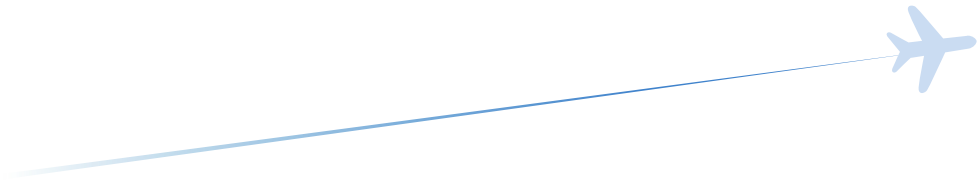

Four Hours of Silence Reduced Annotated Controller and Pilot Utterances for UK

Deliverable ID:	**D3.2**
Dissemination Level:	**PU**
Project Acronym:	**HAAWAII**
Grant:	**884287**
Call:	**H2020-SESAR-2020-2**
Topic:	**SESAR-ER4-18-2019**
Consortium Coordinator:	**DLR**
Edition date:	**10 June 2021**
Edition:	**01-00-00**
Template Edition:	**02.00.02**

Founding Members
EUROPEAN UNION EUROCONTROL
SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Shruthi Shetty (DLR)	Lead of Command Extraction in WP3	20.05.2021
Hartmut HELMKE (DLR)	Project Lead	24.05.2021

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Hartmut HELMKE (DLR)	Project Lead	24.05.2021
Shruthi Shetty (DLR)	Lead of Command Extraction in WP3	20.05.2021

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

Name/Beneficiary	Position/Title	Date
Hartmut HELMKE	Project Lead	10.06.2021
Pavel SMRZ (BUT)	WP4 Leader	By silent approval
Petr MOTLICEK (Idiap)	Contributor	By silent approval
Christian WINDISCH (ACG)	D6.4 Leader	By silent approval
Teodor S. SIMIGANOSCHI (Isavia)	WP1 Leader	By silent approval
Julia HARFMANN (NATS)	WP5 Leader	By silent approval
Martina ERCEG (CCL)	Contact Point CCL	By silent approval

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
0.01	17.05.2021	Draft	S. Shetty	First Version
0.02	18.05.2021	Draft	S. Shetty	Section 4 updates
0.03	19.05.2021	Draft	S. Shetty	Section 5 updates
0.04	20.05.2021	Draft	S. Shetty	Unfinished parts completed
0.05	20.05.2021	Draft	S. Shetty	Final version before HHe review

0.05	24.05.2021	Draft	H. Helmke	Review
0.05	25.05.2021	Draft	S. Shetty	Incorporated HHe's feedback, accepted changes and deleted all comments
1.00	10.06.2021	Final	H. Helmke	Preparation for Delivery to SJU: update of copyright and author list

Copyright Statement

HAAWAII

HIGHLY ADVANCED AIR TRAFFIC CONTROLLER WORKSTATIONS WITH ARTIFICIAL INTELLIGENCE INTEGRATION

This General document is part of a project that has received funding from the SESAR 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. Evaluation of controllers' feedback has been subdued due to the limited recognition performance of the commercial of the shell ASR engines that were used, even in laboratory conditions. Past exploratory research funded project MALORCA, however, has shown (on restricted use-cases) that satisfactory performance can be reached with novel data-driven machine learning approaches. The project builds on very large collection of data, organized with a minimum expert effort to develop a new set of models for complex environments of Icelandic en-route and London TMA. The deliverable is public.

The deliverable D2.2 has described collecting of surveillance data and voice recordings from London ANS approach airspace followed by the transcription process. As already Alan Turing pointed out, speech recognition does not include speech understanding. Therefore, this deliverable concentrates on the semantic interpretation, i.e. the annotation, of the transcribed voice recordings by using the information from corresponding voice recordings. Currently 7.5 hours of manually transcribed pilot and ATCo utterances are available from London airspace. 57 minutes of them are manually annotated, the remaining 6.5 hours are automatically annotated. On command level, recognition rates of 92.5% and 90.34% are achieved for ATCo and pilot utterances, respectively for transcriptions which are manually transcribed. These recognition rates are achieved after one third of project runtime. For automatic transcriptions based on current implementation of speech-to-text transformation, a performance of 84.5% and 78.8% for ATCos and pilots is achieved, respectively if surveillance data is used for callsign extraction. If surveillance data is not available, recognition rates of 75.2% for ATCos and 71.4% for pilots are achieved. The challenge of the next months will be to manually check automatic annotations and to improve the extraction rate especially from automatic transcriptions by improving both extraction performance and speech to text transformation, i.e. reducing word error rate (WER). Recognition performance of above 90% from the output of the speech-to-text process is the objective to enable readback error detection support for ATCos.

Table of Contents

List of Tables

1 Executive Summary

As Alan Turing pointed out, speech recognition does not include speech understanding: Even with a perfect speech to text transformation, i.e. a perfect automatic speech recognition (ASR) system, we do not know the semantics of the following two air traffic controller (ATCo) utterances “good morning speed bird two zero zero zero alfa reduce one eight zero knots until DME four miles contact tower on frequency one one eight decimal seven zero zero” and “one eighty to DME four tower one eighteen seven speed bird two thousand alfa”. Even a perfect ASR system would not be able to identify that both utterances mean the same thing on a semantic level.

Therefore, this deliverable concentrates on the semantic interpretation, i.e., the annotation of the transcribed voice recordings by using the information from corresponding voice recordings. Currently 7.5 hours of manually transcribed pilot and ATCo utterances are available from London airspace. Utterances corresponding to about 57 minutes of voice data are manually annotated, while the remaining 6.5 hours are annotated automatically. On command level, recognition rates of 92.5% and 90.34% are achieved for ATCo and pilot utterances, respectively for transcriptions which are manually transcribed. These recognition rates are achieved after one third of project runtime. For automatic transcriptions based on current implementation of speech-to-text transformation, a performance of 84.5% and 78.8% for ATCos and pilots is achieved, respectively if surveillance data is used for callsign extraction. If surveillance data is not available, recognition rates of 75.2% for ATCos and 71.4% for pilots are achieved. Table 1 provides a good summary of the current command extraction performance and also about the available data.

Type	Total	User-Rec	Relevance	Rec-Rate	Rec-Rate-ASR
AFFIRM	5	5	0.3%	100.0%	50.7%
ALTITUDE	132	125	7.3%	94.7%	75.8%
CALL_YOU_BACK	3	3	0.2%	100.0%	66.7%
CLIMB	160	154	8.8%	96.3%	80.2%
CONTACT	68	64	3.8%	94.1%	70.9%
CONTACT_FREQUENCY	131	121	7.2%	92.4%	80.1%
CONTINUE PRESENT_HEADING	12	11	0.7%	91.7%	
DESCEND	215	211	11.9%	98.1%	61.5%
DIRECT_TO	94	77	5.2%	81.9%	58.0%
DISREGARD	1	1	0.1%	100.0%	
FAREWELL	97	88	5.4%	90.7%	81.8%
FOLLOW_ROUTE	67	63	3.7%	94.0%	
GREETING	114	111	6.3%	97.4%	92.6%
HEADING LEFT	60	60	3.3%	100.0%	100.0%
HEADING RIGHT	42	40	2.3%	95.2%	100.0%
HEADING none	90	84	5.0%	93.3%	36.4%
HOLDING	3	3	0.2%	100.0%	0.0%
INFORMATION ACTIVE_RWY	19	16	1.0%	84.2%	58.3%
INFORMATION QNH	54	54	3.0%	100.0%	56.8%
INIT_RESPONSE	1	1	0.1%	100.0%	84.6%
MAINTAIN ALTITUDE	2	2	0.1%	100.0%	63.9%
MAINTAIN PRESENT_SPEED	4	3	0.2%	75.0%	
MAINTAIN SPEED	2	1	0.1%	50.0%	
NAVIGATION_OWN	3	3	0.2%	100.0%	100.0%
NEGATIVE	4	4	0.2%	100.0%	71.4%
NO_CONCEPT	77	59	4.2%	76.6%	59.3%
NO_SPEED_RESTRICTIONS	12	9	0.7%	75.0%	75.0%
REPORT_MISCELLANEOUS	9	9	0.5%	100.0%	64.0%
REPORT_NOW HEADING	20	20	1.1%	100.0%	
STATION	101	100	5.6%	99.0%	82.2%
Sum of all types	1612	1511	88.9%	93.7%	
Sum of relevant types	1376	1288	75.9%	93.6%	75.5%

Table 1: Summary of current command extraction performance

All command types which were observed at least 10 times are considered as important for the read back error detection application are shown in Table 1. Command types such as NEGATIVE and DISREGARD occur seldom, but are considered to be relevant or important. Currently the data base consists of only 57 minutes of manually verified annotated data.

In Table 1, we see how often each type occurred (column "Total"), how often the command is correctly extracted and the relevance of a given type ("Total" of the given type divided by the number of all command types). Column "Rec-Rate" contains "User-Rec" divided by column "Total", i.e., the command recognition rate for the given type. Column "Rec-Rate-ASR" contains the command recognition rate, when the command extraction [5] is performed on automatically transcribed data and not on manually transcribed and checked transcriptions. We mark command types with recognition rates which are significantly below the average in yellow and command types which are significantly better in green.

Improvable recognition rates for the command types “CLIMB”, “ALTITUDE” and “DESCEND” results from the fact that these commands are often said using sheer altitude values, specially by pilots. Even though the command extractor is capable of extracting such commands, it fails when there is ambiguity between 2 or more command types. Another reason contributing to < 100% recognition rates is that sometimes altitude values are partially said (Eg: “climb to flight level one two”, here “one two” is said instead of “one two zero”). Improvable performance for “DIRECT_TO” type results from conditional clearances, which are not modelled.

The challenge in the coming months will be to manually check automatic annotations and to improve the extraction rate for automatic transcriptions by improving both extraction performance and speech to text transformation, i.e. reducing word error rate (WER). The objective is to achieve a recognition performance of above 90% from the output of the speech-to-text process, in order to enable readback error detection support for ATCos.

2 Introduction

2.1 Purpose of the document

The HAAWAII project builds on a very large collection of data, organized with a minimum expert effort to develop a new set of models for complex environments of Icelandic en-route and London TMA.

This deliverable describes annotation and the automatic annotation process for London's approach airspace data recorded during July to September 2020.

2.2 Scope

The HAAWAII project aims to research and develop a reliable, error resilient and adaptable solution to automatically transcribe voice commands issued by both air-traffic controllers and pilots. To develop new models, large audio data with corresponding transcription is required. The ANSPs of Icelandic en-route and London TMA (Terminal Manoeuvring Area) collect the audio data of the controllers and the pilots.

In order to process the large amount of voice data for building machine learning models, it is important to first transcribe and annotate the commands in the most efficient way possible.

The **transcription** 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". The **annotation** task extracts the semantic concepts from the transcriptions (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".

The transcriptions and annotations are needed on one hand for creating first models for the speech recognizers, which are later improved by automatic transcriptions and annotations. On the other hand, the manual transcriptions and annotations are needed for evaluating the recognition performance of the developed speech recognitions engines. Therefore, high quality transcriptions and annotations are needed. The performance of the speech recognizer is limited by the quality of the available training data.

This deliverable focusses on the annotations, while transcriptions are already described in D2.3 [1].

2.3 Intended readership

This document is mainly intended for:

- **HAAWAII** consortium members in order to have a common and shared view of the command extraction and annotation process
- **SESAR JOINT UNDERTAKING (SJU)** as Horizon 2020 Programme coordinator.

The information HAAWAII consortium members is characterized as follows:

- Chapter 3 is interesting for NATS technicians providing the surveillance data.

- Chapter 4 concentrates on the current performance of command extraction from manually transcribed utterances. It describes the used metrics (e.g. Command Recognition Rate, Callsign Recognition Rate) in detail. It shows performance on command level as well as on callsign level. The chapter is interesting for Idiap, BUT, Isavia and NATS.

- Chapter 5 addresses the performance when command extraction is performed using automatically transcribed data. It is interesting for Idiap and BUT to understand the impact of using automatic transcriptions for speech understanding.

- Table 18 and Table 19 contain a summary of performance on command level and considering recognition of just callsigns by comparing the performance of manual and automatic transcriptions and by also comparing the performances based on whether surveillance information is used or not. These tables might be interesting for all partners and also for the bodies from SJU.

- Section 6 describes the next steps that we plan to follow in order to improve the command extraction performance.

- Last but not least the whole document is important for DLR staff, who are responsible for improving the command extraction (annotation) performance. It shows the current status with quantifiable numbers. It shows that many challenges still exist, but also shows that even more is already achieved.

2.4 Background

The background mainly consists of the recent projects on Assistant Based Speech Recognition

- Helmholtz Validation Fund supported AcListant®, AcListant®-Strips (Active Listening Assistant)
- SESAR2020 Exploratory Research MALORCA (Machine Learning of Speech Recognition Models for Controller Assistance)
- SESAR2020 Industrial Research PJ.16-04-02 ASR (CWP HMI)
- Publication for the DASC 2020 by DLR, which publishes an initial version of the command extraction.

2.5 Structure of the document

The structure of this document is based on the Horizon 2020 template for project deliverables. It is organized as follows:

- **Chapter 1: Executive Summary**. Provides a summary of the key information and elements contained in the Technical Validation Report document.
- **Chapter 2: Introduction** (this chapter). Introduces the document.
- **Chapter 3:** Provides a description of the recorded voice data and the corresponding surveillance data, especially addressing the performance of current implementation of command prediction and the limitations due to other used data formats and missing data items compared to e.g. MALORCA project.
- **Chapter 4:** Provides an overview of the current performance based on manually checked voice transcriptions.

- **Chapter 5:** Provides an overview of the current performance based on automatically transcribed data
- **Chapter 6:** Describes the challenges of the next months with respect to improvements of requirements, speech-to-text transformation and command and callsigns extraction.

2.6 Glossary of terms

The HAAWAII project has more than 20 different deliverables. Therefore, the HAAWAII project decides to have one separate document containing the glossary of terms, so that maintenance of the terms is eased and errors or misunderstandings only need to be changed in one place.

For simplifying the task of the readers, the contents of the master document are shown in the following table.

Term	Definition	Source of the definition
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 [1]
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.	in D1.2

Term	Definition	Source of the definition
	"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.	
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.	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.	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 2 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 [1]
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 [1]
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 [1]
Communication group	Communication group is part of transmission information and it is a generated value or index that is used to identify	in D1.2

Term	Definition	Source of the definition
	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 2 Example of transmission information and identifiers.	
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 [2]
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 2 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 potential application areas in ATM (during the ATM application-oriented research phase).	See definition in [2]
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	

Term	Definition	Source of the definition
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 deliverables is done in support of subsequent contractual deliverables and is described in the PMP.	See definition in [2]
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 [2]
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).	
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
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

Term	Definition	Source of the definition
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 [2]
TRL 3	**Analytical and experimental critical function and/or characteristic proof-of concept:** Proof of concept validation. Active Research and Development (R&D) is initiated with analytical and laboratory studies including verification of technical feasibility using early prototype implementations that are exercised with representative data.	See definition in [2]
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 [2]
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.	
Recall	Recall represents the percentage of actually given commands which were predicted. tp / (tp + fn)	

Term	Definition	Source of the definition
Precision	Precision represents the percentage of true predictions out of all the commands which were predicted. tp / (tp + fp)	
Accuracy	Accuracy represents the prediction rate. It also takes into account the number of commands which were correctly NOT predicted. (tp + tn) / (tp + fp + fn +tn)	
Segment	A part of the audio recording without any specific property	D3.1
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
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

Reference used in Glossary of terms

[1] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, and M. Schulder, “Assistant-Based Speech Recognition for ATM Applications,” in 11th USA/ Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[2] 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

Transmission unique identifier	ATCO/Pilot Transmission	Clearance transmission Identifier	Controlling Working Position Identifier	Communication Group
1	ATCO: XYZ descend flight level three one zero	LLAP	LLAP	1
2	Pilot: XYZ descending level three one zero	LLAP	LLAP	1
3	ATCO: hello london ABC two five nine altitude flight level two zero zero until xyz	TMASOUTH	TMASOUTH	2
4	Pilot:two zero zero xyz ABC two five nine	TMASOUTH	TMASOUTH	2

Table 2 Example of transmission information and identifiers

2.7 Acronyms and terminology

The HAAWAII project has more than 20 different deliverables. Therefore, the HAAWAII project decides to have one separate document containing these acronyms, so that maintenance of the acronyms is eased and errors or misunderstandings only need to be changed in one place.
For simplifying the task of the readers, the contents of the master document are shown in the following table.

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
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
BUT	Brno University of Technology
CBA	Cost Benefit Analysis
CER	Command or Context (Prediction) Error Rate, also used as CtxER

Term	Definition
Cmd	Command (files containing annotations)
CmDER	Command Error Rate
CmDRR	Command Recognition Rate
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
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
ENAIRE	Spanish ANSP
ER	En-Route
Err	Error (files containing errors)
EU	European Union

Term	Definition
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
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
LTCC	London Terminal Control Centre
LTMA	London Terminal Manouvering Area
MALORCA	Horizon 2020 funded project MACHINE LEARNING OF SPEECH RECOGNITION MODELS FOR CONTROLLER ASSISTANCE

Term	Definition
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
Nm	Nautical miles
No.	Number
NOK	Not Ok
NPR	Noise Preferential Route
OA	Open Access
Obj	Objective
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)
REF	Reference
REQ	Requirement
ReTi	Reaction Time
RMA	Radar Manoeuvring Areas
RNAV	Area navigation
RTP	Real Time Protocol
RWY	Runway

Term	Definition
(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
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

Term	Definition
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% - WER
WER	Word Error Rate
WL	Workload
w.r.t.	with respect to
XML	eXtenable Markup Language

3 Description of Used Data

The following Table 3 lists the directories, for which transcribed voice data for London higher and lower approach airspace is available.

Day-time and Sector	# Wav	# Cor	# Cmd	Wav[s]	Cor[s]	Cmd[s]
2020-07-12 TMASOUTH_1401_1434	446	446	446	1392	1392	1392
2020-08-01 LLAP_1457_1551	300	300	300	1101	1101	1101
2020-08-01 TMASOUTH_0659_0742	346	346	41	1079	1079	195
2020-08-01 TMASOUTH_1200_1258	323	323	23	927	927	91
2020-08-01 TMASOUTH_1656_1734	360	360	26	1074	1074	112
2020-08-02 LLAP_0917_1005	126	126	0	482	482	0
2020-08-02 TMASOUTH_0803_0816	119	119	2	415	415	13
2020-08-02 TMASOUTH_1104_1128	240	240	6	742	742	31
2020-08-02 TMASOUTH_1404_1417	82	82	2	219	219	10
2020-08-03 LLAP_0557_0635	107	107	5	496	496	21
2020-08-03 LLAP_0849_0959	276	276	10	1143	1143	40
2020-08-03 LLAP_1000_1159	707	707	15	2739	2739	67
2020-08-03 LLAP_1200_1359	880	880	11	3081	3081	52
2020-08-03 LLAP_1400_1528	419	419	9	1570	1570	42
2020-08-03 LLAP_1530_1651	486	486	11	1724	1724	41
2020-08-03	396	0	0	1414	0	0

LLAP_1651_1803						
2020-08-08 TMASOUTH_0558_0759	861	861	17	2762	2762	79
2020-08-08 TMASOUTH_0800_0959	764	764	19	2442	2442	67
2020-08-08 TMASOUTH_1000_1205	814	814	22	2541	2541	94
Sum	8052	7656	965	07:35:43	07:12:09	00:57:28

Table 3: Transcribed Voice Recordings from London approach airspace from 2020

In total 8052 wave files with voice recordings are available, out of which 7656 of them are transcribed. The duration of all these 8052 wave files is 7 hours and 35 minutes. Currently only 965 files are manually annotated. The duration of these annotations is approximately 57 minutes. Two directories (grey colour in Table 3) are manually annotated completely. The column #Cmd shows the number of manually annotated commands in each of the directories. The remaining files are all automatically annotated.

The following Table 4 shows the number of aircraft, which are in the air or in other words, the aircraft for which surveillance data is available. The last three columns show the number of callsigns which are predicted, i.e. callsigns for which the ATCo may give a command to in the next minute or those aircraft where the pilot may initiate a call with the ATCo.

		Number of Aircraft			Callsigns in Context		
Day-time and Sector	**# Files**	**# min**	**# max**	**#aver**	**# min**	**# max**	**#aver**
2020-07-12 TMASOUTH_1401_1434	446	250	281	264	100	136	122
2020-08-01 LLAP_1457_1551	300	168	253	222	83	127	107
2020-08-01 TMASOUTH_0659_0742	346	41	73	51	32	51	43
2020-08-01 TMASOUTH_1200_1258	323	195	227	211	95	124	109
2020-08-01 TMASOUTH_1656_1734	360	73	115	100	45	72	60
2020-08-02 LLAP_0917_1005	126	175	222	195	84	110	100
2020-08-02 TMASOUTH_0803_0816	119	66	95	79	44	63	52
2020-08-02 TMASOUTH_1104_1128	240	220	259	242	108	130	120
2020-08-02	82	204	220	212	92	117	104

		Number of Aircraft			Callsigns in Context		
Day-time and Sector	**# Files**	**# min**	**# max**	**#aver**	**# min**	**# max**	**#aver**
TMASOUTH_1404_1417							
2020-08-03 LLAP_0557_0635	107	34	57	51	31	54	48
2020-08-03 LLAP_0849_0959	276	104	172	133	64	105	82
2020-08-03 LLAP_1000_1159	707	152	187	170	75	113	100
2020-08-03 LLAP_1200_1359	880	138	190	169	74	115	95
2020-08-03 LLAP_1400_1528	419	136	172	154	70	97	84
2020-08-03 LLAP_1530_1651	486	85	147	122	47	80	69
2020-08-03 LLAP_1651_1803	396	56	96	73	36	62	48
2020-08-08 TMASOUTH_0558_0759	861	38	93	59	35	61	48
2020-08-08 TMASOUTH_0800_0959	764	90	274	194	57	145	105
2020-08-08 TMASOUTH_1000_1205	814	210	297	258	101	154	130
Sum / Average	8052	128.2	180.5	155.7	67	100.8	85.6

Table 4: Number of aircraft with surveillance data and number of callsigns predicted

From the above table, we see that on an average there are about 155.7 aircraft in the air with surveillance information. The average minimum and maximum aircraft with surveillance information are 128.2 and 180.5, respectively. The average number of predicted callsigns in context is about 85.6. From the above table, we see that the number of callsigns predicted correlates with the number of aircraft in the air. On a heavy traffic day like the 2nd of August 2020, there were about 242 aircraft in the air on an average and hence the average number of callsigns predicted (120) were also higher as compared to the other days.

This is the current status of callsign prediction which is expected to improve. We hope that a reduction of 50% in the average number and in the maximal number of predicted callsigns is possible. One of the challenges is that no flight plan information is available compared to MALORCA project, i.e. it is not clear from the surveillance data if an aircraft is an inbound, outbound or an overflight. However, a majority of flights are overflight, which do not land or take-off from a London airport. The decision whether an aircraft is an arrival can only be taken from future surveillance data, i.e., depending on whether it is lands on a London runway or not.

The following Table 5, however, shows a general problem with HAAWAII callsign prediction for London airspace. The table shows the number of issued commands, which are derived automatically for most directories, i.e. no manual checking was performed. We expect that these numbers corresponding to callsign errors would slightly decrease when automatic extraction is improved and further manual checking and correction are done.

Day-time and Sector	Sum Cmds	NO_CALL SIGN	NO_CO NCEPT	# Error s	Area Errors	No Radar	% Err
2020-07-12 TMASOUTH_1401_1434	653	39	27	13	0	13	1.90%
2020-08-01 LLAP_1457_1551	490	18	10	15	0	15	3%
2020-08-01 TMASOUTH_0659_0742	545	13	15	10	0	10	1.80%
2020-08-01 TMASOUTH_1200_1258	439	54	33	19	0	19	4.30%
2020-08-01 TMASOUTH_1656_1734	512	23	25	30	0	30	5.80%
2020-08-02 LLAP_0917_1005	218	14	10	1	0	1	0.40%
2020-08-02 TMASOUTH_0803_0816	197	5	4	15	15	0	7.60%
2020-08-02 TMASOUTH_1104_1128	324	35	10	16	0	16	4.90%
2020-08-02 TMASOUTH_1404_1417	96	20	5	0	0	0	0%
2020-08-03 LLAP_0557_0635	195	10	2	31	0	31	15.80%
2020-08-03 LLAP_0849_0959	499	15	18	11	0	11	2.20%
2020-08-03 LLAP_1000_1159	1279	48	39	0	0	0	0%
2020-08-03 LLAP_1200_1359	1486	50	44	1	0	1	0%
2020-08-03 LLAP_1400_1528	685	22	31	1	0	1	0.10%
2020-08-03 LLAP_1530_1651	752	51	39	0	0	0	0%
2020-08-03	0	0	0	0	0	0	0%

Day-time and Sector	Sum Cmds	NO_CALL SIGN	NO_CO NCEPT	# Error s	Area Errors	No Radar	% Err
LLAP_1651_1803							
2020-08-08 TMASOUTH_0558_0759	1424	33	45	41	0	41	2.80%
2020-08-08 TMASOUTH_0800_0959	1176	93	50	22	14	8	1.80%
2020-08-08 TMASOUTH_1000_1205	1189	94	42	15	0	15	1.20%
Sum / Average	12159	637	449	241	29	212	2.82%
Only for the manually annotated directories	1143	57	37	28	0	28	2.45%

Table 5: Accuracy of command prediction

“NO_CALLSIGN” means that for this file/utterance no callsign was extracted, which in most cases is contributed to the fact that no callsign was said by the ATCo or the pilot. “NO_CONCEPT” means that no command from the ontology [6] was extracted for this utterance, which is the case for utterances like “euro trans three” or “if requesting to deviate north of valdi should be no problem”.

“#Errors” refers to the number of commands with a callsign other than NO_CALLSIGN for which the callsign is currently not predicted. It is the sum of the following two columns.

“Area Errors” refers to the number of errors for which surveillance data is available, but the aircraft is currently outside the expected lat/long rectangle.

“No Radar” refers to the number of commands for which the extracted callsign is not found in the surveillance data. For utterances which are currently automatically annotated, it could also be because a callsign is wrongly extracted by the command extractor. Another commonly occurring reason is an aircraft receiving a clearance from the ATCo a few minutes before it is first seen or a few minutes after it was last seen.

“% Err” is the error percentage, i.e. column “# Errors” divided by column “Sum Cmds”.

4 Current Performance of Automatic Annotation on gold transcriptions

The following Table 6 shows the performance of command extraction [5] for the ATCo utterances which are manually annotated. As mentioned before, two directories are already completely annotated. For the other directories, only the interesting cases are manually annotated. In most cases these are the challenging cases for which the extraction code must be updated. Therefore, the performance is slightly lower as compared to the completely manually annotated folders.

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-07-12 TMASOUTH_1401_1434	330	97.3%	0.91%	1.82%	2489	295	39
2020-08-01 LLAP_1457_1551	226	95.1%	1.33%	3.54%	1917	201	7
2020-08-01 TMASOUTH_0659_0742	16	62.5%	12.5%	31.3%	88	22	4
2020-08-01 TMASOUTH_1200_1258	3	100%	0%	0%	17	2	0
2020-08-01 TMASOUTH_1656_1734	5	100%	0%	0%	31	14	1
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories					
2020-08-02 TMASOUTH_0803_0816	0						
2020-08-02 TMASOUTH_1104_1128	0						
2020-08-02 TMASOUTH_1404_1417	0						
2020-08-03 LLAP_0557_0635	4	100%	0%	0%	50	8	0
2020-08-03 LLAP_0849_0959	3	100%	0%	0%	20	3	0
2020-08-03 LLAP_1000_1159	23	56.5%	26.1%	17.4%	115	13	0
2020-08-03 LLAP_1200_1359	6	16.7%	16.7%	83.3%	34	13	0
2020-08-03	7	62.5%	25%	25%	76	11	0

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
LLAP_1400_1528							
2020-08-03 LLAP_1530_1651	7	71.4%	0%	28.6%	56	17	1
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory						
2020-08-08 TMASOUTH_0558_0759	1	0%	100%	0%	12	6	0
2020-08-08 TMASOUTH_0800_0959	20	95%	5%	0%	102	5	1
2020-08-08 TMASOUTH_1000_1205	12	83.3%	8.3%	16.7%	108	25	2
Sum / Average	664	92.5%	3.01%	5.12%	5115	635	55

Table 6: Accuracy of automatic command extraction for ATCo utterances

Command recognition rates are computed by comparing instructions from **manual** human annotation (gold annotation) to the results of the **automatic** semantic extraction (command extraction). For a given speech utterance, each instruction is treated as one big word. Then, the Levenshtein distance between the gold annotation and the results of command extraction is calculated, resulting in the number of substitutions (subs), insertions (ins) and deletions (del). Table 7 gives an overview about the different metrics and illustrates an example how they are calculated. In the table #gold defines the total number of commands in the gold annotation. #match defines the number of matches, which is #gold – subs – del. The table also shows why the sum of RcR, ErR and RjR can be greater than 100%. This is the case when more commands are recognized than what is really said.

Metric	Calculation	
Command Recognition Rate (**RcR**)	**RcR** = #matches / #gold	
Command Recognition Error Rate (**ErR**)	**ErR** = (subs + ins) / #gold	
Command Rejection Rate (**RjR**)	**RjR** = del / #gold	
Example		
Gold Annotation	**Command Extraction**	
AFR123 INIT_RESPONSE AFR123 TURN LEFT AUA1AB SPEED 140 kt DLH123_NO_CONCEPT	AFR123 DIRECT_TO OKG none AFR123 INIT_RESPONSE AFR123 TURN RIGHT AUA1AB NO_CONCEPT DLH123 NO_CONCEPT	
Result:		
RcR = 2/4 = 50% (green)	*ErR = 2 / 4 = 50% (purple)*	*RjR =1/4 = 25% (yellow)*

Table 7: Metric Definition for Command Recognition Performance

If the result of the command extraction contains either NO_CONCEPT or NO_CALLSIGN, these substitutions and insertions are always calculated as deletions, i.e., these extractions contribute to the rejection rate and not to the error rate (as shown in the example in Table 7).

"#Words" refers to the total number of words in the corresponding manually transcribed utterance. "Unknown Cl" refers to the number of words which are not used for extracting the commands, i.e., they are not classified as callsign, type, value, unit, qualifier or condition. "Unknown 2" refers to the number of consecutive word pairs which are classified as unknown. Three or more consecutive unknown classifications are not counted here. Table 8 provides an example.

one	one nine	one	one nine decimal zero	vista	jet seven zero eight	have	a	good	day
vare	vare	vare	vare	unkn	unkn	unkn	unkn	csgn	csgn	csgn	csgn	csgn	unkn	unkn	type	type
				##########									##########			

Table 8: Example for Number of Consecutive Unknowns

We have 17 words (#Words) in the above example utterance, out of which 6 words are classified with "unkn" (Unknown Cl) as shown in the second row and we have 1 sequence of two consecutive unknown classification in the word sequence "have a" (column Unknown 2).

For calculating the callsign recognition rates CaR, CaE and CaRj, see definitions in Table 9. Here, we just compare the callsigns from the gold annotation and from the automatic extraction. For each utterance we consider the callsign only once, except when different callsigns are annotated or extracted. For the example in Table 7 this results in the three annotated and extracted callsigns AFR123, AUA1AB and DLH123.

Metric	Calculation	
Callsign Recognition Rate (**CaR**)	Same as RcR but only for callsigns without instructions, which is number of all utterances minus the wrong callsign recognitions divided by all utterances (UttCnt -WrongCsgn) / UttCnt	
Callsign Recognition Error Rate (**CaE**)	Same as ErR, but only for callsigns without instructions (InventedCsgn + NoCsgnMissed + BreakBreak) / UttCnt,	
Callsign Rejection Rate (**CaRj**)	Same as RjR, but only for callsigns without instructions NoExtraction / UttCnt.	
If the command extraction results in different callsigns, the calculation is done for each callsign. See example below, which also illustrate that the sum of RcR, ErR and RjR can exceed 100%.		
Example		
Gold Annotation	**Command Extraction**	
AFR123 INIT_RESPONSE AFR123 TURN LEFT AUA1AB SPEED 140 kt DLH123_NO_CONCEPT	AFR123 DIRECT_TO OKG none AFR123 INIT_RESPONSE AFR123 TURN RIGHT AUA1AB NO_CONCEPT AFR1YY NO_CONCEPT	
Result:		
CaR = 2/3 = 67% (green)	*CaE = 1 / 3 = 33% (purple)*	*CaRj =0/3 = 0%*

Table 9: Metric Definition for Callsign Recognition Performance

Table 10 provides the callsign recognition performance for all annotated ATCo utterances.

Day-time and Sector	UttCnt	Wrong Csgn	Invented Csgn	No Extraction	No Csgn missed	Break Break
2020-07-12 TMASOUTH_1401_1434	330	1	0	0	1	0
2020-08-01 LLAP_1457_1551	226	1	0	0	1	0
2020-08-01 TMASOUTH_0659_0742	16	0	0	0	0	0
2020-08-01 TMASOUTH_1200_1258	3	0	0	0	0	0
2020-08-01 TMASOUTH_1656_1734	5	0	0	0	0	0
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories				
2020-08-02 TMASOUTH_0803_0816	0					
2020-08-02 TMASOUTH_1104_1128	0					
2020-08-02 TMASOUTH_1404_1417	0					
2020-08-03 LLAP_0557_0635	4	0	0	0	0	0
2020-08-03 LLAP_0849_0959	3	0	0	0	0	0
2020-08-03 LLAP_1000_1159	23	3	0	0	3	0
2020-08-03 LLAP_1200_1359	6	1	0	1	0	0
2020-08-03 LLAP_1400_1528	7	1	0	1	0	0
2020-08-03 LLAP_1530_1651	7	0	0	0	0	0
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory					
2020-08-08	1	1	1	0	0	0

Day-time and Sector	UttCnt	Wrong Csgn	Invented Csgn	No Extraction	No Csgn missed	Break Break
TMASOUTH_0558_0759						
2020-08-08 TMASOUTH_0800_0959	20	0	1	0	0	0
2020-08-08 TMASOUTH_1000_1205	12	1	1	0	0	0
Sum	398	9	2	2	5	0
Rates	CaR	97.7%	CaE	1.8%	CaRj	0.5%

Table 10: Accuracy of automatic callsign extraction/recognition rates for ATCo utterances

Column "UttCnt" contains the number of utterances considered, i.e., the number of wave files. Column "Wrong Csgn" shows the number of cases for which a callsign was extracted from the utterance but the callsign was wrong. It is the sum of the following four columns: "Invented Csgn", "No Extraction", "No Csgn missed" and "Break Break". Let's take an example utterance "euro trans three" consisting of 3 words. The extracted callsign for this utterance was "BCS3", but the correct callsign is "BCS3998". This is possible because the "BCS3998" is the only euro trans at that time. "Invented Csgn" refers to the number of cases where a callsign was recognized which was not said. "No Extraction" refers to the number of cases in which "NO_CALLSIGN" was wrongly extracted and that a callsign was actually provided. "No Csgn missed" refers to the number of cases in which a callsign was extracted, but no callsign was actually said. "Break Break" represents the number of cases in which more than one different callsign was said by the ATCo, but only one was extracted, e.g. in "lufthansa alfa bravo descend flight level six zero break break **speed bird four alfa nine** call you back stand by".

The following Table 11 corresponds to Table 6, but it contains the numbers only for pilot utterances.

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-07-12 TMASOUTH_1401_1434	373	94.9%	1.34%	4.02%	2483	227	34
2020-08-01 LLAP_1457_1551	298	93%	1.01%	6.04%	2189	238	23
2020-08-01 TMASOUTH_0659_0742	112	93.8%	7.14%	5.4%	622	107	9
2020-08-01 TMASOUTH_1200_1258	64	90.6%	0%	10.9%	318	54	3
2020-08-01 TMASOUTH_1656_1734	73	80.8%	12.3%	6.9%	369	65	7
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories					
2020-08-02	7	85.7%	14.3%	0%	40	13	0

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
TMASOUTH_0803_0816							
2020-08-02 TMASOUTH_1104_1128	27	92.6%	0%	7.41%	103	8	2
2020-08-02 TMASOUTH_1404_1417	8	87.5%	0%	12.5%	36	8	0
2020-08-03 LLAP_0557_0635	4	75%	0%	25%	20	4	0
2020-08-03 LLAP_0849_0959	14	92.9%	0%	7.14%	111	29	1
2020-08-03 LLAP_1000_1159	13	69.2%	30.8%	15.4%	139	38	1
2020-08-03 LLAP_1200_1359	21	76.2%	28.6%	9.52%	161	32	2
2020-08-03 LLAP_1400_1528	10	50%	50%	0%	77	18	5
2020-08-03 LLAP_1530_1651	17	70.6%	11.8%	23.5%	85	18	0
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory						
2020-08-08 TMASOUTH_0558_0759	53	88.70%	1.89%	9.43%	254	32	2
2020-08-08 TMASOUTH_0800_0959	20	85%	10%	10%	135	29	1
2020-08-08 TMASOUTH_1000_1205	35	71.40%	8.57%	20%	215	41	6
Sum / Average	1149	90.3%	4.3%	6.8%	7357	961	96

Table 11: Accuracy of automatic command extraction for Pilot utterances

The table shows the challenges on pilot side. The extraction rate on ATCo side (92.5%) is better than on pilot side (90.3%).

Table 12 corresponds to Table 10 and shows the callsign recognition/extraction rate for pilot utterances.

Day-time and Sector	UttCnt	Wrong Csgn	Invented Csgn	No Extraction	No Csgn missed	Break Break
2020-07-12 TMASOUTH_1401_1434	233	1	0	1	0	0
2020-08-01 LLAP_1457_1551	173	2	0	1	1	0
2020-08-01 TMASOUTH_0659_0742	34	1	0	1	0	0
2020-08-01 TMASOUTH_1200_1258	21	0	0	0	0	0
2020-08-01 TMASOUTH_1656_1734	23	2	1	0	1	0
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories				
2020-08-02 TMASOUTH_0803_0816	2	0	0	0	0	0
2020-08-02 TMASOUTH_1104_1128	6	0	0	0	0	0
2020-08-02 TMASOUTH_1404_1417	2	0	0	0	0	0
2020-08-03 LLAP_0557_0635	2	0	0	0	0	0
2020-08-03 LLAP_0849_0959	9	1	0	1	0	0
2020-08-03 LLAP_1000_1159	7	3	0	1	2	0
2020-08-03 LLAP_1200_1359	8	1	0	0	1	0
2020-08-03 LLAP_1400_1528	5	1	1	0	0	0
2020-08-03 LLAP_1530_1651	6	2	0	2	0	0
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory					
2020-08-08 TMASOUTH_0558_0759	16	0	0	0	0	0

Day-time and Sector	UttCnt	Wrong Csgn	Invented Csgn	No Extraction	No Csgn missed	Break Break
2020-08-08 TMASOUTH_0800_0959	10	0	0	0	0	0
2020-08-08 TMASOUTH_1000_1205	15	3	0	1	2	0
Sum for Pilots	572	17	2	8	7	0
Rates for Pilots	CaR	97%	CaE	1.6%	CaRj	1.4%
Sum for ATCos	398	9	2	2	5	0
Rates for ATCos	CaR	97.7%	CaE	1.8%	CaRj	0.5%

Table 12: Accuracy of automatic callsign extraction/recognition rates for Pilot utterances

From Table 12 and, Table 10 we see that the callsign extraction/recognition rates for pilots and ATCos have similar performances. The improvable challenges are on the command extraction rates for pilots.

5 Current Performance of Automatic Annotation on automatic transcriptions

The previous chapter illustrates the current performance of command extraction/recognition on manually transcribed utterances. Good performance on the output of a speech recognizer is, however, more important.

Therefore, we evaluated the performance of command extraction by using the automatically generated transcriptions provided by BUT in May 2021.

The Word Error Rates (WER) calculated are shown in the following two tables Table 13 and Table 14.

ATCo WER	3.9%
Pilot WER	6.8%
Total	5.4%

Table 13: Average Word Error Rates for Pilot and ATCo calculated for all files

ATCo WER	6.4%
Pilot WER	10.4%
Total	8.4%

Table 14: Average Word Error Rates for Pilot and ATCo calculated for only validation files

WER provided in Table 14 are more realistic with respect to expected rates in the future, because these recordings were excluded from the training data.

Table 15 shows the results of command extraction on automatic transcriptions for ATCo utterances.

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-07-12 TMASOUTH_1401_1434	330	94.2%	3.3%	3.33%	2517	288	34
2020-08-01 LLAP_1457_1551	226	83.6%	13.7%	4.42%	1919	216	7
2020-08-01 TMASOUTH_0659_0742	16	56.3%	18.8%	31.3%	91	21	3
2020-08-01 TMASOUTH_1200_1258	3	66.7%	0%	33.3%	19	2	0
2020-08-01 TMASOUTH_1656_1734	5	40%	60%	0%	35	12	0
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories					

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-08-02 TMASOUTH_0803_0816	0						
2020-08-02 TMASOUTH_1104_1128	0						
2020-08-02 TMASOUTH_1404_1417	0						
2020-08-03 LLAP_0557_0635	4	100%	0%	0%	50	8	0
2020-08-03 LLAP_0849_0959	3	100%	0%	0%	20	3	0
2020-08-03 LLAP_1000_1159	23	26.1%	26.1%	47.8%	82	11	0
2020-08-03 LLAP_1200_1359	6	16.7%	16.7%	83.3%	30	9	0
2020-08-03 LLAP_1400_1528	7	62.5%	25%	25%	76	11	0
2020-08-03 LLAP_1530_1651	7	57.1%	14.3%	28.6%	42	6	1
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory						
2020-08-08 TMASOUTH_0558_0759	1	0%	200%	0%	11	4	0
2020-08-08 TMASOUTH_0800_0959	20	95%	5%	0%	100	4	0
2020-08-08 TMASOUTH_1000_1205	12	50%	50%	16.7%	105	21	2
Sum / Average	664	84.5%	10.1%	7.4%	4763	570	47

Table 15: Accuracy of automatic command extraction for ATCo utterances, when the input does not result from manual transcription but from automatic transcription

Table 16 shows the results of command extraction on automatic transcriptions for Pilot utterances.

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-07-12 TMASOUTH_1401_1434	373	84.5%	10.5%	8.3%	2488	278	33
2020-08-01	298	69.5%	18.1%	15.4%	2166	332	25

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
LLAP_1457_1551							
2020-08-01 TMASOUTH_0659_0742	112	84.8%	15.2%	12.5%	614	107	13
2020-08-01 TMASOUTH_1200_1258	64	87.5%	3.1%	10.9%	317	55	2
2020-08-01 TMASOUTH_1656_1734	73	78.1%	13.7%	8.2%	369	58	5
2020-08-02 LLAP_0917_1005	0	No manually annotated commands in these directories					
2020-08-02 TMASOUTH_0803_0816	7	71.4%	0%	28.6%	35	11	0
2020-08-02 TMASOUTH_1104_1128	27	85.2%	3.7%	11.1%	105	13	2
2020-08-02 TMASOUTH_1404_1417	8	75%	12.5%	12.5%	37	11	0
2020-08-03 LLAP_0557_0635	4	75%	0%	25%	20	4	0
2020-08-03 LLAP_0849_0959	14	71.4%	21.4%	7.1%	110	28	1
2020-08-03 LLAP_1000_1159	13	69.2%	23.1%	23.1%	111	24	1
2020-08-03 LLAP_1200_1359	21	76.2%	28.6%	4.8%	143	24	3
2020-08-03 LLAP_1400_1528	10	30%	60%	10%	74	18	3
2020-08-03 LLAP_1530_1651	17	64.7%	17.6%	23.5%	85	17	0
2020-08-03 LLAP_1651_1803	No gold transcriptions for this directory						
2020-08-08 TMASOUTH_0558_0759	53	86.8%	5.6%	9.4%	255	33	3

Day-time and Sector	# gold	RcR	ErR	RjR	# Words	Unknown Cl	Unknown 2
2020-08-08 TMASOUTH_0800_0959	20	85%	10%	10%	130	25	1
2020-08-08 TMASOUTH_1000_1205	35	74.3%	8.6%	20%	219	41	5
Sum / Average	1149	78.8%	13.3%	11.7%	7278	1079	97

Table 17: Accuracy of automatic command extraction for Pilot utterances, when the input does not result from manual transcription but from automatic transcription

We also calculated the performance when the predicted callsigns extracted from the surveillance data were not used by the command extraction tool for both cases, i.e. when commands are extracted from both manually and automatically transcribed utterances.

		ATCo			Pilot		
		RcR	ErR	RjR	RcR	ErR	RjR
Manual	With Callsigns	92.5%	3.0%	5.1%	90.3%	4.3%	6.8%
	No Callsigns	75.2%	11.4%	14.3%	71.4%	13.4%	17.3%
Auto-matic	With Callsigns	84.5%	10.1%	7.4%	78.8%	13.3%	11.7%
	No Callsigns	73.5%	14.8%	14.0%	66.4%	18.9%	18.9%

Table 18: Performance of automatic command extraction for ATCo and Pilot utterances when manually transcribed (Manual) versus automatically transcribed (Automatic) and when callsign information is provided (With Callsigns) versus when no callsign information from command prediction is provided (No Callsigns)

Table 18 shows that the extraction recognition performance for ATCO commands goes down from 92.5% to 84.2% when automatically generated transcriptions are used. The same behaviour is observed for pilot utterances where the recognition rate decreases from 90.3% to 78.8%. The error rates also increase for both ATCo and Pilot from 3.0% to 10.1% and 4.3% to 13.3%, respectively.

Table 18 also shows that the recognition rates RcR decrease drastically for both pilot and ATCo utterances when no callsign information is provided. The error rates ErR, on the other hand, increase for both cases. This behaviour is observed irrespective of whether manual or automatic transcriptions are used.

The following Table 19 provides the performance when only the callsigns are considered.

		ATCO			Pilot		
		CaR	CaE	CaRj	CaR	CaE	CaRj
Manual	With Callsigns	97.7%	1.8%	0.5%	97%	1.6%	1.4%
	No Callsigns	80.9%	10.3%	8.8%	76%	12.1%	11.9%
Auto-matic	With Callsigns	91.2%	6.8%	2%	92.1%	5.4%	2.4%
	No Callsigns	81.4%	11.1%	7.5%	76.6%	13.5%	10%

Table 19: Performance of automatic callsign extraction for ATCo and Pilot utterances, when manual transcribed (manual) versus automatically transcribed (automatic) and when callsign information of available callsigns is provided (With Callsigns) versus when no callsign information from command prediction is provided (No Callsigns)

The results in Table 19 are similar to that of Table 18. Automatic transcriptions result in a minor performance decrease which can be further improved with reduced word error rates. Moreover, callsign information from the surveillance data are of decisive importance in order to correctly extract callsigns and commands.

6 Next Steps

The next steps to be followed for improving the quality of command extraction is the same as that for Isavia. Please refer deliverable D3.3 [4].

7 References

[1] 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.

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

[3] H. ARILÍUSSON, T. SIMIGANOSCHI, H. HELMKE, J. HARFMANN: HAAWAII project: D6.2: Updated Operational Concept Document, version 01.50.00, intermediate version, 16. May 2021.

[4] H. HELMKE, S. Shetty.: HAAWAII project: D3.3: Four Hours of Silence Reduced Annotated Controller and Pilot Utterances for Iceland, version 01.00.00, 25. May 2021.

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

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


HAAWAII


SESAR
JOINT UNDERTAKING


DLR


BRNO
UNIVERSITY
OF TECHNOLOGY


idiap
RESEARCH INSTITUTE


austro
CONTROL


isavia ANS
Air Navigation Services


NATS


CROATIA
CONTROL