V. Prinzo, “The computation and effects of air traffic control message complexity and message length on pilot readback performance,” in Proceedings of Measuring Behavior 2008, 6th International Conference on Methods and Techniques in Behavioral Research, Maastricht, The Netherlands, 2008.

[9] S. Chen, H. Kopald, R.S. Chong, Y.-J. Wei, and Z. Levonian, “Read Back Error Detection using Automatic Speech Recognition”, Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATMS2017), Seattle, WA, USA, 2017.

[10] G. van Es, “Air-ground communication safety study: an analysis of pilot-controller occurrences,” EUROCONTROL, 2004.

[11] D.G. Morrow, A. Lee, and M. Rodvold, “Analysis of Problems in Routine Controller-Pilot Communication,” The International Journal of Aviation Psychology, 3:4, pp. 285–302, 1993.

[12] D.G. Morrow and O.V. Prinzo, “Improving Pilot/ATC Voice Communication in General Aviation,” Tech. Rep. DOT/FAA/AM-99/21, 1999.

[13] K. Cardosi, B. Brett, and S. Han, “An Analysis of TRACON (Terminal Radar Approach Control) Controller-Pilot Voice Communications,” Tech. Rep. DOT/FAA/AR-96/66, 1996.

[14] Q. Wu, B.R.C. Molesworth, and D. Estival, “An Investigation into the Factors that Affect Miscommunication between Pilots and Air Traffic Controllers in Commercial Aviation,” *The International Journal of Aerospace Psychology*, 29, pp.53–63, 2019.

[15] O.V. Prinzo, A.M. Hendrix, and R. Hendrix, “The Outcome of ATC Message Length and Complexity on En Route Pilot Readback Performance,” Tech. Rep. DOT/FAA/AM-06/25, 2006.

[16] H. Hering, “Technical Analysis of ATC Controller to Pilot Voice Communication with regard to Automatic Speech Recognition Systems,” Tech. Rep. EEC Note No. 01/2001, EUROCONTROL, 2001.

[17] O.V. Prinzo and A.M. Hendrix, “Development of a Coding Form for Approach Control/Pilot Voice Communications,” Tech. Rep. DOT/FAA/AM-95/15, 1995.

[18] J. Bürki-Cohen, “Say Again? How Complexity and Format of Air Traffic Control instructions affect pilot recall,” in 40th Annual Air Traffic Control Association Convention Proceedings, Las Vegas, NV, USA, 1995.

[19] K. Cardosi, P. Falzarano, and S. Han, S., “Pilot-Controller Communication Errors: An Analysis of Aviation Safety Reporting System (ASRS) Reports,” Tech. Rep. DOT/FAA/AR-98/17, 1998.

[20] BEA, ”Erroneous read-back by a crew not detected by ATC, runway incursion,” 2016.

[21] M.D. Ragnarsdottir, H. Waage, and E.T. Hvannberg, ”Language technology in air traffic control,” IEEE/AIAA 22nd Digital Avionics Systems Conference (DASC), Indianapolis, IN, USA, 2003.

[22] 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,” 2020.

[23] T. Rozenbroek, “Sequence-to-Sequence Speech Recognition for Air Traffic Control Communication,” 2020.

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

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

[26] 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,” *Journal of Air Transport Management, Elsevier*, expected publication in 2021.

[27] H. Helmke, M. Kleinert, O. Ohneiser, H. Ehr, S. Shetty, “Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications,” IEEE/AIAA 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020.

[28] Y. Lu, Y. Shi, G. Jia, and J. Yang, “A new method for semantic consistency verification of aviation radiotelephony communication based on LSTM-RNN”, IEEE International Conference on Digital Signal Processing, pp. 422-426, Beijing, China, 2016.

[29] G. Jia, Y. Lu, W. Lu, Y. Shi, and J. Yang, “Verification method for Chinese aviation radiotelephony readbacks based on LSTM-RNN”, *Electronic Letters*, 53(6), pp. 401-403, 2017.

[30] G. Jia, F. Cheng, J. Yang, and D. Li, “Intelligent checking model of Chinese radiotelephony read-backs in civil aviation air traffic control”, *Chinese Journal of Aeronautics*, 31(12), pp. 2280-2289, 2018.

[31] F. Cheng, G. Jia, J. Yang, and D. Li, “Readback Error Classification of Radiotelephony Communication Based on Convolutional Neural Network”, *Biometric Recognition*, Springer International Publishing, pp. 580-588, 2018.

[32] HAAWAII homepage: www.haawaii-project.de, Highly Automatic Air Traffic Controller Workstation with Artificial Intelligence Integration, n.d.

[33] O. Ohneiser, H. Gürlük, M.-L. Jauer, A. Szöllősi, and D. Balló, "Please have a Look here: Successful Guidance of Air Traffic Controller's Attention," 9th SESAR Innovation Days, Athens, Greece, 2019.

[34] STARFiSH, research project funded by the German Federal Ministry of Education and Research, see for further information https://www.softwaresysteme.pt-dlr.de/de/ki-in-der-praxis.php, in German, n.d.

[35] PJ.05-97-W2 SESAR2020 funded industrial research projects under the European Union's grant agreement 874464, see for further information https://www.remote-tower.eu/wp/?page_id=888, and https://www.sesarju.eu/index.php/projects/DTT, n.d.

[36] PJ.10-96-W2: SESAR2020 funded industrial research projects under the European Union's grant agreement 874470, https://cordis.europa.eu/programme/id/H2020_SESAR-IR-VLD-WAVE2-10-2019/de, n.d.

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

[38] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies. 2 (1): pp. 37–63, 2011.

[39] Y. LeCun, Y. Bengio et al., "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol. 3361, no. 10, 1995.

[40] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in Interspeech, 2018, pp. 3743–3747.

[41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in IEEE workshop on automatic speech recognition and understanding, no. CONF. IEEE Signal Processing Society, 2011.

[42] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in Interspeech, 2016, pp. 2751–2755.

[43] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[44] S. Pigeon, W. Shen, A.D. Lawson, and D. van Leeuwen, "Design and characterization of the non-native military air traffic communications database (nnmatc)," in Eighth Annual Conference of the International Speech Communication Association, 2007.

[45] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The hiwire database, a noisy and non-native english speech corpus for cockpit communication," Online available: http://www. hiwire. org, 2007.

[46] K. Hofbauer, S. Petrik, and H. Hering, "The atcosim corpus of non-prompted clean air traffic control speech," in LREC, 2008.

[47] E. Delpech, M. Laignelet, C. Pimm, C. Raynal, M. Trzos, A. Arnold, and D. Pronto, "A Real-life, French-accented Corpus of Air Traffic Control Communications," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[48] J. Godfrey, "The Air Traffic Control Corpus (ATC0) - LDC94S14A," 1994. Online available: https://catalog.ldc.upenn.edu/LDC94S14A.

[49] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in Proc. of the 18th Annual Conference of the International Speech Communication Association, 2017.

[50] J.R. Bellegarda, "Statistical language model adaptation: review and perspectives," Speech communication 42.1: pp. 93–108, 2004.

[51] H. Arilíusson, T.S. Simignnoschi, H. Helmke, J. Harfmann: HAAWAII project: D1-1 Operational Concept Document; version 1.0, July 2020.

[52] H. Arilíusson, T.S. Simignnoschi, H. Helmke, J. Harfmann: HAAWAII project: D6-2 Updated Operational Concept Document; version 1.01, April 2021.

AUTHOR BIOGRAPHY

Hörður Arilíusson is working as Air Traffic Controller at Isavia ANS since 1985 and since 2012 managing different projects within R&D of Isavia´s ANS having an extensive experience of Project Management and system implementation of ATM Systems.

Heiko Ehr is Mathematic-technical Assistant. He joined DLR's Institute of Flight Guidance in 1997. Mr. Ehr worked on several projects such as 4D-Planer and its successor 4D-CARMA, Flow Monitor, and Traffic Monitor. He is an expert in system integration and interface development.

Julia Harfmann received her Master's degree in Psychology from the University of Graz (Austria). She worked for the Safety Department at Austro Control before joining NATS in 2017 where Julia focuses on advancing automation in air traffic control both internally as well as in the SESAR body of work.

Hartmut Helmke holds a Diploma degree in Computer Science and a doctor degree in Chemical Engineering from Stuttgart University. Since 1989 he is with DLR. He led the Speech Recognition Projects AcListant®, AcListant®-Strips, MALORCA, PJ.16-04-ASR. Currently he is leading the HAAWAII project. Prof. Helmke is an assistant professor for Computer Science since 2001.

Matthias Kleinert received his Master degree in Computer Science from Technical University Braunschweig (Germany). After finishing his semester at the University of Wisconsin Parkside, in Kenosha, USA he joined DLR's Institute of Flight Guidance in 2012. Currently he concentrates on machine learning ASR applications.

Karel Ondřej is a PhD student at the Faculty of Information Technology, Brno University of Technology, Czech Republic. His research interests include advanced machine learning approaches, natural language processing, and human-machine interaction.

Petr Motlicek (Senior Member, IEEE) received a M.S. degree in electrical engineering and a Ph.D. degree in computer science from the Brno University of Technology, Brno, Czechia, in 1999 and 2003, respectively. He is currently a Senior Research Scientist with Idiap Research Institute, Martigny, and holds a position of an Assistant Professor with the Brno University of Technology.

Oliver Ohneiser received his master degree in Computer Science and his doctor degree (PhD) in Aerospace Engineering from the Technical University of Braunschweig (Germany) in 2011 and 2017, respectively. He joined DLR in 2006 and is a scientific assistant in the department "Controller Assistance" of the Institute of Flight Guidance in Braunschweig since 2010. Mr. Ohneiser investigates modern interaction technologies at controller working positions.

Amrutha Prasad received her Master's degree in Artificial Intelligence from Distance University, Switzerland in 2020. Currently, she is working as a Research Assistant at Idiap Research Institute, while pursuing her PhD in Speech Processing at Brno University of Technology (BUT), Czech Republic.

Shruthi Shetty received her master's degree in Computer Science from Technical University Braunschweig (Germany) in 2019. In 2020, she joined DLR at the Institute of Flight Guidance. Currently, she develops software for ASR and machine learning applications.

Teodor S. Simiganoschi received his Bachelor of Engineering degree in Computer Science from the University of Suceava, Romania in 2000. He worked for different companies like Rohde and Schwarz, Euroweb, and Bit Telecom Romania. Currently he is working as Project Manager in R&D at Isavia ANS.

Pavel Smrz is an associate professor at the Faculty of Information Technology, Brno University of Technology, Czech Republic. He leads the Knowledge Technology Research Group, focusing on multimedia processing, hardware-accelerated machine learning, and human-machine interaction, embedded intelligence, and big data processing. He also leads the BUT's team in HAAWAII.

Karel Veselý is a junior researcher at the Faculty of Information Technology, Brno University of Technology, Czech Republic. He is one of the co-authors of OpenSource speech recognition toolkit Kaldi. He specializes on machine learning and customizing automatic speech recognition technology for applications.

Christian Windisch is working at Austro Control as a Senior-Expert for Planning & Development in ATC Operations and as a Supervisor and Air Traffic Controller for Approach Vienna. He studied Computer Science at Technical University Vienna, and worked for IBM as a Technical Trainer for UNIX Systems. He is Senior Specialist for Planning of ATM Systems.