Final Project Results Report

Deliverable ID:	**D5-5**
Dissemination Level:	**PU**
Project Acronym:	**HAAWAII**
Grant:	**884287**
Call:	**H2020-SESAR-2020-2**
Topic:	**SESAR-ER4-18-2019**
Consortium Coordinator:	**DLR**
Edition date:	**21st October 2022**
Edition:	**02.00.00**
Template Edition:	**02.00.04**

EUROPEAN PARTNERSHIP

Co-funded by
the European Union

Authoring & Approval

Authors of the document

Name / Beneficiary	Position / Title	Date
Hartmut HELMKE (DLR)	Project Lead	03.10.2022
Oliver OHNEISER (DLR)	Dissemination Manager	13.06.2022
Pavel SMRZ (BUT)	WP4 Leader	18.07.2022
Petr MOTLICEK (Idiap)	WP3 Leader	08.08.2022

Reviewers internal to the project

Name / Beneficiary	Position / Title	Date
Hartmut HELMKE	Project Lead	21.10.2022
Julia HARFMANN	WP5 Leader	05.08.2022
Johannes Blümel	WP6.1 Leader	05.08.2022
Shruthi Shetty	Contributor	05.08.2022

Reviewers external to the project

Name / Beneficiary	Position / Title	Date

Approved for submission to the S3JU By - Representatives of all beneficiaries involved in the project

Name / Beneficiary	Position / Title	Date
Hartmut HELMKE	Project Lead	21.10.2022
Pavel SMRZ (BUT)	WP4 Leader	06.08.2022
Petr MOTLICEK (Idiap)	WP3 Leader	19.10.2022
Johannes Blümel (ACG)	New D6.1 Leader	via email to HHe 5.8.2022 and by review
Teodor S. SIMIGANOSCHI (Isavia)	WP1 Leader	20.10.2022
Julia HARFMANN (NATS)	WP5 Leader	via email to HHe 5.8.2022 and by review
Martina ERCEG (CCL)	Contact Point CCL	05.08.2022

Rejected By - Representatives of beneficiaries involved in the project

Name and/or Beneficiary	Position / Title	Date

Document History

Edition	Date	Status	Name / Beneficiary	Justification
1.00	08.08.2022	Final	HHe/DLR	Prepared for final submission to SJU, reviewed comments of PM and OO, included a sentence that MA was based TRL1 criteria, but STELLAR contains TRL MA
1.01	03.10.2022	Draft	HHe/DLR	Accepted all previous changes, delete all previous comments, deleted old history
			HHe/DLR	Update of MA tables to the correct table and made consistent with contents on STELLAR, correcting spelling errors
			HHe/DLR	Changed PER.ER.2 and PER.ER.3 from “Not Applicable” to “Partial Non-Blocking”
			HHe/DLR	Addition “ASR as a service” to the recommendations, addition link to EUROCAE working group to recommendation
2.00	21.10.2022	Final	HHe/DLR	Prepared for submission to SJU

Copyright Statement

HAAWAII

HAAWAII

This document is part of a project that has received funding from the SESAR3 Joint Undertaking under grant agreement No 884287 under European Union's Horizon 2020 research and innovation programme.

Abstract

Advanced automation support developed in Wave 1 of SESAR IR includes using of automatic speech recognition (ASR) to reduce the amount of manual data inputs by air traffic controllers (ATCos). Evaluation of ATCos' feedback has been subdued back then due to the limited recognition performance of the commercial-off-the-shelf ASR engines that were used, even in laboratory conditions.

The HAAWAII project successfully addresses all these problems and developed a reliable, error resilient, and adaptable solution to automatically transcribe voice utterances of both ATCos and pilots and automatically extract their semantics. The project builds on a large collection of ATCo and pilot utterances including the corresponding surveillance data of Icelandic en-route airspace and London Terminal Manoeuvring Area (TMA).

The results of the proof-of-concept trials of the HAAWAII project show that automatic speech recognition and understanding (ASRU) using the HAAWAII framework and architecture is ready for use for applications such as

- Prefilling Controller-Pilot Data Link Communications (CPDLC) message supported by ASRU
- Prefilling radar labels by ASRU
- ATCo workload prediction
- Callsign highlighting
- Post-Evaluation of historic data with respect to readback errors.

In detail the following performance results are achieved within the project on voice data directly from the noisy operational environment, which include all the phraseology deviations of real pilots and ATCos:

- Word error rates (WER) of 3% for ATCo utterances and below 7% for NATS pilots and below 11% for Isavia pilots,
- Command recognition rates of 94% for NATS and 86% for Isavia ATCo utterances,
- Command recognition rates of 82% for NATS and 71% for Isavia pilot utterances,
- Callsign recognition rates of 98.5% and 96.9% for NATS and Isavia ATCo utterances, respectively,
- Callsign recognition rates of 95% and 88% for NATS and Isavia pilot utterances, respectively.

These results enable readback error detection rates of 80%, an easy adaption of the HAAWAII architecture to other airports by machine-learning based training of the speech recognition models, ATCo workload reduction by early callsign highlighting and pre-filling the radar label contents. Standardization of the ontology based semantic extraction is prepared by extending the ontology to enroute traffic and specially to pilot utterance. Standard logical interfaces, are defined between the building blocks of an ASRU architecture (Speech-to-Text, Text-To-Concept, Command Prediction, HMI). Ontology and standard interface enable a comparison on performance level of different ASRU implementations and an easy exchange of implementations, which both strengthen the competitiveness of European ATM industry.

This is the second version of the report, considering updates of the recommendation for next R&D phase.

Table of Contents

List of Tables

List of Figures

1 Executive Summary

The proof-of-concept trials running as part of HAAWAII exercises for using speech recognition and understanding for radar label maintenance and integration with CPDLC were successful. Both offline evaluations on historical voice and surveillance data from the ops room and proof-of-concept trials involving ATCos from NATS and Isavia ANS have been conducted in May and June 2022.

Overall, command recognition rates of 94% and 86% were achieved for voice recordings from the ops room for NATS and Isavia, respectively. The threshold resulting from the objectives was set to 85%. This is much better than the baseline of the HAAWAII project, described in the Grant Agreement (i.e., mentioned as part of KPIs of the project). The baseline includes command recognition rates of 70% for Prague approach and of only 55% for Vienna approach, , resulting from the lab environment trials from solution PJ16-04 of SESAR-2.

In terms of word error rates (WER), WERs of 3% for ATCos and below 7% for pilots in the UK airspace and below 11% for pilots in the Icelandic airspace were achieved in HAAWAII. Performance on word level correlates with performance on the semantic level, i.e. to command recognition rates and to callsign recognition rates. On callsign level the HAAWAII project achieved 88% for pilot data of Isavia and 95% of NATS, which corresponds to the 71% and 82% on command level for the noise pilot voice recordings.

The extraction performance for ATCo is even better. For NATS the performance of 98.5% for callsign recognition for ATCo data is achieved, whereas for Isavia we get 96.9%. The difference might be surprising, because the results on word level are better for Isavia than for NATS. The reason, however, is that some aircraft callsign are not present in the surveillance data provided by Isavia, which makes it impossible to guess that “one alfa delta” corresponds to “arctic eagle one alfa delta” or to correct wrong recognitions on word level. “ice air four to nine” or “eyes air for two nine” could be easily mapped to ICE429, if it is known, that this callsign is in the air.

Two approaches for read back error detection have been implemented. The first data driven approach is capable of fast adaptation to a new environment but suffers from very few training samples. Only 90 readback error use cases were extracted from the four hours of annotated Isavia data. Of these use cases only one third were real readback errors, the rest were mostly missing readbacks of the pilot. Nevertheless, the second ontology-based approach (i.e., so-called the rule-based approach) discovers 4 out of 5 readback error use cases. The results obtained from the exercises in HAAWAII clearly show the potential of a readback error detection assistant (REDA), which is achievable even with the current implementation. Validation trials first in lab environment, but very soon in ops room are recommended to figure out whether a false alarm of 10% is really needed or if a false alarm rate of 50% is also acceptable. This seems to be achievable, when more voice recordings from pilots are included into the training data sets and when the speech quality values from the voice receivers is included in the decision process. Possible readback alerts resulting from bad voice quality should be shown to the ATCo.

Furthermore, the ontology for semantic interpretation of air traffic control (ATC) utterances was extended in the project, considering pilot utterances and en-route ATCo-pilot conversation. Standardization and usage by other SJU-funded speech recognition projects is recommended to enable comparison on the semantic level. HAAWAII project has also standardized the exchange between the different building blocks, i.e. between Speech-To-Text, Text-To-Concept, Command Prediction and the ATC application, so that the ambiguity and plausibility values are also considered.

The HAAWAII project suggested new enablers (ER) and also new OI steps (operational improvement steps), which consider workload prediction and readback error detection support by automatic speech recognition and understanding.

Standardization of the ontology based semantic extraction is prepared by extending the ontology to enroute traffic and specially to pilot utterance. Standard logical interfaces, are defined between the building blocks of an ASRU architecture (Speech-to-Text, Text-To-Concept, Command Prediction, HMI). Ontology and standard interface enable a comparison on performance level of different ASRU implementations and an easy exchange of implementations, which both strengthen the competitiveness of European ATM industry.

Last, but not least: the HAAWAII ASRU system was successfully tested even in the operational environment of Isavia ANS, the Icelandic air navigation service provider. The HAAWAII ASRU was directly connected with the controller working position of the voice communication system receiving direct audio of the ATCO and Pilot communication. At the same time live surveillance data was streamed towards the ASRU over ethernet from the Isavia ANS tracker to improve callsign recognition. The tests were performed in the Isavia ANS Air Traffic Control Centre at the 29th of June 2022, a great success for a Research Project aiming for TRL2, which is achieved as shown in the performed assessment available via STELLAR.

2 Project Overview

2.1 Operational/Technical Context

The HAAWAII project contributed to the application area 1 "**Advanced automation support for en-route and TMA"** of the topic "**Automation and CWP**" of the SESAR Exploratory Research Call 4 (ER4).

The Controller Working Position (CWP) is the centre of the human-machine interactions for Air Traffic Controllers (ATCos). The tools and platforms, used by controllers, shape and inform their way of operating, their mental models and ultimately their performance. This in turn promotes the safe and efficient management of air traffic. Voice communications between ATCos and pilots is still not digitalised and, therefore, not accessible for machine analysis.

To change the current status, the SESAR Master Plan vision includes increasing digitalisation and automation of Air Traffic Control (ATC). While data communication or text-based transmission of data, between ATCos and pilots, is envisioned to supplement radio communications in future operating environments, this capability is unlikely to completely replace radio communications in the near term [43]. Furthermore, data link and voice communication should not be seen any more as competitors, but as complementary: Even the data elements exchanged via data link need to be input into the system. This could be done for example by mouse, keyboard, touch-pad or also via voice.

Applications from digitalisation of voice communication must be identified and exploited. Automatic Speech Recognition (ASR) technologies are increasingly deployed across many industries. Recently, solution PJ.16-04 of Wave-1 has demonstrated that the current ASR systems are mainly targeting the every-day consumer market and are not suited to safety and time critical application areas such as Air Traffic Management (ATM). As shown in the past, trust and acceptance by ATCos of low fidelity tools is an obstacle for deployment.

Nevertheless, several past projects have indicated the usefulness and utility of ASR: (i) AcListant® quantified the benefits with respect to workload reduction and performance increase [AcListant, Helmke17] [1], (ii) MALORCA project demonstrated commercial viability and enhanced models [MALORCA, Kleinert18], (iii) SESAR 2020 solution PJ.16-04 explored industrial integration and requirements building.

So far, these precursor projects have focused on research and development of a set of models, further deployed in relatively simple and controlled application domains. The HAAWAII project has targeted complex and challenging environments and, more importantly, wider applications of automatically recognized voice communications. More specifically, HAAWAII has proposed in the grant agreement two general objectives [51]:

1. Research and develop data-driven (machine learning oriented) approaches to be deployed for novel and complex environments from two large ANSPs, demonstrating an increased validity of the tools; Emphasis is on ops room data and not on voice and surveillance data from the laboratory environment.

---

[1] A list of references and abbreviations is provided in the appendix after section 5.

2. Demonstrate the wider applicability of the tools in ATM, focussing on generating benefits for the ATCos and the ANSPs, i.e., reducing workload, increasing both efficiency and safety.

Overall, HAAWAII focused on the following applications:

- Pilot readback error detection,
- Callsign highlighting especially from pilot utterances,
- Pre-filling radar labels and CPDLC messaging using the automatic speech recognition,
- Human performance metric extraction to predict ATCo's workload.

2.2 Project Scope and Objectives

The main messages, i.e. the project scope of the HAAWAII project are:

- HAAWAII aims to increase air traffic controller (ATCo) productivity by pre-filling radar labels and call-sign highlighting.
- HAAWAII aims to do the first, but significant step to increase air traffic management (ATM) safety by automatically detecting readback errors.
- HAAWAII aims to show that modern speech recognition techniques can bring a high accuracy, even when applied on pilot speech from a potentially noisy channel.
- HAAWAI aims to reduce ATCOs' workload. The digitization of controller and pilot voice utterances can be used to pre-fill entries into electronic flight strips and CPDLC messages.
- HAAWAII aims to objectively estimate ATCOs' workload utilising digitized voice recordings of the complex London TMA.

From these high-level messages of the project the following objectives were derived:

- Obj1: Exploiting massive amounts of unlabelled voice data through new unsupervised learning algorithms.
- Obj2: Automatic speech recognition of the controller and pilot communication for the very complex London TMA and for Isavia's enroute airspace.
- Obj3: Automatic detection of pilot readback errors.
- Obj4: Pre-filling of radar labels and CPDLC messaging.
- Obj5: Improve ATCo staffing, rostering and flow management planning and reaction for the London TMA by measuring and anticipating the workload from voice communication.
- Obj6: Sufficiently consider data privacy issues.

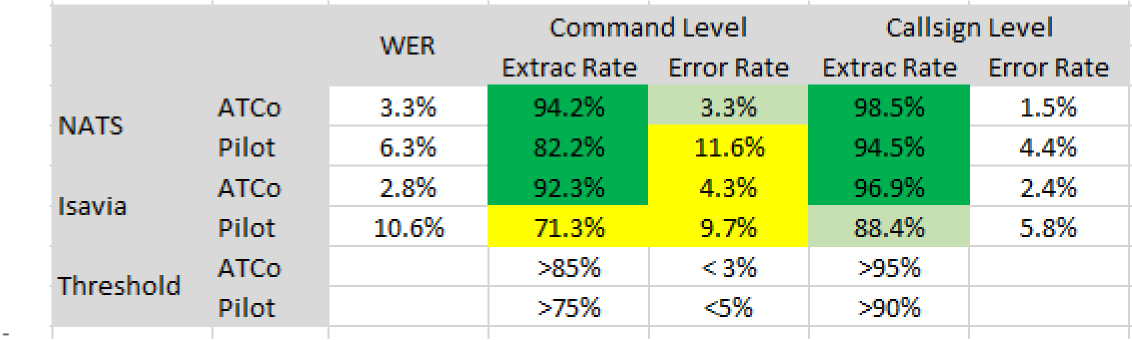
2.3 Work Performed

The following table detail the HAAWAII objectives of section 2.2 and mentions the deliverables which cover these objectives. If baseline performance from the MALORCA project is provided, it is related to Prague approach. In case of two performance values, the latter one is related to the Vienna approach. The KPIs are detailed in the "Glossary of terms" in the appendix A.1. For ease of reading the tables (coped from Part B of the grant agreement) are extend by one row, containing the results. The main results are summarized in the next section 2.4

Obj1: Exploiting massive amounts of unlabelled voice data through new unsupervised learning algorithms to training an Assistant Based Speech Recognition (ABSR) system		
Key performance indicators (KPI)	**Baseline** from MALORCA	**Target**
KPI 1.1: Amount of voice and surveillance data	150 hours	appr. 10'000 hours
KPI 1.2: Improvement of automatic speech recognition models through the deployment of unlabelled data	Only labelled data was used	10% absolute; 25% relative
Relevant deliverables	D2.2, D2.3, D3.4, D3.5	
Results: • NATS voice data cover a time frame of 1005 hours, which corresponds to 403 hours of speaking time, i.e. voice channel loading is 40%. In total 383,450 voice wave files are recorded • Isavia voice data cover a time frame of 2125 hours. The voice channel loading is here 11%, so that we can expect 235 of speaking time, when silence is removed • The improvement of using unlabeled data is described in D3-5. The finally achieved word error rates are, nevertheless, excellent. For Isavia 2.8% WER for ATCos, 10.6% for pilots; for NATS 3.3% for ATCOs and 6.3% for pilots		

Obj2: Automatic recognition of the controller and pilot communication for London TMA and Isavia enroute airspace		
Key performance indicators	**Baseline** from MALORCA	**Target**
KPI 2.1: average command recognition rate	92% / 85%	> 85% for ATCos; > 75% for pilots
KPI 2.2: average command recognition error rate	0.6% / 3.2%	< 3% for ATCos < 5% for pilots
KPI 2.3: average callsign recognition rate	98.5 / 95%	> 95% for ATCos; > 90% for pilots
Relevant deliverables	D3.4, D3.5, D5.3	

Results:



		WER	Command Level		Callsign Level	
			Extrac Rate	Error Rate	Extrac Rate	Error Rate
NATS	ATCo	3.3%	94.2%	3.3%	98.5%	1.5%
	Pilot	6.3%	82.2%	11.6%	94.5%	4.4%
Isavia	ATCo	2.8%	92.3%	4.3%	96.9%	2.4%
	Pilot	10.6%	71.3%	9.7%	88.4%	5.8%
Threshold	ATCo		>85%	< 3%	>95%	
	Pilot		>75%	<5%	>90%	

The green cells show the achieved threshold. Light green shows the nearly achieved threshold. Not all metrics got threshold in the grant agreement.

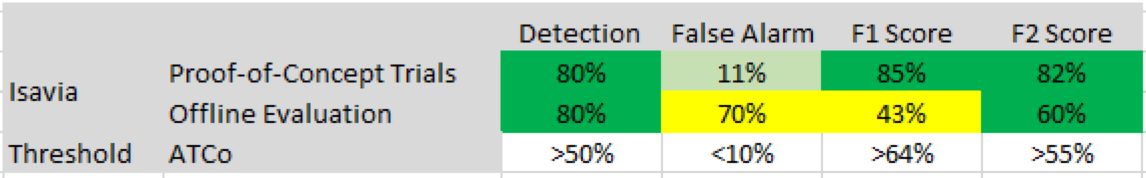
The recognition rates are mostly achieved, whereas the error rates show possibilities for improvements.

Obj3: Automatic detection of pilot readback errors

Key performance indicators	**Baseline**	**Target**
KPI 3.1: Read-back error detection rate	not established	at least 50%
KPI 3.2: False alarm rate for readback error detection	not established	below 10%
Relevant deliverables	D5.1, D5.2	

KPI3.1 and KPI3.2 can be combined to an F1 score of at least 64% or a F2 score of 55%, respectively.

Results:

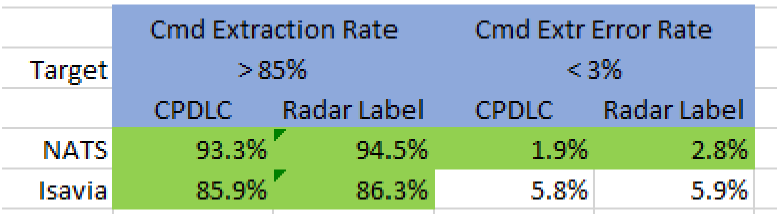


		Detection	False Alarm	F1 Score	F2 Score
Isavia	Proof-of-Concept Trials	80%	11%	85%	82%
	Offline Evaluation	80%	70%	43%	60%
Threshold	ATCo	>50%	<10%	>64%	>55%

The challenge is to achieve a lower false alarm rate. The detection rate, however, is very good.

Obj4: Pre-filling of radar labels and CPDLC messaging via non-integrated application (proof-of-concept)		
Key performance indicators	**Baseline**	**Target**
KPI 4.1: Accuracy on pre-filled radar label contents (with command type and command value) and CPDLC messages ready to be sent to on board systems	70% for Prague in PJ.16-04 and 55% for Vienna in PJ.16-04	90%
KPI 4.2: Requirements for system integration	not available	defined
Relevant deliverables	D1.1, D1.2, D6.2, D6.3	

Results:



	Cmd Extraction Rate		Cmd Extr Error Rate	
Target	> 85%		< 3%	
	CPDLC	Radar Label	CPDLC	Radar Label
NATS	93.3%	94.5%	1.9%	2.8%
Isavia	85.9%	86.3%	5.8%	5.9%

Obj5: Improve ATCo staffing, rostering and flow management planning and reaction for the London TMA by measuring and anticipating the workload from voice communication		
Key performance indicators	**Starting point**	**Target**
KPI 5.1: User acceptance and usability analysis confirms that ASR performance to be sufficient for ATCo shift planning	not established	>75% of supervisors and ATCOs
KPI 5.2: Perceived workload reduction when predicting task load, also used as indicator of objective metrics validity	not established	confirmed by > 80% of controllers
Relevant deliverables	D1.1, D4.1, D5.4	
The performed proof-of-concept trials concentrate on different metrics, which are explained in D5.4.		

Obj6: Data privacy issues are sufficiently considered, i.e. minimum amount of anonymized data stored		
Key performance indicators	**Baseline**	**Target**
KPI 6.1: Controllers' and pilots' recordings are anonymized so that handover to HAAWAII is possible	not established	Anonymization legally approved
KPI 6.2: Protection of personal data concept by all relevant data protection authorities and ethical advisor	not available	approved
Relevant deliverables	D8.1, D8.2, D2.4	
Fully achieved and documented in D8.1 and D8.2.		

Additionally, it turns out during the life-time of the project that a general approach for the semantic interpretation of ATC utterances is needed. As already Alan Turing has pointed out in the early 1950s, speech recognition, i.e. transform a voice signal to a sequence of words, is NOT speech understanding. Therefore, the HAAWAII project has also extended the ontology, being started by PJ.16-04 of Wave-1, by pilot utterances, en-route communication and conditional clearances.

The speech-to-text itself was also more challenging than expected. Due to Covid-19 restrictions no direct access to ops room data was possible, i.e. voice utterance training data was processed by the ANSPs before handover to research. The speaker information, i.e. ATCo or pilot speaking was lost, and push-to-talk information was not available. A voice signal detection algorithm was needed to decide when speech starts and ends. Additionally, automatic speaker classification was needed. Both challenges will not occur in the operational environment, when HAAWAII results go to the next TRL phase.

2.4 Key Project Results

Key HAAWAII results include:

- Several types of conventional automatic speech recognition models were researched and developed based on the Iceland enroute airspace and London TMA for both ATCo and pilot speech. The models reach excellent results (the word error rate – WER – is for unseen data below 6% for controllers' speech and below 12% for pilots' utterances) (D3-4, D3-5).

- Proof-of-concept implementation of the real time system with continuous recognition and recognized ATC concepts according to ATC ontology as outputs, which can run in Docker on an ordinary notebook (D5-3).
- The applications for callsign highlighting, supporting ATCos in generating CPDLC messages, and pre-filling recognized ATC concepts and command values into radar labels to reduce ATCos' manual workload were evaluated in D5-3. The callsign recognition rate is 97.7% with an error rate of 1.8% for callsigns spoken by the air traffic controller. For command types relevant for pre-filling CPDLC message or pre-filling radar labels, the system achieves a command extraction rate of 94% with an error rate of 2.5% for the NATS ground data.
- Two alternative implementations for readback error detection were done. One is based on the ontology and manually implemented rules. The second one, the data-driven approach, based on a attention-based neural Network architecture, borrowed from natural language processing domain. It is automatically trained on artificially generated data. The ontology -based approach is documented in D5-1 and D5-2. Results of the data-driven approach can be found in D5-2.
- Proof-of-Concept trials for the Readback-Error-Detection Assistant (REDA) were performed with Isavia's ATCos. Artificial situations with challenging ATCo and pilot utterances were analysed. Recording conditions were good. Two third of the communication event contain readback error use cases. Readback Error Detection Rate (Recall) was 80%. The False Alarm Rate was only 11%, i.e. Precision was 89% (D5-2), resulting in an F1 score of 85%.
- Repeating the REDA performance evaluation experiment on noisy real-life data from Isavia's ops room environment from 2020 resulted in a Readback Error Detection Rate of 81%, whereas the False Alarm Rate drops down to 69%. Due to the fact, that readback error use cases are seldom events, i.e. only 2% of the communication events contain readback error use cases, the probability of a false alarm is still only 5%, see D5-2 for more details.
- The ontology for ATC utterance annotation, being started by PJ.16-04 of Wave-1, was extended by pilot utterances, en-route communication and conditional clearances. The ontology was made available for the IR solution #96 and solution #97 (D1-1, D3-1, D6-2.
- A voice signal detection algorithm and a speaker classification algorithm were implemented (D2-2).

- Proof-of-concept trials of understanding and predicting changes in ATCo's workload show that speech recognition data is reliable enough to be used to extract workload related data. Objective data correlates with subjective workload scores (D5-4).

2.5 Technical Deliverables

Reference	Title Description	Delivery Date[2]	Dissemination Level[3]
D1.1	This document contains the operational concept of the HAAWAII project. It addresses the high-level Automatic Speech Recognition use cases read-back error detection, ATCO workload assessment, callsign highlighting, and integration of speech recognition with CPDLC, radar label prefilling, and consistency checking of manual versus verbal input. It is a living document. The final version will be submitted as D6.2 at the end of the project.	19/08/2020	PU
D1-2	This document is not legally binding the consortium partners to fully comply with the written requirement. It just, contains the System Requirements of the HAAWAII project describing in detail each requirement. D1-2 shall be used as a guidance during the development and implementation process of the HAAWAII speech recognition and understanding system. It is a living document. The final version will be submitted as D6.3 at the end of the project.	12/10/2020	CO
D1-3	This deliverable details the architectures of the module interaction for all HAAWAII applications. This includes the architecture of ABSR for online applications, which requires analysis of a continuous voice stream from controllers and pilots, decision on where a command starts/ends and who speaks, etc. In most of the HAAWAII applications this requires a reaction from ABSR in real-time. Besides that, the document also describes the architectures for model training from collected data (offline mode) and the architecture for the online applications itself. Furthermore, the necessary interfaces building on a JSON format are detailed.	05/11/2020	CO

---

[2] Delivery data of latest edition

[3] Public or Confidential

D2-1	For testing the algorithms that automatically derive meta data, test data is necessary. The labelled data from NATS and Isavia were not available during the first months of the project. Therefore, it was decided to reuse data already recorded and labelled by Austro Control. Deliverable D2-1 describes the data formats of the speech data, surveillance data, and, especially, the static data (e.g., runway names, waypoints), which are needed to transform voice utterances into word sequences and into meaningful ATC concepts (e.g., callsigns, command types, and command values).	03/08/2020	CO
D2-2	This deliverable consists of the surveillance data and voice recordings from London TMA. The text consists of a short report, which contains a description of the recorded data from London TMA. The surveillance data was available for 46 days for August and September 2020 always from 6 a.m. to 6 p.m. GMT. Voice recordings were available for the same time frame and for two frequencies, i.e. TMA SOUTH and Heathrow Approach sector (LLAP).	25/02/2021	CO
D2-3	This deliverable consists of the surveillance data and voice recordings from Isavia's enroute airspace. Similarly, to D2-2, it also consists of a short report, which contains a description of the recorded data from Isavia's enroute airspace. The surveillance data was available for 46 days in July to September 2020 always from 0 a.m. to 23:59 p.m. GMT. Voice recordings were available for the same time frame for two working positions.	04/03/2021	CO
D2-4	The document provides a detailed description of how the HAAWAII projects manages data sharing. It outlines the nature of the data as well as the FAIR principles, which was adhered to by all project partners. Furthermore, data management is presented in detail. This covers the whole project life cycle from data acquisition to storage, preservation and access. It also describes data sharing principles towards the overall data management plan. This document served as a reference for all project partners to ensure consistent data management throughout the project.	25/01/2021	CO
D3-1	This document provides a detailed description of the transcription and annotation of air traffic control utterances in the HAAWAII project, i.e., it describes the process from the audio files via word-by-word content files to words-to-concepts files in the respective formats. This includes all the important features of BUT's transcription tool SpokenData as well as DLR's transcription and annotation tool CoCoLoToCoCo. It also provides a complete guide to help new users to use the tool and carry out transcriptions and partly annotations with ease.	05/11/2020	PU

D3-2	This deliverable concentrates on the semantic interpretation, i.e., the annotation of the transcribed voice recordings by using the information from corresponding voice recordings from NATS. At the time of its submission, 7.5 hours of manually transcribed pilot and ATCo utterances were available from London airspace. Utterances corresponding to about 57 minutes of voice data were manually annotated, while the remaining 6.5 hours were annotated automatically.	10/06/2021	PU
D3-3	This deliverable concentrates on the semantic interpretation, i.e. the annotation, of the transcribed voice recordings by using the information from corresponding voice recordings from Isavia. At the time of its submission, 7.5 hours of manually transcribed pilot and ATCo utterances were available from Isavia airspace. 90 minutes of them were manually annotated, the remaining 6 hours were automatically annotated.	25/05/2021	PU
D3-4	This deliverable provides a detailed overview of the architecture developed for automatic processing of speech, as well as for command prediction and extraction tasks in HAAWAII project. Although suggested by the title, this deliverable describes and evaluates the built models dealing with ATC speech in general, not only with the ATCo's speech exclusively. To be complete, it also reports preliminary results for the pilots' speech, that was originally planned to be included in deliverable D3.5 only. Most of the components developed, implemented, and described in this deliverable are based on very recent machine learning (data driven) methods. The deliverable demonstrates the architectures for machine learning in an off-line mode, as well as the real-time architecture for the online applications.	31/01/2022	CO
D3-5	This deliverable is a follow-up of D3.4 bringing updates to the described and evaluated models dealing with ATC speech. It focuses mainly on evaluation of proposed models developed earlier, concretely: an update of different speaker role classification models and discussion of their possible combination, effective model transfers to a new airspace area, as well evaluation of different interaction modes with focus on speaker role specific acoustic models proposed in D1-3; an update on automatic speech recognition models, and their use in domain transfer tasks, including the use of very recent end-to-end ASR frameworks; the evaluation of several interaction models for online speech recognition.	13/05/2022	CO
D4-1	This document provides a detailed outline of the evaluation concept for Human Performance (HP) Metrics assessments as well as other applications of the	24/09/2021	CO

	system. It provides an overview of the approach that was taken to ensure appropriate evaluation exercises take place which enable the gathering of the evidence required to demonstrate the feasibility of the system. D4-1 also describes background information on HP metrics currently used in ANSP operations and how the speech recognition technology is foreseen to be embedded as another puzzle piece to gain better insights into human performance.		
D5-1	This deliverable focuses on the first results with respect to Readback Error Detection Assistant (REDA) prototype, which relies on the semantic interpretation of the air traffic controller (ATCo) to pilot communication, i.e. the ontology implementation. The results show that the simple use cases “command, readback correct/wrong, corrected or not” need to be extended, because the pilot’s or ATCo’s answer could also be “say again” or a request for more information.	18/11/2021	CO
D5-2	This deliverable presents the proof-of-concept results of the HAAWAII project with focus on data-driven readback error detection that can be used as support indication to improve safety in ATC environment. During the proof-of-concept trials different findings were discovered and issues that must be considered in future projects.	06/07/2022	CO
D5-3	This deliverable presents proof-of-concept results of the HAAWAII project with focus (i) on real time capabilities, i.e., continuous recognition and output of the recognized ATC concepts, (ii) callsign highlighting (iii) on supporting ATCos for generating CPDLC messages and (iv) on pre-filling recognized ATC concepts and command values into radar labels to reduce ATCos’ manual workload. A callsign recognition rate of 97.7% with an error rate of 1.8% for callsigns spoken by the air traffic controller is achieved. For command types relevant for pre-filling CPDLC message or pre-filling radar labels a command extraction rate of 94% with an error rate of 2.5% for the NATS ground data is observed.	22/06/2022	CO
D5-4	The deliverable presents the proof-of-concept results of the HAAWAII project with focus on understanding and predicting changes in Air Traffic Controller Officer (ATCO) workload and, therefore, overall human performance. The findings show that speech recognition data is reliable enough to be used to extract workload related data. The results indicate that this objective data correlates with subjective workload scores. It can be concluded that the successful proof-of-concept results need to trigger further validation studies to ensure the data is of value to the operation.		CO

D5-5	This deliverable		PU
D6-1	This deliverable summarizes the dissemination of the HAAWAII project by conducting Stakeholder workshops. It gives a report summarizing the first Stakeholder workshop conducted at the end of June 2021 and of the second Stakeholder Workshop conducted end of September 2022.	Part of exploitation phase, submission after D5-5	PU
D6-2	This deliverable is an update of D1-1 and contains the findings added to D1-1 during the project. The document was updated during the last months considering the feedback from SJU and especially from IFATCA.	Part of exploitation phase, submission after D5-5	PU
D6-3	Updated Requirements Document	Part of exploitation phase, submission after D5-5	PU
D6-4	This document contains and details the Dissemination, Communication, and Exploitation Plan of the HAAWAII project based on the Project Management Plan. It defines objectives, outlines the strategy, explains the process to declare high-level messages, identifies the target audience, proposes a schedule, gives an overview on concrete activities as well as their formats, and presents success criteria.	04/09/2020	PU
D6-5	Results of Dissemination, Communication and Exploitation. D6-4 is a living documented being updated during the lifetime of the HAAWAII project. D6-5 is the latest version of D6-4.	Part of exploitation phase, submission after D5-5	PU
D7-1	Project Management Plan	06/07/2020	CO
D7-1.010	The first progress report of the HAAWAII project described the achievements and the status of the first seven months of the project. WP1 was closed, the operational concept and requirements were available. Nevertheless, they were detailed during the project resulting in two public documents being available also for other Speech Recognition Projects, e.g. PJ-05-W2-Solution-97 or PJ-10-W2-Solution-96-ASR. Data recordings from NATS and Isavia ANS were available. The legal issues and data privacy concerns were solved. Very first speech recognition models even for the pilot's voice were available supporting the ANSPs during transcription.	22/12/2020	CO
D7-1.020	The second progress report of the HAAWAII project described the achievements and the status of 2021. WP1, WP2 and WP8 were closed now, i.e. the operational concept and requirements are available. Speech Recognition models have been developed for	20/02/2022	CO

	Iceland enroute airspace and for London TMA for both ATCo and pilot utterances. Command prediction predicted the command types and values for the 10 most frequently used command types. A REDA (Readback Error Detection Assistant) was implemented as offline analysis tool.		
FPR1	The first periodic report summarized the project achievements in the first reporting period and presented the plans for the second one.	27/07/2021	CO
D8-1	This document addresses the ethical principles, research integrity or relevant legislation of activities performed during the HAAWAII project and is part of the ethics deliverables documents.	04/09/2020	CO
D8-2	This document outlines requirements specifically aimed at the protection of personal data used in the HAAWAII project. It explains how these requirements are addressed and compliance is achieved. Particular attention is given to the storage, processing, access and sharing of data utilised throughout the project life cycle.	09/12/2020	CO

Table 1: Project Deliverables

3 Links to SESAR Programme

3.1 Contribution to the ATM Master Plan

The project started with the following Operational Improvement Steps (ON) and Enabler (EN).

Code	Name	Project contribution	Maturity at project start	Maturity at project end
ER APP ATC 180	*ER APP ATC 180 Controller productivity enhancements by Automatic Speech Recognition at the ER/APP CWP/HMI)*	*HAAWAII suggested to split into part A, B and C*	*V-level / TRL*	*V-level / TRL*

Table 2: Project Maturity

The current EN is too broad. HAAWAII partners agreed with SJU in October 2021 to split into parts A, B and C:

- ER APP ATC 180-A: Prefilling of radar labels by Automatic Speech Recognition at the ER/APP CWP/HMI.
- ER APP ATC 180-B: Prefilling of radar labels by Automatic Speech Recognition at the ER/APP CWP/HMI with AI-enhanced speech recognizer
- ER APP ATC 180-C: Automatic Speech Recognition at the ER CWP/HMI in support of CPLDC

Reason/background

HAAWAII and before MALORCA have shown that significant ASR performance improvements are possible, if the ASR engines are adapted to local environment constraints (e.g. waypoints, accents, local phraseology deviations). These deviations can be automatically learned/trained from recorded surveillance and voice data. HAAWAII even addresses the fact to train from massive amounts of data. Further controller productivity enhancements could be reached by using ASR as an additional input modality (alternative to mouse, keyboard or multi-touch input devices) to create CPDLC messages:

ER APP ATC 180-A was not addressed by HAAWAII. HAAWAII starts with the more challenging EN ER APP ATC 180-B.

Co-funded by the European Union

Code	Name	Project contribution	Maturity at project start	Maturity at project end
ER APP ATC 180-B	***Prefilling of radar labels*** *by Automatic Speech Recognition at the ER/APP CWP/HMI with* ***AI-enhanced speech recognizer***	*See D5-3*	*V-level / TRL*	*V-level / TRL*
ER APP ATC 180-C	*Automatic Speech Recognition at the* ***ER*** *CWP/HMI in support of* ***CPLDC***	*See D5-3*		

Table 3: Project Maturity for Pre-Filling Information by Automatic Speech Recognition and Understanding

The HAAWAII project also suggested the new OI step to SJU. Controllers are supported by Automatic Speech Recognition and Understanding to detect pilot readback errors as an additional safety net

Code[4]	Name	Project contribution	Maturity at project start	Maturity at project end
CM-XXXX	*Increased safety through pilot readback error detection on the ground*	*Summarize in one paragraph (~100 words)*	*TRL 0*	*TRL 1*

Table 4: New OI-Step to increase ATM safety by a REDA

As long as voice is used as main or at least as one communication means between ATCos and pilots, misunderstanding between both is possible due to e.g. different mother languages, bad communication channels. Readback errors or missing readbacks of pilots can result in severe incidents or even accidents if undetected. Automatic Speech Recognition and Understanding can support ATCOs in detecting readback errors and reduce the number of uncorrected readback errors (so called hearback errors), which results in workload reduction, but more important, it is an additional safety net, which increases ATM safety

Due to the seldom events of readback errors (<2% of the communication), the usage of ASR integrated into a REDA (Readback Error Detection Assistant) requires very low command recognition error rates (<1%), to keep false alarm rates in acceptable ranges (<20%).

---

[4] The code is not specified, yet (status 2022-07-18).

The HAAWAII project also suggested the new OI Step to SJU, which could reduce costs of ATCo training on the job by using Automatic Speech Recognition and Understanding and a pilot-readback model are used to automated controller training.

Code[5]	Name	Project contribution	Maturity at project start	Maturity at project end
CM-YYYY	*Increased ATCO productivity through ASR-enhanced ATCO training*	*Suggestions of new OI Step*	*HAAWAII has suggested this new OI Step, but does not explicitly work on this OI Step.*	

Table 5: New OI-Step to decrease of ATCo training on the job

Training costs are reduced as fewer resources of ATCo On-the-Job-Training-Instructors (OJTI) and simulation pilots are needed. Increased flexibility of training enables unscheduled training sessions when ATCos are not needed in the OPS room. More flexible training with reduced costs enables controllers to maintain licences for more sectors and, therefore, a more flexible assignment of controllers.

Cheaper training also enables more flexibility for the ATCos, so that e.g. a Vienna controller can work in Prague or at least in Linz or Graz

The HAAWAII project also suggested the new enabler to SJU, which suggests the integration of xMAN (and Conflict Detection) with ASR. Spoken controller commands are used to update system planning or to check if previously entered commands are not implemented by the pilot.

Code[6]	Name	Project contribution	Maturity at project start	Maturity at project end
ER APP ATC ZZZZ	*Ground system input via ASR at the ER/APP CWP/HMI*	*Suggestions of new OI Step*	*HAAWAII has suggested this new OI Step, but does not explicitly work on this OI Step.*	

Table 6: New Enabler to add to new sensor for xMANs

Knowing, what the ATCo has communicated to the pilot can update system planning better and earlier. In case the ATCo made a manual digital input in the system before, the consistency with the spoken command can be checked. While the freezing approaches of the extended AMAN and AMAN philosophy do not foresee plan adaptations, it should be up to the human operator to decide, what (s)he prefers in the current situation. Plan freezing and plan adaptation have both its advantages and disadvantages. Plan adaptions, however, requiring improved input channels in addition to surveillance data for immediate plan adaptations. ASR and Understanding can provide this additional input channel.

---

[5] The code is not specified, yet (status 2022-07-18).

[6] The code of this new enabler is not specified, yet (status 2022-07-18).

In addition to suggesting and working on new Enabler and OI enablers, SJU and the HAAWAII project also suggested to offer ASR and Understanding as a service. It could be delivered by an ATM Data Service Providers (ADSP).

The future European ATM Architecture will be based on common EU-wide ATM Data Services. Automatic Speech Recognition and Understanding could be one of these services, which offers to extract commands and pilot readbacks based on communication and surveillance data provided as input.

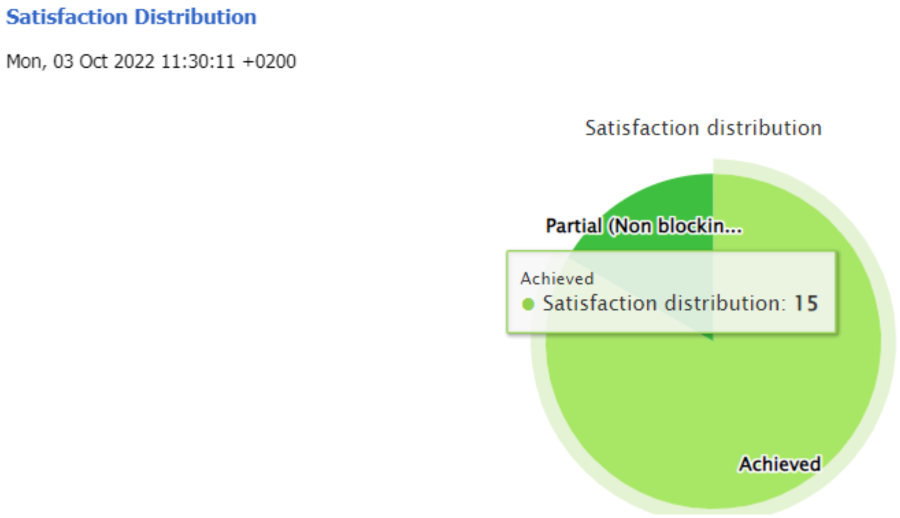
The project started with the following Operational Improvement Steps (ON) and Enabler (EN).

Code	Name	Project contribution	Maturity at project start	Maturity at project end
ER APP ATC 180	*ER APP ATC 180 Controller productivity enhancements by Automatic Speech Recognition at the ER/APP CWP/HMI)*	*Summarize in one paragraph (~100 words)*	*V-level / TRL*	*V-level / TRL*

Table 7: Project Maturity

3.2 Maturity Assessment

HAAWAII fully achieved TRL 2 and partly TRL3/4. On the STELLAR portal the necessary TRL 2 maturity assessment was done. The following figures show the results of the STELLAR tools (after having changed PER.ER.2 and PER.ER.3 from “Not Applicable” to “Partial Non-Blocking”.



Satisfaction Distribution

Mon, 03 Oct 2022 11:30:11 +0200

Figure 1 Satisfaction Distribution of TRL-2 Maturity Assessment performed on STELLAR

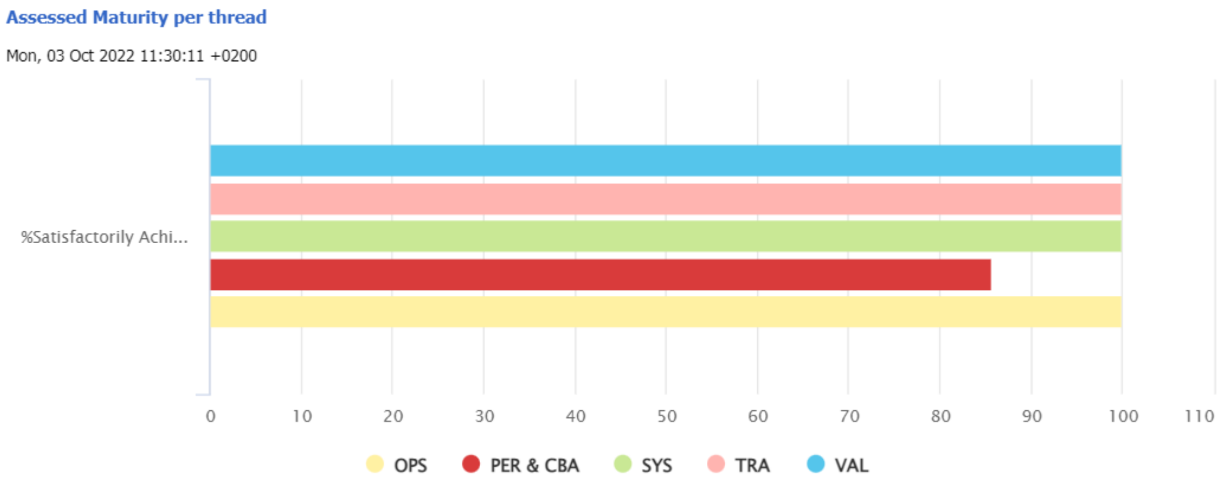


Figure 2 Assessed Maturity per Thread of TRL-2 Maturity Assessment performed on STELLAR

Table 8: ER Fund / AO Research Maturity Assessment

ID	Criteria	Satisfaction	Rationale - Link to deliverables - Comments
OPS.ER.6	Are there potential (sub)operating environments identified where, if deployed, the concept would bring performance benefits?	Achieved	The valuable feedback from industry and air navigation service providers motivates to exploit potential performance benefits of ASRU in a number of use cases such as pseudo-pilot support or replacement, callsign highlighting, radar label pre-filling, etc., see D5-2, D5-3 Additional Enablers and OI Steps are communicated to SJU and also summarized in D5-5, e.g. combination of xMAN and ASRU.
OPS.ER.1	Has a potential new idea or concept been identified that employs a new scientific fact/principle?	Achieved	The concept (and implementation) of an automatic readback error detection assistant (REDA) has been implemented in HAAWAII. The functionality uses a comparison of the semantics (annotation with ontology) of controller-pilot conversation to spot possible readback errors next to an additional data-based approach (D5-1, D5-2) Furthermore, the ontology has been enhanced to fit for further ATC communication content, e.g., pilot utterances, oceanic clearances, greetings/farewells, and a more complex values/conditions (D3-1) The usage of Automatic Speech Recognition and Understanding (ASRU) for ATCo workload prediction (for the future) is also new (D5-4)
OPS.ER.2	Have the basic scientific principles underpinning the idea/concept been identified?	Achieved	The three scientific partners of HAAWAII have described the scientific principles, e.g., ontology (Excel table shared with other European ASR projects such as Sol96, Sol97, STARFiSH), described the mechanisms and results in various deliverables (e.g., D3.x), and

			published the results in scientific papers at international conferences.
OPS.ER.3	Does the analysis of the "state of the art" show that the new concept / idea / technology fills a need?	Achieved	Yes, the state-of-the-art research shows that automatic readback error detection remains a huge challenge especially in the speech understanding (not speech recognition) process. One avoided aviation accident can already justify the effort to develop such a system even if we are just on the way to acceptable low false alarm rates (see D5-1, D5-2) The achieved word error rates (D3-4, D3-5) outperform the results of Sol. 16-04 and MALORCA even for pilot utterances. The same applies for speech understanding i.e. text-to-concept, which is summarized in D5-2 and D5-3.
OPS.ER.4	Has the new concept or technology been described with sufficient detail? Does it describe a potentially useful new capability for the ATM system?	Achieved	The technology has been described and tested in simulations at Isavia air navigation service provider in Iceland. Great potential of the ASRU technology for a possible use in the operations room of an ANSP have been confirmed and was demonstrated in Isavia ops room 29-06-2022, see D5-2. Speech-to-Text is detailed in D3-4 and D3-5. Text-to-Concept is detailed in D5-2 and D5-3 and in papers, see reference list in D5-5. Workload prediction is detailed in D4-1 and D5-4. REDA is detailed in D5-1 and D5-2.
OPS.ER.5	Are the relevant stakeholders and their expectations identified?	Achieved	The relevant stakeholders were clearly identified in D6.4. Furthermore, an overwhelming number of different stakeholders from different continents took part in HAAWAII's stakeholder workshops and demonstration days. The stakeholder feedback showed that their

			expectations are fully satisfied for the current TRL. This will again be addressed in D6-1.
PER.ER.1	Has a feasibility study been performed to confirm the potential feasibility and usefulness of the new concept / idea / Technology being identified?	Achieved	Isavia ANS performed proof-of-concepts trials with air traffic controllers in May 2022 using operations room data (surveillance and speech) to demonstrate the feasibility. The different elements of the ASRU applications show different levels of technology readiness, but definitely demonstrate their technological feasibility, see results documents in D5-2. NATS performed feedback studies with their ATCos in May, June 2022. The results are documented in D5-4.
PER.ER.2	Is there a documented analysis and description of the benefit and costs mechanisms and associated Influence Factors?	Partial (Non-blocking)	A detailed cost-benefit analysis has not been performed in the early TRL phase of the HAAWAII project. Nevertheless, Isavia ANS wants to continue with ASRU in SESAR-3, which already shows that there is more than a feeling that the CBA is positive.
PER.ER.3	Has an initial cost / benefit assessment been produced?	Partial (Non-blocking)	A detailed cost-benefit analysis has not been performed in the early TRL phase of the HAAWAII project. Nevertheless, Isavia ANS wants to continue with ASRU in SESAR-3, which already shows that there is more than a feeling that the CBA is positive.
PER.ER.4	Have the conceptual safety benefits and risks been identified?	Achieved	The potential safety benefits are hints for controllers regarding potential readback errors of pilots and (as experience with ASRU in ATC shows) controllers to stick closer to phraseology due to the ABSR support. However, the safety effect (e.g., also from higher false alarm rates) has not been systematically analysed at the current TRL.

PER.ER.5	Have the conceptual security risks and benefits been identified?	Partial (Non-blocking)	The ASRU component is a software module of the controller working position (CWP). Thus, the security aspects of the CWP itself are relevant as the ASRU component is "behind" the CWP security wall. These results are not explicitly reported in a deliverable. We refer here to the results of Sol.96 of SESAR2-IR. Therefore, we decided not to fully achieve and analyse the security risks and benefits.
PER.ER.6	Have the conceptual environmental impacts been identified?	Achieved	The environmental impacts can very roughly be extrapolated/anticipated from the AcListant®-Strips project results where saving of more than 50 litters kerosene per flights could be shown, when ASRU is used to support ATCos for radar label pre-filling. However, it would not be scientifically correct to do a one-to-one mapping on Icelandic en-route and UK approach airspace with different circumstances. The criteria is achieved, because ASRU performance even on ops-room data has been increased with respect to AcListant®-Strips project. These criteria, however, has not been explicitly addressed by HAAWAII.
PER.ER.7	Have the conceptual Human Performance aspects been identified?	Achieved	The ASRU functionality is also used (based on the selected data set) to estimate controller workload for better rostering. Here HAAWAII again relies on the AcListant(r)-Strips projects and just extrapolates our results which worked with ops room data. The REDA aspects of HMI have been addressed during the first Stakeholder Workshop (June 2021, see D6-1) and in D5-2. There are readback errors according to the book, some of them are brought to the ATCos' attention and even less are communicated to the pilot. The challenge is addressed, and the final report (D5-5) clearly addresses the need for further research.

SYS.ER.1	Has the potential impact of the concept/idea on the target architecture been identified and described?	Achieved	EATMA already includes a functional block and functions for speech recognition and understanding. The additional functionalities developed in HAAWAII are just some further instances of e.g., an application that uses the ASRU output.
SYS.ER.2	Have automation needs e.g. tools required to support the concept/idea been identified and described?	Achieved	The ABSR and ASRU functionality itself is a tool that offers support for a controller. However, further human factors analysis should be applied when moving to higher TRLs on how to present the ABSR output to the controller. REDA is a new functionality of the CWP (D5-1, D5-2) CoCoLoToCoCo and SpokenData (see D3-1) support to transcribe and annotate the data acquisition process also for annotation of readback error use cases. D5-4 addresses the need for ASRU output analysis with respect to ATCO workload prediction and post-analysis.
SYS.ER.3	Have initial functional requirements been documented?	Achieved	The functional requirements of an ASRU component have been documented during PJ.16-04 as well as Solution 96 and 97 projects. Nevertheless, HAAWAII project created a separate Operation Concept Document D1-1 that was reviewed by IFATCA with SJU support. The feedback is used in the Final Operational Concept Document D6-2. D1-1 was used as Input for a HAAWAII requirements document D1-2, which was improved/updated during the lifetime of the project. The final version is D6-3.
TRA.ER.1	Are there recommendations proposed for completing V1 (TRL-2)?	Achieved	The Final Project Results Report lists the most important recommendations to move beyond TRL2 and to complete minor aspects of reaching TRL2. It also addresses new enablers (EN) and OI steps resulting from the HAAWAII project results. These are already communicated to SJU in early Nov. 2021.

VAL.ER.1	Are the relevant R&D needs identified and documented? Note: R&D needs state major questions and open issues to be addressed during the development, verification and validation of a SESAR Solution. They justify the need to continue research on a given SESAR Solution once Exploratory Research activities have been completed, and the definition of validation exercises and validation objectives in following maturity phases.	Achieved	The R&D needs are mainly documented in the conclusions and recommendations of the Final Project Results Report. This includes human-in-the-loop simulations for REDA on higher TRL, developing the functionalities closer to industrial needs, and analysing non-nominal situations. Automatic learning of Speech Understanding (from sequence of words to sequence of ATC concepts) from given transcription/annotation pairs is a hot research topic, which would enable easier adaptation to new environments and ease the maintenance when ATCO phraseology evolves.

4 Conclusion and Lessons Learned

4.1 Conclusions

HAAWAII project (running over the last two years) aimed to advance an automatic support for end-route TMA. Specifically, we explored new ways of digitalisation of voice communication between ATCos and pilots by using machine-learning based automatic speech recognition solutions. The application of the technology was aligned with following areas: (a) pilot readback error detection, (b) modelling the controller behaviour, (c) pre-filling radar labels and CPDLS messaging and (d) human performance metric extraction.

More particularly, six objectives were related to the abovementioned applications, all related to researching, developing and integration the automatic speech recognition engine (and other important blocks) to the level of accuracy required for its applicability. Most of these objectives were fulfilled, although due to Corona restriction less voice data was generated and recorded than expected. As related to objective 2, word error rates of 3% for ATCos in UK and Icelandic airspace and below 7% for pilots in UK airspace and below 11% for pilots in Icelandic airspace were achieved. The lower word recognition performance on Isavia data has two reasons: 10% of the aircraft were not in the provided surveillance data for Isavia airspace and maybe more important, Isavia covers a much bigger airspace than the included sectors at NATS, so that the voice receivers are sometimes far away from the aircraft calling the ATCo. This causes the decrease of the quality of the signal (i.e., lower signal to noise ratio).

Lower performance on word level of course also results in lower performance on semantic level, i.e. in the command recognition rates and callsign recognition rates. On callsign level we achieved for Isavia data 88% for pilot data and for NATS 95%, which corresponds to the 71% and 82% on command level. With respect to objective 2 and as mentioned above, both pilot and ATCos were analysed and used for recognition. More surprising is the performance analysis on ATCo data: For NATS, a performance of 98.5% for callsign recognition for ATCo data is achieved, whereas for Isavia we get 96.9% (which is significantly better than planned by a relevant KPI). The same applies when considering command level: 94% for NATS compared to 92% for Isavia. The difference might be surprising, because the results on word level are better for Isavia than for NATS, but on semantic level the NATS' results are better. The reason, however, is again that some aircraft are not visible in the provided surveillance data, i.e. if the callsign is not fully spoken or is not fully recognized, the callsign extractor has no chance to extract the correct callsign. It is not possible to detect AFR2A25 from the wrong word sequence "air france to alfa three five", when it is not known that the AFR2A25 is in the air.

As part of the objective 3 two approaches for read back error detection have been researched and implemented: (i) ontology based- and (ii) data-driven approach. However, the data driven approach suffers from very few training samples. Only 90 readback error use cases were extracted from the four hours of annotated Isavia data, of which only one third were real readback errors, the rest were mostly missing readbacks of the pilot. Nevertheless, the ontology-based or the rule-based approach discovers 4 out of 5 readback error use cases. The achieved false alarm rate of 70% is far beyond the desired 10% (i.e., compare to objective 3 and related KPIs), although the false alarm rate could be slightly reduced by combining with the ontology-based approach. The results clearly show the potential of a readback error detection assistant (REDA), which is achievable even with the current implementation.

Validation trials first in lab environment, but very soon in ops room are recommended to be organised with interesting ANSPs to figure out whether a false alarm of 10% is really required or if a false alarm rate of 50% is also acceptable for read-back error detection. This seems to be achievable, when more voice recordings from pilots are included into the training data sets and when the speech quality from the voice receivers is included. Possible readback alerts resulting from bad voice quality should be shown to the ATCo.

Due to the scope of the project NATS data were mainly used for workload prediction, whereas Isavia ANS has concentrated on REDA. It is recommended that Isavia data is also analysed for predicting ATCos workload, but more important, NATS data should be analysed with respect to possible readback error use cases. Better results can be expected here, because both ATCo and pilot recognition rates on semantic level are better than for Isavia's airspace.

Using speech recognition and understanding for radar label maintenance and integration with CPDLC was again successfully proven (see objective 4 and related KPIs). The command recognition rates even on voice recordings from the ops room achieved 94% for NATS and 86% for Isavia, when we concentrate only on the commands relevant for CPDLC and radar label maintenance. This is much better than the achieved 70% for Prague approach or 55% for Vienna approach, respectively. Both numbers result from the lab environment trials from PJ16-04.

4.2 Technical Lessons Learned

Validation trials need to consider whether the PTT (Push-to-Talk) signal for ATCo utterances is available or not. At the same time, a clear separation of ATCos and pilots (physically in different channels) is of high importance as it is impossible to run the speaker detection (i.e., segmentation) in real-time (i.e., the simplest solution is to run two parallel ASR engines to mitigate the algorithmic delay for real-time streaming which increases the complexity of the whole systems significantly) The online applications of Callsign Highlighting, Pre-Filling Radar Labels and CPDLC message and REDA could expect that a PTT is available. Therefore, start and end of an utterance are known and also no speaker detection is necessary. Offline evaluation of ATC workload, however, does not rely on the availability of the PTT . This application employs a speaker detection and a voice activity detection algorithm. The same applies for callsign highlighting on initial call of pilots or pilot readbacks. With PTT, it is known if ATCo or pilot speaks and when ATCo starts and stops speaking. It is, however, not known, when the pilot starts and stops speaking if PTT of the pilot is not transmitted from the aircraft to the ground station.

PTT signal detection should rely on ED-137 standard, which means that a full integration of the speech recognition software into the ops room environment also reduces some needed steps in the offline validations.

Flight information should be included into the surveillance data to avoid the AFR2A25 problem mentioned above. HAAWAII results showed, that recognition rate on command level improves by 10% absolute, when surveillance data is combined with speech data.

4.3 Plan for next R&D phase (Next steps)

The Readback Error Detection Assistant (REDA) needs to be evaluated in human-in-the-loop simulations.

> Stakeholder Workshops and discussions with ATCos have clearly shown, that everybody agrees (in principle) what a readback error is, if the phraseology would have been used according to the book. Agreements, however, are missing when discussing, which (possible) readback error use cases should be brought to the ATCo's attention. And even more questions were open

when trying to get agreements on readback error use cases, which should be communicated to the pilot.

The HAAWAII project could only evaluate the quantitative numbers of Readback Error Detection Rates and False Alarm Rates. It was shown, that automatic understanding of ATCo-pilot-communication in both directions are feasible with sufficient accuracy, which make the cases with detection of readback error of interest for large community even though still a high number of false alarms is obtained. It is currently not clear, if the achieved detection rate together with the false alarm rate increases safety or decreases it. The simulation can answer the question, which readback error use cases should be brought to the ATCo's attention and which one should be communicated to the pilot. Is also gives hints for the needed user interface for this task.

The Readback Error Detection Assistant (REDA) needs to be evaluated on ops-room data.

Evaluation of the REDA in human-in-the-loop simulations can only be a very first step. Readback error use cases are luckily still seldom events. Therefore, very long simulation runs would be necessary. In any case the ATCo will be more focusing on readback error use cases in the lab environment. Shadow mode trials or post-evaluations and analyses on ops-room data could be first step, before evaluation of the REDA in the ops room starts.

Standardization of ontology for transcription and annotation

- The standardization avoids that other ASR related projects "re-invent the wheel".
- The standardization enables the technical (not the legal) exchange of voice utterance transcriptions and annotations.
- The standardization enables the exchange of building blocks (e.g. speech-to-text, text-to-concept, command prediction) from different suppliers.
- The standardization enables to compare the performance on commonly available test data sets.

Standardization of logical exchange format between building blocks

The standard should define

- Annotation ontology
- Transcription ontology
- More than one speaker in one utterance (under-splitting of controller-pilot communication)
- An utterance is split into multiple parts (over-splitting of an utterance)
- N-Best-Lists of speech-to-text output, i.e., the output of STT is not always unique (e.g., "descend two seven zero" or "descend to seven zero")

- N-Best-List of text-to-concept output ("one seven zero" could be the part of a callsign (e.g., DLH170) or the readback of a flight level, speed value or heading (e.g., ALTITUDE 170 none, SPEED 170 none, HEADING 170).
- N-Best-List for speaker classification
- Plausibility values for STT and TTC output

All these items could also be combined.

HAAWAII proposes a solution for the above challenges and tested it already in exercise 002 of Solution 96, exercise 006 of Solution 97, and the German STARFiSH project.

A first step is initiated. ASR is added to the technical Work Program of EUROCAE. This could result in a new working group starting the standardization activities in 2023. Automatic Speech Recognition and Understanding will also be of importance with SESAR-3 starting mid of 2023. These solutions need to coordinate or better to work together with this new working group.

Automatic learning of TTC transformation rules

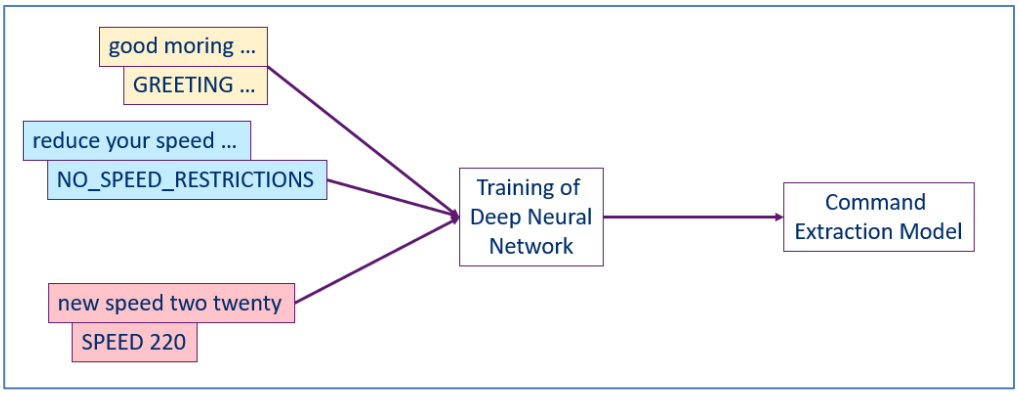


Figure 3 Automatic Learning of Command Extraction Rules

Extension of the enablers as described in 3.1

Splitting ER APP ATC 180 into:

- ER APP ATC 180-A: Prefilling of radar labels by Automatic Speech Recognition at the ER/APP CWP/HMI.
- ER APP ATC 180-B: Prefilling of radar labels by Automatic Speech Recognition at the ER/APP CWP/HMI with AI-enhanced speech recognizer
- ER APP ATC 180-C: Automatic Speech Recognition at the ER CWP/HMI in support of CPLDC

Create new Enabler: **ASR-based workload assessment in real time and workload prediction, which** needs to be plugged into an Operational improvement. This could be CM-0103-B, which is under development by Sol 44.

Create new Enabler “**Ground system input via ASR** at the ER/APP CWP/HMI”, which uses ASR and Speech Understanding as an additional sensor for e.g. xMAN and conflict detection.

Extension of the Operational Improvement step (OI Steps) as described in 3.1

Creating a new OI Step “**Increased safety through pilot readback error detection on the ground**”.

Creating a new OI Step “**Increased ATCO productivity through ASR-enhanced ATCO training**”

The ATCO workload prediction application needs to be validated in larger scale studies

The results of the proof-of-concept evaluation exercises are promising in terms of being able to show changes in subjective workload using objective speech recognition data. However, further work is needed to identify cut-off values that allow the parameters, based on speech recognition output, to be used for operational decision making (e.g., when to reconfigure sectors). Additionally, work is needed to develop a human machine interface (HMI) that is fit for purpose and presents the data to the supervisory staff in a meaningful way. Larger scale validation studies will also help to identify and further refine which parameters are best suited for workload prediction. Evaluation results showed for example that speech ratio and number of transmissions per minute provide better insights than words spoken per minute. This work could result in a new enabler **ASR-based workload assessment in real time and workload prediction** or an integration with solution 44**.**

The future European ATM Architecture will be based on common EU-wide ATM Data Services. **Automatic Speech Recognition and Understanding** could be one of these services offering speech to text and semantic interpretation **as a service**.

5 References

5.1 Project Deliverables

[1] ARILÍUSSON, T. SIMIGANOSCHI, H. HELMKE, J. HARFMANN: HAAWAII project: D1.1: Operational Concept Document, version 01.00.00, 19. August 2020.

[2] T. SIMIGANOSCHI, H. HELMKE, P. SMRZ, P. MOTLICEK, H. PÁLSSON: HAAWAII project: D1.2: System Requirements Document, version 01.00.00, 12. October 2020.

[3] Matthias KLEINERT, Hartmut HELMKE, Petr MOTLICEK: HAAWAII project: D1-3 Architecture Design Document, version 01.00.00, 05. November 2020.

[4] Matthias KLEINERT, Hartmut HELMKE: HAAWAII project: D2-1 Surveillance, voice and associated meta data for Vienna approach, version 01.00.00, 03 August 2020.

[5] H. Helmke; A. Prasad, T. S. Simiganoschi, J. Harfmann: HAAWAII project: D2.2 One Month of Surveillance and Voice Data from NATS, version 1.00, 23rd February 2021.

[6] H. Helmke; A. Prasad, T. S. Simiganoschi, J. Harfmann: HAAWAII project: D2.3 One Month of Surveillance and Voice Data from Isavia, version 1.00, 4th March 2021.

[7] Petr MOTLICEK, Julia HARFMANN, Hartmut HELMKE, Teodor S. Simiganoschi: HAAWAII project: D2-4 Data Management Plan, version 01.00.00, 25 January 2021.

[8] H. HELMKE, Shruthi SHETTY, Hartmut HELMKE, Oliver OHNEISER, Frantisek GREZL et al.: HAAWAII project: D3.1: Transcription and Annotation Handbook - CoCoLoToCoCo, version 1.00, 5th November 2020.

[9] Shruthi SHETTY, Hartmut HELMKE: HAAWAII project: D3-2 Four Hours of Silence Reduced Annotated Controller and Pilot Utterances for UK, version 01.00.00, 10 June 2021.

[10] Hartmut HELMKE, Shruthi SHETTY: HAAWAII project: D3-3 Four Hours of Silence Reduced Annotated Controller and Pilot Utterances from Iceland, version 01.00.00, 25 May 2021.

[11] Amrutha PRASAD, Shruthi SHETTY, Frantisek GREZL, Pavel SMRZ, Karel ONDREJ, Bolaji YUSUF, Petr MOTLICEK, Hartmut HELMKE: HAAWAII project: D3-4 Final Acoustic, Language and Command Prediction Models for ATCo' Speech, version 01.00.00, 31 January 2022.

[12] Amrutha PRASAD, Shruthi SHETTY, Frantisek GREZL, Petr MOTLICEK, Hartmut HELMKE: HAAWAII project: D3-5 Final Acoustic, Language and Com-mand Prediction Models for Pilots' Speech, version 01.00.00, 13 May 2022.

[13] Julia HARFMANN, Teodor SIMIGANOSCHI, Hartmut HELMKE: HAAWAII project: D4-1 Human Performance Metrics Evaluation Concept, version 01.00.00, 24 September 2021.

[14] Hartmut HELMKE, Teodor S. SIMIGANOSCHI, Shruthi SHETTY, Julia HARFMANN: HAAWAII project: D5-1 Ontology Based Readback Error Detection, version 01.00.00, 18 November 2021

[15] Teodor SIMIGANOSCHI, Hartmut HELMKE, Karel ONDREJ, Shruthi SHETTY, Petr MOTLICEK, Hjalti PALSSON, Hörður ARILIUSSON: HAAWAII project: D5-2 Data Driven Readback Error Detection Evaluation Report, version 01.00.00, 06 July 2022

[16] Hartmut HELMKE, Shruthi SHETTY, Karel ONDREJ, Pavel SMRZ, Teodor S. SIMIGANOSCHI: HAAWAII project: D5-3 Real Time Capability and Pre-filling Evaluation Report, version 01.00.00, 22 June 2022

[17] Julia HARFMANN: HAAWAII project: D5-4: Human Performance Metrics Evaluation , version 01.00.00, August 2022 in preparation

[18] Hartmut HELMKE, Oliver OHNEISER, Pavel SMRZ: HAAWAII project: D5-5: Final Project Results Report , version 01.00.00, August 2022 this document

[19] Hartmut HELMKE, Christian WINDISCH: HAAWAII project: D6.1 Report on Stakeholder Workshop Results, version 00.02.00, August 2021, final version expected November 2022

[20] Hörður ARILÍUSSON, Teodor SIMIGANOSCHI, Hartmut HELMKE, Julia HARFMANN: HAAWAII project: D6-2 Updated Operational Concept Document, version 01.50.00, May 2021, final version expected October 2022

[21] Teodor SIMIGANOSCHI, Hartmut HELMKE, Pavel SMRZ, Petr MOTLICEK, Hjalti PÁLSSON:: HAAWAII project: D6-3 Updated Requirements Document , version 01.05.00, July 2021, final version expected October 2022

[22] Oliver OHNEISER, Teodor S. SIMIGANOSCHI, Christian WINDISCH, Jules HARFMANN, Pavel SMRZ, Petr MOTLICEK: HAAWAII project: D6.4 Dissemination, Communication, and Exploitation Plan, version 2.03, 17th May 2022.

[23] Oliver OHNEISER, Teodor S. SIMIGANOSCHI, Christian WINDISCH, Jules HARFMANN, Pavel SMRZ, Petr MOTLICEK: HAAWAII project: D6.5 Results of Dissemination, Communication, and Exploitation, in preparation, final results expected in November 2022.

[24] Hartmut HELMKE: HAAWAII project: D7-1 Project Management Plan of HAAWAII Project, version 02.00.00, 5. November 2020

[25] Hartmut HELMKE, Petr MOTLICEK, Christian WINDISCH, Pavel SMRZ: HAAWAII project: D7-1.010 The first progress report of the HAAWAII project described the achievements and the status of the first seven months of the project, version 01.00.00, 21. December 2020.

[26] Hartmut HELMKE, Julia HARFMANN, Pavel SMRZ, Petr MOTLICEK: HAAWAII project: D7-1.020 The second progress report of the HAAWAII project described the achievements and the status of 2021 , version 01.00.00, 23. February 2022.

[27] Sine MYEZWA, Julia HARFMANN: HAAWAII project: D8-1 H-Requirements No. 1, version 01.00.00, 4. September 2020.

[28] Hartmut Helmke, Julia HARFMANN, Teodor S. SIMIGANOSCHI: HAAWAII project: D8-2 POPD-Requirements No. 2, version 01.00.00, 9. Dezember 2020.

5.2 Project Publications

[29] Helmke, Hartmut; Kleinert, Matthias; Shetty, Shruthi; Ohneiser, Oliver; Ehr, Heiko; Arilíusson, Hörður; Simiganoschi, Teodor S.; Prasad, Amrutha; Motlicek, Petr; Veselý, Karel; Ondřej, Karel; Smrz, Pavel; Harfmann, Julia; Windisch, Christian (2021) Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety. 14th ATM Seminar, 20.-24. Sep 2021, Virtual.

[30] Helmke, Hartmut; Shetty, Shruthi; Kleinert, Matthias; Ohneiser, Oliver; Prasad, Amrutha; Motlicek, Petr; Cerna, Aneta; Windisch, Christian (2021) Measuring Speech Recognition Understanding Measuring Speech Recognition Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates. 11th SESAR Innovation Days, 07.12.2021 - 09.12.2021, Virtual Conference.

[31] Martin Kocour, Karel Veselý, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, Jan Černock, Dietrich Klakow, Petr Motlicek: Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition, Special Session on "Automatic Speech Recognition in Air Traffic Management", 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, INTERSPEECH 2021, 30. Aug. - 03. Sep. 2021, Brno, Czechia (hybrid conference).

[32] Kleinert, Matthias; Helmke, Hartmut; Shetty, Shruthi; Ohneiser, Oliver; Ehr, Heiko; Prasad, Amrutha; Motlicek, Petr; Harfmann, Julia (2021) Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In: 40th AIAA/IEEE Digital Avionics Systems Conference, DASC 2021. 40th Digital Avionics Systems Conference (DASC), 03.-07. Oct 2021, San Antonio, Texas, USA. ISSN 2155-7195.

[33] Kleinert, Matthias; Venkatarathinam, Narasimman; Helmke, Hartmut; Ohneiser, Oliver; Strake, Maximilian; Fingscheidt, Tim (2021) Easy Adaptation of Speech Recognition to Different Air Traffic Control Environments using the DeepSpeech Engine. 11th SESAR Innovation Days, 07.-09. Dec 2021, Virtual Conference.

[34] Iuliia Nigmatulina, Rudolf Braun, Juan Pablo Zuluaga, Petr Motlicek: Improving callsign recognition with air-surveillance data in air-traffic communication,Satelitte Workshop "Automatic Speech Recognition in Air Traffic Management (ASR-ATM)" at the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, INTERSPEECH 2021, 30. Aug. - 03. Sep. 2021, Brno, Czechia (hybrid conference)

[35] Ohneiser, Oliver; Helmke, Hartmut; Shetty, Shruthi; Kleinert, Matthias; Ehr, Heiko; Murauskas, Sarunas; Pagirys, Tomas (2021) Prediction and Extraction of Tower Controller Commands for Speech Recognition Applications. Journal of Air Transport Management, 95. Elsevier. doi: 10.1016/j.jairtraman.2021.102089. ISSN 0969-6997.

[36] Ohneiser, Oliver; Sarfjoo, Saeed; Helmke, Hartmut; Shetty, Shruthi; Motlicek, Petr; Kleinert, Matthias; Ehr, Heiko; Murauskas, Sarunas (2021) Robust Command Recognition for Lithuanian Air Traffic Control Tower Utterances. In: 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, 1, pp. 266-270. INTERSPEECH 2021, 30. Aug. - 03. Sep. 2021, Brno, Czechia (hybrid). doi: 10.21437/Interspeech.2021-935. ISBN 978-171383690-2. ISSN 2308-457X.

[37] Prasad, Amrutha; Zuluaga-Gomez, Juan Pablo; Motlicek, Petr; Ohneiser, Oliver; Helmke, Hartmut; Sarfjoo, Saeed; Nigmatulina, Iuliia (2021) Grammar Based Identification Of Speaker Role For Improving ATCO And Pilot ASR. Interspeech 2021 Satellite Workshop "Automatic Speech Recognition in Air Traffic Management (ASR-ATM)", 30. Aug 2021, Brno, Czechia (hybrid conference)

[38] Shetty, Shruthi; Helmke, Hartmut; Kleinert, Matthias; Ohneiser, Oliver (2022) Early Callsign Highlighting using Automatic Speech Recognition to Reduce Air Traffic Controller Workload. accepted for 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022), New York, USA, July 24-28, 2022.

[39] Igor Szöke, Santosh Kesiraju, Ondřej Novotný, Martin Kocour, Karel Veselý, Jan Černocký: Detecting English Speech in the Air Traffic Control Voice Communication, Special Session on “Automatic Speech Recognition in Air Traffic Management”, 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, INTERSPEECH 2021, 30. Aug. - 03. Sep. 2021, Brno, Czechia (hybrid conference).

[40] Juan Zuluaga-Gomez, Iuliia Nigmatulina, Amrutha Prasad, Petr Motlicek, Karel Veselý, Martin Kocour, Igor Szöke: Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems: Special Session on “Automatic Speech Recognition in Air Traffic Management”, 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, INTERSPEECH 2021, 30. Aug. - 03. Sep. 2021, Brno, Czechia (hybrid conference).

[41] Project website www.haawaii.de

5.3 Other

[42] Project Execution Guidelines for SESAR Exploratory Research, Edition 01.00.00, 08/02/2016

[43] Helmke, Hartmut; Ohneiser, Oliver; Mühlhausen, Thorsten; Wies, Matthias (2016) Reducing Controller Workload with Automatic Speech Recognition. In: DASC 2016 35th Digital Avionics Systems Conference, pp. 1-10. 35th Digital Avionics Systems Conference DASC 2016, 25.-29. Sep 2016, Sacramento, USA.

[44] S. Chen, H.D. Kopald, R. Chong, Y. Wei, and Z. Levonian: Read Back Error Detection using Automatic Speech Recognition. In 12th USA/ Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 2017

[45] Helmke, Hartmut; Ohneiser, Oliver; Buxbaum, Jörg; Kern, Christian (2017) Increasing ATM Efficiency with Assistant Based Speech Recognition. In: Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2017, Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), 27.-30. June 2017, Seattle, Washington, USA.

[46] Helmke, Hartmut; Slotty, Michael; Poiger, Michael; Ferrer Herrer, Damián; Ohneiser, Oliver; Vink, Nathan; Cerna, Aneta; et al., (2018) Ontology for Transcription of ATC Speech Commands of SESAR 2020 Solution PJ.16-04. 37th AIAA/IEEE Digital Avionics Systems Conference (DASC), 23.-27. Sep 2018, London, UK.

[47] Ohneiser, Oliver; Helmke, Hartmut; Kleinert, Matthias; Siol, Gerald; Ehr, Heiko; Hobein, Stephanie; Predescu, Andrei-Vlad; Bauer, Jakob (2019) Tower Controller Command Prediction for Future Speech Recognition Applications. 9th SESAR Innovation Days, 02.-05. Dec 2019, Athens, Greece.

[48] Kleinert, Matthias; Helmke, Hartmut; Moos, Sylvain; Hlousek, Petr; Windisch, Christian; Ohneiser, Oliver; Ehr, Heiko; Labreuil, Aline (2019) Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. SESAR Innovation Days 2019, 2.-5. Dec 2019, Athens, Greece.

[49] Rataj, Jürgen; Helmke, Hartmut; Ohneiser, Oliver (2021) AcListant with Continuous Learning: Speech Recognition in Air Traffic Control (EIWAC 2019). In: Springer Journal Springer Verlag. pp. 93-112. doi: 10.1007/978-981-33-4669-7. ISBN 1876-1100. ISSN 1876-1119.

[50] Helmke, Hartmut; Kleinert, Matthias; Ohneiser, Oliver; Ehr, Heiko; Shetty, Shruthi (2020) Machine Learning of Air Traffic Controller Command Extraction Models for Speech Recognition Applications. In: 39th AIAA/IEEE Digital Avionics Systems Conference, DASC 2020. 39th Digital Avionics Systems Conference DASC, 11.-16. Oct 2020, Virtual Conference. ISBN 978-172819825-5. ISSN 2155-7195.

[51] HAAWAII project: Grant Agreement, 24. Jan 2020.

Appendix A

A.1 Glossary of terms

Term	Definition	Source of the definition
Accuracy	Accuracy represents the prediction rate. It also considers the number of commands which were correctly NOT predicted. (tp + tn) / (tp + fp + fn +tn)	
AcListant®	Venture Capital funded project Active Listening Assistant being conducted by DLR and Saarland University from 2013 to 2015.	PJ.16-04
Annotation	This task extracts the semantic concepts from the Transcription (i.e. text-to-concepts transformation), e.g., “DLH2BA DESCEND 80 FL, DLH2BA REDUCE 220 kt” and “AFR273 CORRECTION, AFR273 CONTACT VIENNA_RADAR, AFR273 CONTACT_FREQUENCY 129.500”.	D3.1
Assistant Based Speech Recognition (ABSR)	Special Instance of Automatic Speech Recognition which needs an assistant system to provide context in order to improve recognition rate and/or reduce error rate	See definition in [52]
Automatic Speech Recognition	An Automatic Speech Recognition (ASR) system gets an audio signal as input and transforms it into a sequence of words, i.e. “speech-to-text” following the recognition process. The sequence of words is transcribed into a sequence of ATC concepts (“text-to-concepts”) using an ontology. The word sequence “lufthansa two alpha altitude four thousand feet on qnh one zero one four reduce one eight zero knots or less turn left heading two six zero” is transcribed into “DLH2A ALTITUDE 4000 ft, DLH2A INFORMATION QNH 1014, DLH2A REDUCE 180 OR_LESS, DLH2A HEADING 260 LEFT”. The resulting concepts can be used for further applications such as visualization on an HMI.	PJ.16-04
Callsign (Recognition) Error Rate	The number of callsign, which are wrongly recognized by ABSR and which are not rejected divided by the number of total given callsigns; in other words: the percentage of given callsigns wrongly shown on the controllers’ HMI. “oscar kilo one” must be mapped to “OACK1” if this is the only “O..K1” in the air. Otherwise it is counted as an error.	in D1.2

Term	Definition	Source of the definition
	We are not counting as an error when the plausibility for the extracted callsign is below a given threshold (e.g. 90%). In this case we are counting as a rejection.	
Callsign Recognition Rate	The number of callsigns, which are correctly recognized by ABSR and are not rejected before divided by the number of total given callsigns; in other words: the percentage of given callsigns correctly shown on the controllers' HMI. "oscar kilo one" must be mapped to "OACK1" if this is the only "O..K1" in the air. If no callsign is extracted from the utterance and also no callsign is provide by the ATCo or pilot, this is counted as a recognition.	in D1.2
Callsign Rejection Rate	The number of callsigns, which are said by the ATCo, but mapped to NO_CALLSIGN divided by the number of total given callsigns; in other words: the percentage of given callsigns not shown at all on the controllers' HMI. We also count as a rejection when the plausibility for the extracted callsign is below a given threshold (e.g. 90%). If no callsign is extracted from the utterance and also no callsign is provide by the ATCo or pilot, this is counted as a recognition.	in D1.2
Chunk		D3.1
Clearance transmission identifier	The Clearance transmission identifier is part of the readback information and represents the Transmission unique identifier from the Transmission information. This will be used to trace and check a specific transmission from the multiple transmissions. See example in Table 10 Example of transmission information and identifiers	in D1.2
CoCoLoToCoCo	Controller Command Logging Tool for Context Comparison that provides a user-friendly interface to carry out transcriptions and various annotations for air traffic control voice commands.	D3.1
Command Prediction Error Rate	The number of controller commands which are given but not predicted (by the Command Hypotheses Predictor) divided by number of total given commands; in other words: the percentage of errors of the Command Hypotheses Predictor.	See definition in [52]
Command Recognition Rate	The number of controller commands which are correctly recognized by ASR and are not rejected before divided by number of total given commands; in other words: the percentage of given commands correctly shown on the controllers' HMI.	See definition in [52]

Term	Definition	Source of the definition
Command (Recognition) Error Rate	The number of controller commands which are wrongly recognized by ASR and which are not rejected divided by number of total given commands; in other words: the percentage of given commands wrongly shown on the controllers' HMI.	See definition in [52]
Communication group	Communication group is part of transmission information and it is a generated value or index that is used to identify and group multiple ATCO/Pilot transmissions that represent a single communication/dialogue. The single communication/dialogue is for example when pilot asks for higher flight level and the ATCO provides clearance for that flight level. See example of multiple transmissions grouped into communication groups in Table 10 Example of transmission information and identifiers.	in D1.2
Concept of Operations [ConOps]:	Concept of Operations [ConOps]: The ConOps is jointly elaborated by all ATM stakeholders, from the civil and military airspace users and service providers, to airports and the manufacturing industry to gain common understanding of the ATM system. It describes the operational targets, to move ATM towards trajectory-based operations whereby aircraft can fly their preferred trajectories, considering the matching between constraints and optimization. The ConOps allows all ATM stakeholders, from the civil and military airspace users and service providers, to airports and the manufacturing industry to gain common understanding of the ATM system. In this context, the ConOps is the operational answer to reach the ATM Performance improvements targeted by the ATM MP. Furthermore, the ConOps is an important reference for global interoperability and harmonization, as it has been adapted for Europe from the ICAO Global Air Traffic Management Operational Concept.	See definition in [53]
Controlling Working Position Identifier	The controlling working position identifier is part of the Transmission information and represents a name or index to identify the position that generated that specific transmission. See example in Table 10 Example of transmission information and identifiers.	in D1.2
Exploratory Research	The exploratory research investigates relevant scientific subjects (during the ATM Excellent Science & Outreach phase) and conducts feasibility studies looking for	See definition in [53]

Term	Definition	Source of the definition
	potential application areas in ATM (during the ATM application-oriented research phase).	
Diarization **Diarization Error Rate (DER)**	This is the process of partitioning an input audio stream into homogeneous segments according to the speakers. It is used to answer the question "who spoke when?". Speaker diarization is a combination of speaker segmentation and speaker clustering. The first aims at finding speaker change points in an audio stream. The second aims at grouping together speech segments on the basis of speaker characteristics. In HAAWAII project the term refers to assignment the labels “speech” and “silence” to a continuous wave stream. It is the sum of Miss DER, DER-FA **Miss DER** is the percentage of speech segments which are accidently classified as silence, but contain speech. An error rate Miss DER of 2% means that from 100 seconds of utterances containing speech 2 seconds are lost, because they are classified as silence. **DER-FA (False Alarm):** is the percentage of silence segments, which are wrongly labelled as speech. An error rate DER FA of 4% means that from 100 seconds of utterances containing just silence 4 seconds are not deleted, because they are classified as containing speech.	See Wikipedia.com
False alarm rate	The number of alarms (e.g. read read-back errors), which are not a correct alarm, divided by the total number of alarms see also probability of false alarm: fp / (tp + fp) The “False alarm rate” is 1 – minus Precision tp / (tp + fp)	
Horizon 2020	The EU Framework Programme for Research and Innovation.	SESAR 1, WP14, SESAR 2020
MALORCA	Machine Learning of Speech Recognition Models for Controller Assistance, Horizon 2020 funded project from 2016 to 2018	
PMP deliverable	Output produced by the projects that is submitted to the SJU via the SESAR 2020 collaborative platform and that is subject to quality assessment by the SJU. However, these deliverables do not appear in the grant agreement as contractual deliverables. The production of PMP	See definition in [53]

Term	Definition	Source of the definition
	deliverables is done in support of subsequent contractual deliverables and is described in the PMP.	
Precision	Precision represents the percentage of true predictions out of all the commands which were predicted. tp / (tp + fp)	
Probability of false alarms	This is the probability of a false alarm, which is defined as the number of alarms (e.g. read read-back errors), which are not a correct alarm, divided by the sum of cases correctly not resulting in an alarm (true negative) and the false alarm number (false positive).	
Project Management Plan	Formal, approved document, provided by each SESAR 2020 Solution Project, used to manage its execution. It defines how the project is executed, monitored, controlled, and closed.	See definition in [53]
Read-back error detection rate	The number of correctly detected read-back errors (with or without correction) divided by the total number of read-back errors (with or without correction).	
Read-back error false alarm rate	The number of detected read-back errors, which are not a read-back error, divided by the total number of read-back errors (with or without correction), see also probability of false alarm	
Recall	Recall represents the percentage of actually given commands which were predicted. tp / (tp + fn)	
Segment	A part of the audio recording without any specific property	D3.1

Term	Definition	Source of the definition
SESAR 2020	The SESAR 2020 (Single European Sky ATM Research) Research and Innovation (R&I) Programme will demonstrate the viability of the technological and operational solutions already developed within the SESAR R&I Programme (2008-2016) in larger and more operationally-integrated environments. At the same time, SESAR 2020 will prioritise research and innovation in a number of areas, namely integrated aircraft operations, high capacity airport operations, advanced airspace management and services, optimised network service performance and a shared ATM infrastructure of operations systems and services. SESAR 2020 will retain its founding members, the European Union and Eurocontrol.	SESAR 1, WP14, SESAR 2020, PJ.17-03
Speaker (classification) Error	SpkE: A speech segment is wrongly classified as a pilot segment although the ATCo is speaking and vice versa.	
SpokenData	A generic web based tool which allows to transcribe the speech recordings, while transcribers are supported by several functions to minimise their effort.	D3.1
Transcription	This task involves the speech-to-text transformation, writing down word-by-word, what the ATCo has said. Examples are: “lufthansa two bravo alfa descend flight level eight zero and reduce speed two two zero knots” and “bonjour air_france two seven three [unk] confirm vien* correction contact vienna radar on one two nine decimal five”.	D3.1
Transmission Direction	This is either “ATCo” when the ATCo (ground) speaks to the pilot or “Pilot”, if the pilot (air) speaks to the ATCo.	D1.2
Transmission unique identifier	Transmission unique identifier is part of transmission information and represents a generated unique value or index that is used to distinguish one single transmission from either ATCO or Pilot.	D1.2
TRL 2 (V1)	Technology concept and/or application formulated: Applied research. Theory and scientific principles are focused on very specific application area(s) to perform the analysis to define the concept. Characteristics of the application are described. Analytical tools are developed for simulation or analysis of the application.	See definition in [53]
TRL 3	**Analytical and experimental critical function and/or characteristic proof-of concept:** Proof of concept validation. Active Research and Development (R&D) is	See definition in [53]

Term	Definition	Source of the definition
	initiated with analytical and laboratory studies including verification of technical feasibility using early prototype implementations that are exercised with representative data.	
TRL 4 (V2)	Component/subsystem validation in laboratory environment: Standalone prototyping implementation and test with integration of technology elements and conducting experiments with full-scale problems or data sets.	See definition in [53]
True Positives (tp)	The total number of correctly predicted commands, i.e., the number of commands which were predicted which were actually given.	
False Positives (fp)	The total number of falsely predicted commands, i.e., the number of commands which were predicted but actually NOT given.	
False Negatives (fn)	The total number of commands which were falsely not predicted, i.e., the number of commands which were NOT predicted but were actually given.	
True Negatives (fn)	The total number of commands which were correctly not predicted, i.e., the number of commands which were NOT predicted and actually NOT given.	
Utterance	Segment of an audio file, which consists of a complete message by only one speaker to the other dialogue participants . In case of ATC it contains complete message of ATCO to one pilot or complete answer of pilot to ATCO. Utterance can contain one or more sentences e.g. "Good morning. Speed bird one three seven descend flight level eighty". Utterance segments can be automatically or manually created.	D3.1
WER	WER = (S+D+I)/N, where S -> no. of substitutions, D -> no. of deletions, I -> no. of insertions, C -> no. of correct words, and N -> total no. of words in reference (S+D+C)	D3.4

Table 9: Glossary

References used in Glossary of terms

[52] 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.

[53] SESAR 2020 Execution guidance of ER4 projects : https://ec.europa.eu/research/participants/data/ref/h2020/other/guides_for_applicants/jtis/h2020-guide-project-handbook-er4-sesar-ju_en.pdf

Table 10 Example of transmission information and identifiers.

Transmission unique identifier	ATCO/Pilot Transmission	Clearance transmission Identifier	Controlling Working Position Identifier	Communication Group
1	ATCO: XYZ descend flight level three one zero	1	CWP1	1
2	Pilot: XYZ descending level three one zero	1	CWP1	1
3	ATCO: ASD here Reykjavik control 1, 2,3,4,5 audio check.	NULL	CWP1	2
4	Pilot: I hear you 5 by 5.	NULL	CWP1	2
5	ATCO: ABC descend flight level three one zero	2	CWP1	3
6	Pilot: ABC level one three zero	2	CWP1	3
7	Pilot : ABC correction descending flight level three one zero	2	CWP1	3
8	ATCO: XYZ descend flight level one zero zero	3	CWP1	4
9	Pilot: XYZ descending level one zero zero	3	CWP1	4
10	Pilot: And how is the weather in Keflavik?	NULL	CWP1	4
11	ATCO: Its always still wind and sunny.	NULL	CWP1	4

A.2 Acronyms and Terminology

Term	Definition
ABSR	**Assistant Based Speech Recognition**
ACC	**Area Control Centre**
ACG	**Austro Control Österreichische Gesellschaft für Zivilluftfahrt (Austrian ANSP)**
ADS-B	**Automatic dependent surveillance–broadcast**
AEC	**Approach executive controller**
AFIS	**Aerodrome Flight Information Service**
AG	**Attention Guidance**
AI	**Artificial Intelligence**
AM	**acoustic model**
ANRIC	**Aeronautical Radio Incorporated**

ANSP	**Air Navigation Service Provider**
ANS-CR	**Air Navigation Services of the Czech Republic**
APC	**Approach planning controller**
APP	**Approach**
ARR	**Arrival**
ARTAS	**ATM suRveillance Tracker And Server**
ASR	**Automatic Speech Recognition**
ASTERIX	**All Purpose Structured Eurocontrol Surveillance Information Exchange**
ASW	**Air situation window**
ATC	**Air Traffic Control**
ATCo	**Air Traffic Controller; also ATCO used, but ATCo preferred in HAAWAII project**
ATM	**Air Traffic Management**
Avg	**Average**
BERT	**Bidirectional Encoder Representations from Transformers**
BUT	**Brno University of Technology**
CBA	**Cost Benefit Analysis**
CER	**Command or Context (Prediction) Error Rate, also used as CtxER**
Cmd	**Command (files containing annotations)**
CmDER	**Command Error Rate**
CmDRR	**Command Recognition Rate**
CNN	**Convolution neural network**
CoCoLoToCoCo	**Controller Command Logging Tool for Context Comparison**
Cor	**Correct (files containing transcriptions)**
COTS	**Commercial of the shell**
CPP	**Context Portion Predicted**
CONOPS	**Concept of operations**
CPDLC	**Controller Pilot Data Link Communications**

CTA	Control area
CTR	Controlled traffic region
CtxER	See CER
CV	Clearance verification
CWP	Controller Working Position
DASC	Digital Avionics Systems Conference
DEC	Departure executive controller
DEP	Departure
DER	Diarization Error, see also glossary
DFS	Deutsche Flugsicherung GmbH (German ANSP)
DLR	German Aerospace Center, Deutsches Zentrum für Luft- und Raumfahrt e.V.
DNN	Deep neural network
DPO	Data Protection Officer
DVI	Direct Voice Input
DVO	Direct Voice Output
EATMA	European Air Traffic Management Architecture, An architectural Model of European ATM for each SESAR Concept Story board step containing information relating to Operational activities.
EDR	Event Detection Rate
EML	European Media Laboratory
EN	Enabler
ENAIRE	Spanish ANSP
ER	En-Route
Err	Error (files containing errors)
EU	European Union
EXE	Exercise
FAA	Federal Aviation Administration

FANS	**Future Air Navigation System**
FDPS	**Flight Data Processing System**
FL	**Flight level**
FIR	**Flight Information Region**
ft	**Feet**
G2P	**grapheme-to-phoneme**
GDPR	**General Data Protection Regulation**
GUI	**Graphical User Interface**
HF	**Human factors**
HMI	**Human Machine Interface**
HUP	**Human Performance**
IB	**Information Bottleneck**
ICAO	**International Civil Aviation Organization**
ICE	**Intelligent Communications Environment**
ID	**Identifier**
Idiap	**Idiap Research Institute**
IEC	**Information executive controller**
ILS	**Instrument landing system**
IFR	**Instrument Flight Rules**
ISA	**Instantaneous self assessment**
JSON	**JavaScript Object Notation**
khz	**Kilo hertz**
KPA	**Key Performance Area**
kt	**Knots**
KWA	**Keyword Spotting Algorithm, special implementation of callsign recognition**
LAC	**London Area Control**
LF-MMI	**lattice free maximum mutual information**

LM	language model
LTCC	London Terminal Control Centre
LTMA	London Terminal Manouvering Area
MALORCA	Horizon 2020 funded project MACHINE LEARNING OF SPEECH RECOGNITION MODELS FOR CONTROLLER ASSISTANCE
MFCC	Mel-Frequency Cepstral Coefficients
MMI	maximum mutual information
MWM	Mental Workload Model
N/A	Not applicable
NASA TLX	NASA Task load index
NATS	United Kingdom ANSP
NAT OTS	NORTH ATLANTIC ORGANIZED TRACK SYSTEM
NER	named entity recognition
Nm	Nautical miles
No.	Number
NOK	Not Ok
NPR	Noise Preferential Route
OA	Open Access
Obj	Objective
OI (Step)	Operational Improvement (Step)
OOV	out-of-vocabulary
OSED	Operational services and environment description
OTS	ORGANIZED TRACK SYSTEM
PC	Prestwick Centre
PEC	Director executive controller
PERF	Performance
PJ	Project
POK	Partly Ok

PST	Performance Stability
PSS	Paperless Strip System
PTT	Push to talk
R/T	Radio Telephony
RabbitMQ	is an open-source message-broker software (sometimes called message-oriented middleware)
RED	Readback Error Detection
REDA	Readback Error Detection Assistant
REF	Reference
REQ	Requirement
ReTi	Reaction Time
RMA	Radar Manoeuvring Areas
RNAV	Area navigation
RTP	Real Time Protocol
RWY	Runway
(S)VFR	(Special) Visual Flight Rules
S2T	Speech-To-Text
SA	Situation Awareness
SAD	Speech Activity Detection
SAF / SAFE	Safety
SAR	Safety assessment report
SASHA	Situation Awareness for SHAPE (Solutions for Human Automation Partnerships in European ATM)
SC APP	Approach Senior Controller
Scn	Scenario
SDK	Software Development Kit
SDDS	Surveillance Data Distribution
SESAR	Single European Sky ATM Research

SID	**Standard instrument departure**
SJU	**SESAR Joint Undertaking**
SME	**Subject Matter Experts**
SOL	**Solution**
STAR	**Standard terminal arrival route**
STCA	**Short Term Conflict Alerting**
T2C	**Text-to-Concept**
T2S	**Text-to-Speech**
TC	**Terminal Control**
TDNN-F	**factorized time delay neural network**
TMA	**Terminal Manoeuvring Area**
TRL	**Technology Readiness Level**
TS	**Technical Specification**
TSWR	**Tower**
TTC	**Text-to-Concept**
TTS	**Text-to-Speech**
TVALP	**Technical Validation Plan**
TVALR	**Technical Validation Report**
V2T	**Voice to Text**
V&V	**Validation & Verification**
VAD	**Voice activity detection**
VCS	**Voice communication system**
VFR	**Visual flight rules**
VieAPP	**Vienna Approach**
VRR	**Voice Recognition and Response**
VTT	**Voice to Text**
WDR	**Word Detection Rate, approx. 100% minus WER**

WER	**Word Error Rate**
WL	**Workload**
w.r.t.	**with respect to**
XML	**eXtenable Markup Language**
Term	**Definition**
ATM	Air Traffic Management
SESAR	Single European Sky ATM Research Programme
S3JU	SESAR3 Joint Undertaking (Agency of the European Commission)

Table 11: Acronyms and technology

DLR

BRNO UNIVERSITY OF TECHNOLOGY

idiap RESEARCH INSTITUTE

austro CONTROL

isavia ANS Air Navigation Services

NATS

CROATIA CONTROL