



Article

Ensuring Safety for Artificial-Intelligence-Based Automatic Speech Recognition in Air Traffic Control Environment

Ella Pinska-Chauvin ^{1,*}, Hartmut Helmke ², Jelena Dokic ¹, Petri Hartikainen ¹, Oliver Ohneiser ²
and Raquel García Lasheras ³

¹ Integra Consult A/S, Staktoften 20, 1., 2950 Vedbaek, Denmark; jdj@integra.dk (J.D.); pha@integra.dk (P.H.)

² German Aerospace Center (DLR), Institute of Flight Guidance, Lilienthalplatz 7, 38108 Braunschweig, Germany; hartmut.helmke@dlr.de (H.H.); oliver.ohneiser@dlr.de (O.O.)

³ ATM Research and Development Reference Centre (CRIDA A.I.E.), Las Mercedes Business Park, C/de Campezo 1, 28022 Madrid, Spain; rglasheras@e-crida.enaire.es

* Correspondence: epc@integra.dk

Abstract: This paper describes the safety assessment conducted in SESAR2020 project PJ.10-W2-96 ASR on automatic speech recognition (ASR) technology implemented for air traffic control (ATC) centers. ASR already now enables the automatic recognition of aircraft callsigns and various ATC commands including command types based on controller–pilot voice communications for presentation at the controller working position. The presented safety assessment process consists of defining design requirements for ASR technology application in normal, abnormal, and degraded modes of ATC operations. A total of eight functional hazards were identified based on the analysis of four use cases. The safety assessment was supported by top-down and bottom-up modelling and analysis of the causes of hazards to derive system design requirements for the purposes of mitigating the hazards. Assessment of achieving the specified design requirements was supported by evidence generated from two real-time simulations with pre-industrial ASR prototypes in approach and en-route operational environments. The simulations, focusing especially on the safety aspects of ASR application, also validated the hypotheses that ASR reduces controllers' workload and increases situational awareness. The missing validation element, i.e., an analysis of the safety effects of ASR in ATC, is the focus of this paper. As a result of the safety assessment activities, mitigations were derived for each hazard, demonstrating that the use of ASR does not increase safety risks and is, therefore, ready for industrialization.

Keywords: safety assessment; air traffic control; automatic speech recognition; workload; situational awareness; en-route sector; approach sector



Citation: Pinska-Chauvin, E.; Helmke, H.; Dokic, J.; Hartikainen, P.; Ohneiser, O.; Lasheras, R.G. Ensuring Safety for Artificial-Intelligence-Based Automatic Speech Recognition in Air Traffic Control Environment.

Aerospace **2023**, *10*, 941. <https://doi.org/10.3390/aerospace10110941>

Academic Editor: Michael Schultz

Received: 11 September 2023

Revised: 25 October 2023

Accepted: 29 October 2023

Published: 3 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Automatic speech recognition (ASR) in the air traffic management (ATM) domain is seen as a promising technology for improving efficiency and safety [1]. Use of ASR technology in ATC environments consists of three conceptual steps. First, a speech-to-text conversion is performed, i.e., an ATC utterance such as “lufthansa two seven victor descend flight level two hundred” is transcribed from the speech signal into a sequence of words. This is followed by the text-to-concepts transformation, i.e., the semantics of the transcription are extracted as machine-readable ontology-conforming annotations with aircraft callsigns and various command elements such as “DLH27V DESCEND 200 FL”. In the third step, the output of the two preceding steps is directly presented on the air traffic controllers' (ATCO) human machine interface (HMI) enabling, amongst other benefits, the replacement of manual HMI inputs by the ATCOs. The following ASR functionalities were covered by the safety assessment relevant to this paper:

1. Recognition of relevant aircraft callsigns from ATCO and pilot utterances as well as highlighting the callsigns on the controller working position (CWP) HMI display.

2. Recognition of ATCO commands and input of the command contents into the aircraft radar data labels displayed at the ATCO CWP HMI.

As the ASR technology is relatively new, it is not yet deployed in the various ATM systems developed by the industry. This is due in part to the fact that European regulation [2] requires any new technology introduced into the operational environment undergoes a rigorous safety assessment conducted at the design phase, providing relevant evidence that the physical design satisfies the design requirements. The safety of two ASR prototypes for air traffic control (ATC) purposes supported by artificial intelligence (AI) has been assessed by Integra in accordance with the Single European Sky Air Traffic Management Research (SESAR) safety reference materials [3] in the course of SESAR2020 project PJ.10-W2-96-ASR [4,5]. The safety assessment includes several steps, starting at the design level with the assessment of the introduction of a new system—or a change to an existing system—for the identification of any hazards introduced by the new system elements, and possible associated increase in risks. The system must be proven to be safe in a specific environment by demonstrating that the level of safety is not degraded, and that at least the same level of safety can be achieved as prior to the introduction of the change.

The considered ASR prototypes were designed to improve ATCOs' situational awareness, reduce their workload and increase their productivity. The scope of the safety assessment considered application of ASR supporting ATCOs with the aforementioned goals in approach and en-route sectors of medium traffic complexity.

The main goal of this article is to evaluate possible safety risks introduced by the implementation of ASR and the possible impact of these risks on ATC operations, in order to derive mitigations formulated as system design requirements. In the next step, these system design requirements are implemented into two pre-industrial prototype platforms to demonstrate the feasibility of the evaluated design and expected safety levels of system performance.

In the next section, we provide background work of ASR applications in the ATM domain, including previous safety-related work performed on the topic. In Section 3, the safety assessment performed in the context of ASR application developed in project PJ.10-W2-96 to derive the safety requirements for the design is described. In this section, we also describe the setup of the human-in-the-loop simulations that were conducted to gather the evidence required for the safety assessment. Section 4 presents the results of the two simulations and the safety assessment related results and also the limitations of this study, which is based on only two validation experiments. Concluding remarks are given in Section 5. Lastly, two appendices are included containing additional information for the hazard analysis performed.

2. Background

The SESAR2020 project PJ.16-04 "CWP HMI" (*Controller Working Position Human Machine Interface*) [6] investigated the feasibility of ASR with early prototypes applied in the air traffic control domain. Those prototypes were validated in laboratory-like environments equivalent to technology readiness level (TRL) 4 as per industrial research development standards [7]. Technology readiness levels (TRLs) are a method for estimating the maturity of technologies during research and development phase, that enables consistent and uniform assessment. TRLs are based on a scale from 1 to 9 with 9 being the most mature technology.

The basis of this paper, i.e., the follow-up project PJ.10-W2-96 [4,5] continued developing the application with the aim of demonstrating the technology feasibility in relevant operational environment (TRL6). In parallel, the project PJ.05-W2-97 *HMI Interaction modes for Airport Tower* aimed to develop an ASR system for an aerodrome control tower environment [8]. The ATCO2 platform [9] aims at collecting, pre-processing, and pseudo-anonymizing ATC communications' audio databases of more than 5000 h of audio data with the objective of increasing robustness of speech recognition in the air traffic management domain. The ATCO2 corpus has also been used to detect speaker roles in voice

communication, i.e., pilot or ATCO, and clustering speakers [10]. Given enough training data, automatic speech recognition and understanding systems also build the base to train ATCOs [11]. A common goal of prior presented ASR research projects was to define an initial ontology for the annotation of ASR recognized ATC concepts such as command types, values, units, and qualifiers, to be later coordinated and agreed between major European ATM stakeholders enabling industrialization of the technology [12].

Early results demonstrated that ASR facilitated safety in operational environments by detection of read-back errors from comparison of controller and pilot radio communication at aerodrome control towers [13]. ASR together with deep-learning-based methods were applied as safety monitoring function by translating the pilot-controller voice communications into texts, which were then converted to contextual data to be analyzed for flight conformance verification, and potential conflict detection [14]. The Venture capital funded project AcListant[®] [15] and AcListant[®]-Strips for ATC approach areas focused on ASR with the aim to significantly reduce controllers' workload [1] and increase ATM efficiency [15]. The exercise of DLR and Austro Control of SESAR2020 project PJ.10-W2-96-ASR, described in detail in Section 3.4, has the same objectives, using Vienna approach and not Dusseldorf as validation airspace. The main difference is that the focus of project PJ.10-W2-96 is on investigating safety aspects regarding the number of erroneous recognitions of ASR that are undetected by the ATCO [5]. These results are summarized in Section 4.2.2.

The STARFiSH (Safety and Artificial Intelligence Speech Recognition) [16] project integrated AI-based speech recognition into an A-SMGCS (Advanced Surface Movement Guidance and Control System) for ground traffic control and monitoring at Frankfurt Airport. The joint application of ASR and A-SMGCS recognized the instructions given by apron controllers to pilots, extracted the commands contained therein and integrated the outputs to the user interface of the A-SMGCS. An additional safety net for AI applications in ASR is intended to ensure that errors in AI-based speech recognition do not have any negative effects on the overall system [16]. The benefits of callsign recognition from flight crew utterances and highlighting the callsign at an en-route CWP HMI were investigated in [4]. The study demonstrated feasibility of the integrated ASR system for the identification of callsigns from flight crew utterances which provide benefits in terms of workload and situational awareness. The papers also highlighted the importance of further work on recognition and timeliness of outputs.

Incorporating ASR into ATC specifically as a safety enhancing feature has been researched by various practitioners, especially in conjunction with integrating ASR into various other safety features contained in an ATM system such as conformance monitoring and trajectory prediction. Karlsson et al. previously hypothesized in 1990 on this basis that the introduction of ASR technology into ATC could result in a reduced occurrence of human-generated errors enabling in turn increased safety of the overall system [17]. More recently in 2023, the use of ASR as a safety enhancing application in ATC operations was investigated and noted that the solutions investigated can improve the safety of ATC operations and can contribute to the reduction in ATCO workload [18]. Zhou et al. argued that ASR represents a gateway between the ATM system and the ATCO in converting speech signal to text inputs and that after spoken instruction understanding (SIU) is applied to the converted text the output information can be used to support safety-critical applications (SCA), enabling safety and reducing possible human errors [19]. The European Union Aviation Safety Agency recently developed a roadmap for the approval and deployment of safety-related AI systems for end-user support (pilots and ATCOs) [20]. In following guidance [21], the process of safety assurance for AI level 1 (assistance to human) and AI level 2 (human machine teaming) is developed with the further classification for different level of safety analysis depending on the AI application level. According to the guidance, AI-supported ASR application can be classified as AI Level 2A: human/machine teaming representing human and AI-based system cooperation. The work presented in this paper did not explicitly address the trustworthiness of AI application as the elements of the safety analysis. This paper concentrates on the initial part of safety assessment in the design

phase for such an application. The safety assessment must be continued “in-service” via a data-driven AI safety risk assessment based on operational data and occurrences.

3. Materials and Methods

The validated applications demonstrated ASR capability to effectively support ATCOs by showing evidence appropriate at the pre-industrial feasibility level. This section describes the safety assessment process conducted in accordance with the SESAR Safety Reference Material [3] and its guidance [22] at the design phase to ensure that the proposed implementation of ASR in ATM operations is capable in satisfying the performance requirements as stipulated by European regulation [2]. The SESAR safety assessment process has to demonstrate that the design is safe by using two different approaches:

1. A *success* approach, in which the effectiveness of the new concepts and technologies is assessed, when they are working as intended, i.e., how much the pre-existing risks that are inherent and already present in aviation will be reduced by the changes to the ATM system under assessment, i.e., defining the positive contribution to aviation safety that the ATM changes under assessment may deliver in the absence of failure.
2. A *failure* approach, in which the ATM system generated risks, induced by the ATM changes under assessment are evaluated. This approach defines the negative contribution to the risk of an accident that the ATM changes under assessment may induce in the event of failure(s), however caused.

This paper focuses on the process of deriving the performance requirements for the failure approach based on identification of potential hazards presented by the introduction of ASR, thus ensuring safe implementation of ASR technology to ATC operations.

3.1. Selected Use Cases

The safety assessment covered TRL 5 system development phase, representing technology validated in relevant operational environment, and TRL 6—technology demonstrated in relevant operational environment. For this reason, the operational use cases were selected by a group of subject matter experts from the field who represent the possible users of the technology. The scope of the assessment described was limited to the following uses cases:

3.1.1. Use Case “Highlight of Callsigns (Aircraft Identifier) on the CWP Based on the Recognition of Pilot Voice Communications”

In the scope of this use case, the pilot’s voice signal was extracted, processed by ASR for callsign recognition and further verified against contextual flight plan data. This type of use of ASR technology supports the ATCO by identifying new flights entering the sector and making initial contact on the ATC VHF channel, and flight crews requesting actions from ATCOs, e.g., trajectory change, flight level change or information.

3.1.2. Use Case “Highlight of Callsigns on the CWP Based on the Recognition of ATCO Voice Communications”

The ATCO voice signal was extracted, processed by ASR for callsign recognition and further verified against contextual flight plan data. This type of ASR application, where the aircraft callsign is highlighted and is based on the recognition of ATCO voice communications, provides a safety check to the ATCO who will be able to detect, whether there is a difference between the aircraft callsign mentioned and the flight radar data label on the CWP HMI, for which commands are being input.

3.1.3. Use Case “Annotation of ATCO Commands”

The ATCO voice command was extracted, processed, and verified against contextual data to provide the annotation of a specific command on the CWP HMI. This type of ASR application, where annotation of given commands is made available to ATCOs for consultation, enables increased situational awareness and provides a safety check of

clearances and instructions given to flights. This use case is an intermediate step prior to the semi-automatic/automatic input of commands in the CWP using ASR.

3.1.4. Use Case “Pre-Filling of Commands in the CWP”

The recognized (and validated) commands were presented to the ATCOs together with the command values in the CWP. ATCOs were able to accept, reject or correct the commands.

Two validation exercises were selected, which jointly address all of the use cases noted above:

1. The exercise performed by CRIDA, Indra and ENAIRE places emphasis on a very low callsign recognition error rate (approx. 0%). Consequently, a lower callsign recognition rate (between 50% and 85%) is foreseen. This exercise will be referred to as the *Callsign Highlighting* exercise in the rest of the text.
2. The second exercise performed by DLR and Austro Control attempts to identify a compromise between low callsign recognition error rate (<1%) and acceptable callsign recognition rate (>97%). This exercise will be referred to as the *Radar Label Maintenance* exercise in the rest of the text.

Both approaches are different with regard to solving the callsign highlighting use cases and are, therefore, very interesting from the perspective of safety considerations. More details of the two validation exercises are provided in Section 3.3 for the *Callsign Highlighting* exercise and in Section 3.4 for the *Radar Label Maintenance* exercise. The safety assessment methodology is addressed before the exercise descriptions.

3.2. Safety Assessment Methodology

The focus of this paper is to present the “failure” part of the assessment, thus the contribution of ASR to the risk of an accident that the ATM changes under assessment may induce in the event of failure(s). The process starts with a hazard identification based on the analysis of the use cases using the walk-through technique supported by sequence diagrams as shown in Figure 1.

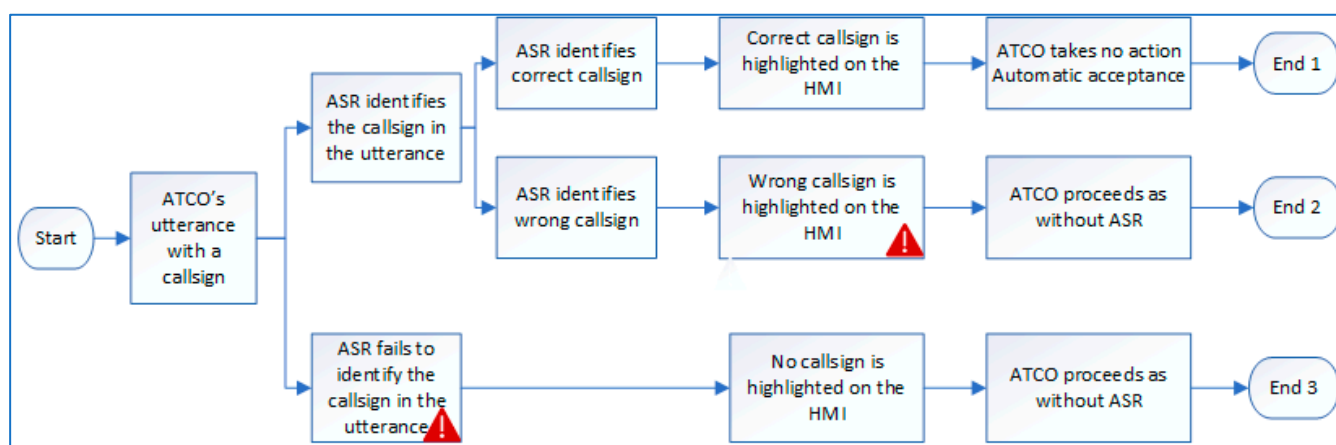


Figure 1. Sequence diagram of use case 2—“Highlight of callsigns on the CWP from ATCO utterances” with indications of the hazard’s occurrence, (indicated by the exclamation marks).

Sequence diagrams were produced and analyzed for each use case. The sequence diagrams were used for the walkthrough with subject matter experts to identify potential hazards, meaning each situation that could trigger the unsafe situation. The identified hazards were further assessed according to Functional Hazard Assessment as per Safety Assessment Methodology [23] by applying the following steps:

1. Identification of hazards’ effects on operations, including the effect on aircraft operations.
2. Assessment of the severity of each hazard effect.

3. Specification of target performance (safety objectives), i.e., determination of the maximum tolerable frequency of the hazard's occurrence.

A top-down causal analysis was performed for each functional hazard, their causes and associated mitigations. The identified mitigations refer to preventive mitigations for a functional hazard, which either prevent a basic cause from occurring or protect against the propagation of the basic cause effect up to the functionality hazard occurrence.

A complementary bottom-up analysis of the failure modes of the ASR elements/element-to-element interfaces and of their effects was performed in order to determine potential common cause failures.

3.3. Callsign Highlighting Exercise with Focus on Low Callsign Recognition Error Rates

ENAI, Indra and CRIDA, conducted an exercise to validate the performance of the pre-industrial ASR prototype covering the following use cases [4]:

- Use Case 1. Highlight of callsigns ... based on the recognition of pilot voice,
- Use Case 2. Highlight of callsigns ... based on the recognition of ATCO voice,
- Use Case 3. Annotation of ATCO commands.

This validation exercise used two complementary approaches aiming at providing evidence of ASR applications' performance by providing the following outputs:

1. Collection of subjective operational feedback from ATCOs gathered by means of questionnaires, debriefings and observations. This was achieved through a real-time human-in-the-loop simulation.
2. Collection of statistically significant objective data regarding ASR performance. This was achieved through the analysis of operational recordings of real-life communications between ATCOs and flight crew. Audios from different Spanish en-route sectors were processed by the ASR system to obtain the accuracy on callsign identification and command annotation.

The human-in-the-loop validation exercise simulated two en-route sectors of Madrid Flight information Region (FIR) during the nighttime. The sectors are presented in Figure 2 obtained from Enaire's Aeronautical Information web application [24], each sector in a different color. The sectors are quite wide and have several entry points where flight crew performs their first call (related to use case 1). There are very different traffic flows that require different type of control commands (related to use case 3) and facilitate the creation of situations where the traffic is focused in one area or dispersed along the whole sector (related to use cases 1 and 2).

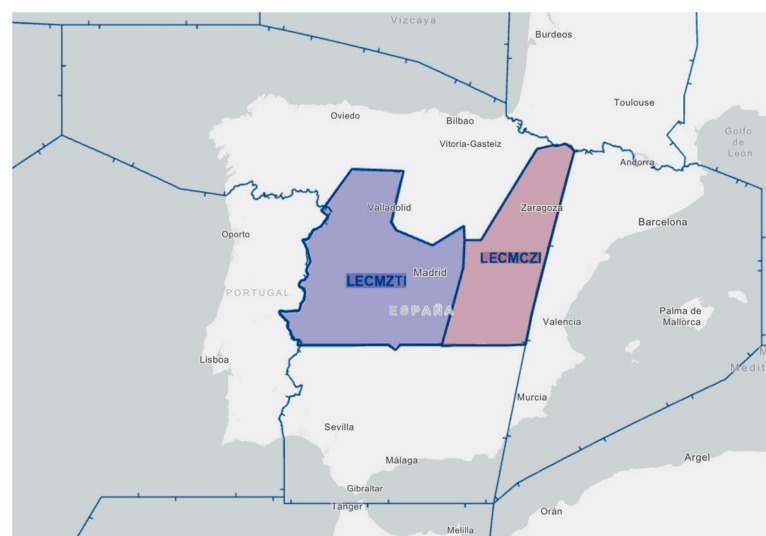


Figure 2. Madrid FIR Simulated sectors in the real-time human-in-the-loop simulation.

The validation exercises were performed in an integrated sector, where one ATCO performs both the executive controller and planning controller roles as is typical in many ACCs during the nighttime. One simulation pilot was assigned for each sector [4]. The exercises were designed with medium-to-high traffic load. A total of two ATCOs from Enaire took part in the simulations in November 2021.

To overcome the limitations of the validation activity (the low number of scheduled runs and participating ATCOs and (simulation) pilots, and the locations specificity to the validated operational environment) a statistical approach was applied. The statistical approach included the analysis of operational recordings from different types of sectors and several actors, both controllers and flight crew. The operational data also serve as a reference to compare performance between laboratory data with real-life data.

During the real-time simulation, it was possible to enable only ATCO speech recognition, pilot speech recognition or both. The different ASR functionalities (callsign recognition and command history window) were activated or deactivated to assess the three use cases separately.

Communications between ATCO and pilot were performed using COMETA, the communication system that ENAIRE has deployed in Spanish ATC units. COMETA uses VoIP and the version used for the exercise was the latest available.

When a radio voice communication is performed the ASR is triggered. The ASR system identifies the callsign in the communication and highlights the corresponding aircraft radar track symbol on the CWP. The ASR also extracts relevant information from ATCO utterances and proceeds to annotate them in a separate window that the ATCO is able to consult.

Context information, i.e., information regarding flight plans and their updates, were sent to the ASR prototype by the simulation platform to reduce callsign recognition error rate. Only callsigns that were completely recognized and present in the flight plans were displayed to the ATCOs, i.e., wrong or partial callsign recognition was considered as not identified, and no flight was highlighted on the screen.

As presented in Figure 3, the callsign recognition is indicated by displaying a white circle around the radar track symbol. The circle flashed for five seconds before disappearing. The functionality allowed highlighting several aircraft at the same time by flashing the indicator around their respective radar track symbols simultaneously.

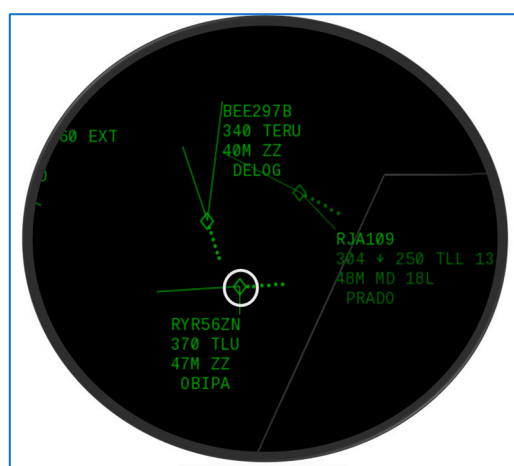


Figure 3. Callsign illumination by flashing.

The annotation window, shown in Figure 4, contains information regarding the commands provided by the ATCO. It includes the callsign of the addressed aircraft, the issuing time, the command annotation in accordance with the standard agreed between the SESAR partners, and an action column. The text in the action column and the colors in the annotation window are coherent with other elements in the CWP. As presented in the figure,

if a callsign was not identified by the ASR system, it appeared as NO_IND but the transcription was available in the text field. An annotation window per flight, where only the communications exchanged with the selected flight appeared, was also available.



INDICATIV	HORA	TEXTO	ACCION
RYR6094	1333	RYR6094 RADAR_CONTACT	
RYR6094	1333	RYR6094 CONTACT_FREQUENCY 133.35	TRF
PGT400A	1332	PGT400A DIRECT_TO LASPO	DCT LASPO
AFL2500	1332	AFL2500 CONTACT_FREQUENCY 118755	TRF
NO_IND	1332	212500 CONTACT_FREQUENCY 118755	TRF

Figure 4. Annotation window with callsign, time, and command content.

The annotation window was displayed only for consultation. ATCOs did not update the information displayed but were able to navigate and sort it.

A mixture of subjective and objective data was used to assess the achievement of the objectives of the exercise. Subjective data were collected via:

- Individual questionnaires: standard and specific questionnaires were developed to assess the validation objectives. The questionnaires were agreed with the subject matter experts participating in dedicated Safety and Human performance workshops.
- Debriefing sessions: after each simulation run the findings, i.e., opportunities, difficulties, general findings observed during the exercise were discussed among all participants (operational and simulation staff).
- Over the shoulder observations: direct and non-intrusive over-the-shoulder observation were carried out by human factors expert, during the runs. This non-intrusive observation had the purpose of providing detailed, complete and reliable information on the way the activity is carried out, especially, if further commented and discussed with the observed users during the debriefing.

Additionally, objective data were obtained from system data recorded during each session by the replay and post-analysis tools. These data contained information on callsign transcription and command annotation generated during the simulation. Data on system reaction times were also recorded. Further quantitative data were obtained from the analysis of operational recordings. The recognized callsigns were compared to the correct callsigns resulting from manual annotations (gold standard annotations).

Two statistical analyses of the outputs were performed: The first one used objective data collected from the real-time simulation (RTS) screen and audio recordings. The second one used operational recordings from different Spanish en-route ATC sectors. These sectors were selected taking into account their complementary characteristics that provided a wide sample of technical (i.e., signal-to-noise ratio, native speakers origin) and operational (i.e., type of commands) characteristics. The statistical analyses were obtained by manually transcribing the recordings, creating the callsign and command annotation standards, and then comparing them against the ASR outcome.

Further details can be found in the project final report [25].

3.4. Radar Label Maintenance Exercise with Focus on High Callsign Recognition Rates

DLR together with Austro Control conducted a real-time human-in-the-loop simulation at DLR's premises in Braunschweig to validate the performance of the pre-industrial ASR prototype covering the following use cases:

- Use Case 2. Highlight of callsigns . . . based on the recognition of ATCO voice,
- Use Case 3. Annotation of ATCO commands,
- Use Case 4. Pre-filling of commands in the CWP.

The focus of the simulation was to quantify the benefits of ASR with respect to operational safety and ATCO workload. The traffic scenarios consider inbound flights to

Vienna airport runway 34. Departures and overflights were not modelled. The ATCO, however, was responsible for the four adjacent approach sectors BALAD, MABOD, PESAT and NERDU plus the terminal maneuvering area (TMA) including the landing clearance roughly 6 to 10 miles before touch down, see Figure 5 taken from [5].

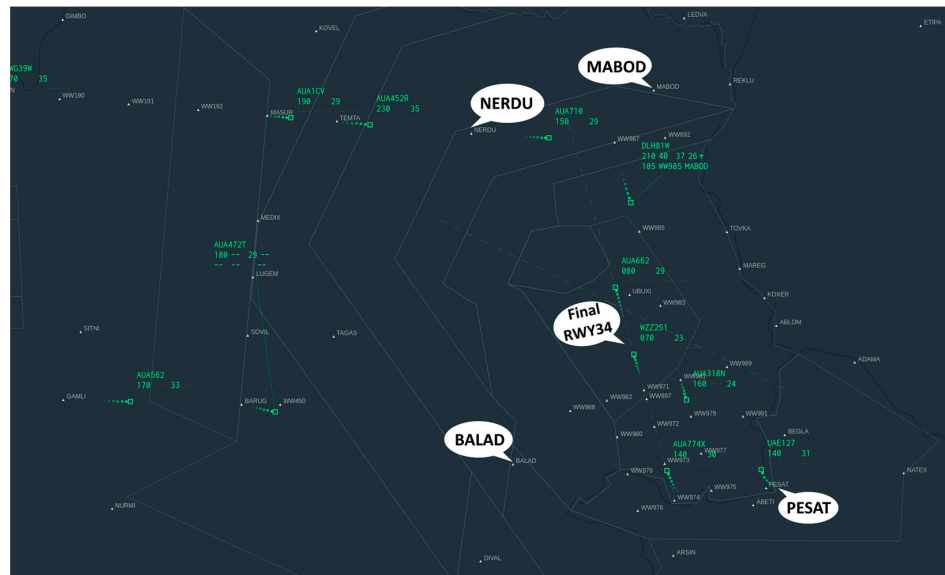


Figure 5. Approach chart of Vienna TMA with the four sectors BALAD, NERDU, PESAT, MABOD around the four metering fixes with the same names, taken from [5].

Simulation pilots managed flights and interacted with the ATCO via voice communication. Subjective feedback was gathered by means of questionnaires, debriefings and observations. Objective data regarding system performance were recorded (e.g., flown trajectory length and command recognition rates).

Two different scenarios were created: a medium-density traffic scenario with 30 arriving aircraft per hour and a heavy-density traffic scenario with 42 arriving aircraft per hour. A total of 12 ATCOs from Austro Control took part in the simulations lasting from September to November 2022.

In the baseline scenario, the ATCO was not supported by ASR, but was working and inputting the various commands manually using the current operating method consisting of mouse inputs. The ATCO had to click on one of the nine underlined data fields of the radar labels shown in Figure 6, taken from [5]. The click opens a drop-down menu and the ATCO needs to manually enter the given clearance values, e.g., for altitude, speed, heading, waypoint, etc.

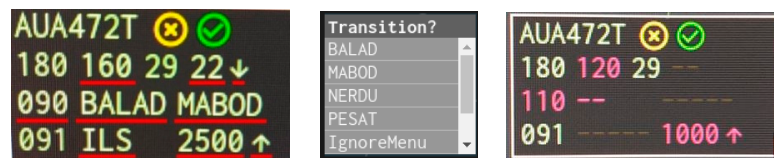


Figure 6. Left: Interactive radar label cells (red underlined); Middle: Drop-down menu to enter given transition names. Right: Radar label showing recognized command values (purple).

In solution runs, the values of the ATCO commands are extracted from the radio telephony utterance by ASR and are automatically input to the radar label cells appearing in purple color. The right part in Figure 6 shows the appearance when a flight level of 120, a heading of 110 and a descent rate of 1000 feet per minute or greater were extracted by the ASR. The dotted line “—” in waypoint field means that a recognized heading value overwrites a previously recognized waypoint value. Thus, the ATCO only needs to check

and confirm the automatically generated input with a mouse click on the green checkmark in the first radar label line, or alternatively correct any values in cases of misrecognition. Accepted cell values turned into light green as soon as the ATCO accepted them. The command values are automatically accepted after ten seconds, if the ATCO does not reject or correct them. More details with respect to the HMI design used are described in [5].

Each ATCO participated in four simulation runs of 35 min duration each. Two runs were conducted in baseline mode and two in solution mode with ASR support, so that all combinations of heavy and medium traffic with baseline and solution modes were simulated by each ATCO. To compensate for sequence effects, i.e., training effects, five ATCOs started with the baseline run. In addition, seven ATCOs started with the solution runs. After the baseline run, two solution runs followed or two baseline runs followed, if the ATCO started with the solution runs. Nevertheless, there were sequence effects. A technique to compensate for the sequence effects by subtracting or adding the mean value of all first, second, third and fourth simulation runs was implemented [5].

After each simulation run, the ATCOs filled out several questionnaires, and after the last validation exercise the ATCOs completed an additional final questionnaire. An informal semi-structured debriefing with the ATCOs followed the final validation simulation runs. Table 1 shows the 10 questions, which are safety related and taken from [5].

Table 1. Questions gathering feedback related to safety issues.

Question ID	Content
1	How insecure, discouraged, irritated, stressed, and annoyed were you? (Stress annoyed)
2	What was your peak workload? (Peak workload)
3	In the previous run I . . . started to focus on a single problem or a specific aircraft. (Single aircraft)
4	In the previous run there . . . was a risk of forgetting something important (such as inputting the spoken command values into the labels). (Risk to Forget)
5	In the previous run, how much effort did it take to evaluate conflict resolution options against the traffic situation and conditions? (Conflict resolution)
6	In the previous run, how much effort did it take to evaluate the consequences of a plan? (Consequences)
7	In the previous working period, I felt that . . . the system was reliable. (Reliable)
8	In the previous working period, I felt that . . . I was confident when working with the system. (Confidence)
9	I . . . found the system unnecessarily complex. (Complexity)
10	Please read the descriptors and score your overall level of user acceptance experienced during the run. Please check the appropriate number. (User Acceptance)

The results from the two validation exercises with focus on safety are presented in the next section.

4. Results

The first part of this section provides the hazards derived from the assessment and the requirements to mitigate the hazards. The second part describes two validation activities conducted to demonstrate the completeness of the design.

4.1. Hazard, Severity and Corresponding Design Requirements

A total of eight functional hazards (FHz) were identified based on the analysis of the use cases.

1. FHz#01: Significant delay in ASR callsign/command recognition and/or display (relevant for use cases 3 and 4)
2. FHz#02: ASR fails to identify an aircraft callsign from pilot's utterance, i.e., no aircraft is highlighted (relevant for use case 1)
3. FHz#03: ASR fails to identify an aircraft callsign from controller's utterance, i.e., no aircraft is highlighted (relevant for use case 2)
4. FHz#04: ASR erroneously identifies an aircraft callsign from pilot's utterance, i.e., the wrong aircraft is highlighted (relevant for use case 1)
5. FHz#05: ASR erroneously identifies an aircraft callsign from controller's utterance, i.e., the wrong aircraft is highlighted (relevant for use case 2)
6. FHz#06: ASR fails to identify a command from controller's utterance, i.e., no given command is shown to the ATCO (relevant for use cases 3, 4)
7. FHz#07: ASR erroneously identifies a command from controller's utterance, i.e., a wrong command or a command never given is shown to the ATCO (relevant for use cases 3, 4)
8. FHz#08: ASR recognizes an incorrect aircraft callsign, and the (correct or wrong) command is displayed for the incorrect flight in the CWP HMI (relevant for all use cases)

Safety assessment requires that the operational effects of identified hazards are classified in accordance with a Risk Classification Scheme (RSC) based on the severity of the operational effect the hazard may trigger [22,26]. The RSC classifies the hazards and provides the safety target, i.e., the maximum tolerable frequency (MToF) for each hazard's occurrence per flight hour in a specific unit.

- Severity Class 1: Accidents (max safety target with a probability of less than 10^{-9} , i.e., one catastrophic accident per one billion flight hours attributable to ATM).
- Severity Class 2: Serious Incidents (max safety target with a probability of less than 10^{-6})
- Severity Class 3: Major Incidents (max safety target with probability of less than 10^{-5})
- Severity Class 4: Significant Incidents (max safety target with a probability of less than 10^{-3})
- Severity Class 5: No Immediate Effect on Safety (no target).

Table 2 provides a list of the identified hazards with their causes, the assessed operational effect, and mitigations considered in defining the operational effect protecting against the functional hazard's effects propagation. The severity classification for each hazard was derived during a workshop with three ATCOs, concept designers and safety experts. The severity of the hazard determines the tolerable frequency of hazard occurrence.

As demonstrated in Table 1, based on the discussion with ATCOs participating in the session, it was recognized that the impacts of the ASR functional hazards are not significant from a safety perspective according to the Risk Classification Scheme, (RCS) [26]. Therefore, the requirements set as a mitigation do not require the safety target as such derived from the RCS and can be derived from operational needs, ensuring performance acceptable for ATCOs and ensuring no degradation in the execution of the ATCOs' tasks.

The impact of never attaining a perfect ASR may lead to situations, which are also present in the current operating method and working procedures impacting human performance negatively, i.e., increased workload and decreased situational awareness. With the support of various tools already used in current operations (such as monitoring aids), these events will in the current mode and with ASR support not escalate to safety relevant events.

Table 2. Hazards' causes, operational effect, possible mitigation and severity of hazards derived during Functional Hazard Assessment workshop.

Functionality Hazard & Severity	Potential Causes & Operational Effect	Mitigations Protecting against Propagation of Effects
<p>FHz#01 Significant delay in ASR callsign/command recognition and/or display</p> <p>No Immediate Effect on Safety</p>	<p>-Design issue -ASR provides delayed output</p> <p>If the use of ASR introduces delays in the usage of speech information (display of inputs, identification of aircraft, etc.) this may cause the ATCOs to focus on specific flight/area of the Area of Responsibility, until they can verify that the action induced by ASR has been correctly processed and displayed. This may have a negative impact on ATCO situational awareness.</p>	<p>Contingency measure to switch off ASR.</p>
<p>FHz#02 ASR fails to identify an aircraft from pilot's utterance—no aircraft is highlighted</p> <p>No Immediate Effect on Safety</p>	<p>-Pilot utters a non-understandable callsign or noise environment -Pilot utters a legal and understandable callsign, but ASR fails to recognize it.</p> <p>If the pilot performs the radio call and the flight is not highlighted, ATCO may have to scan the area of responsibility (AoR) to locate the aircraft. However, if ASR functionality to highlight the callsign is defined in the new operating method, there is a default expectation by the ATCO that it is functional and assisting in locating aircraft, resulting in minor workload increase and situational awareness reduction.</p>	<p>ATCO may have to scan the Area of Responsibility to locate the aircraft. No difference to current operating method.</p>
<p>FHz#03 ASR fails to identify an aircraft from controller's utterance—no aircraft is highlighted</p> <p>No Immediate Effect on Safety</p>	<p>-ATCO utters an illegal/non-understandable callsign -ATCO utters a legal and understandable callsign, but ASR fails to recognize it.</p> <p>If ATCO performs the radio call, it is assumed the impact is minor, because the ATCO's attention is focused on the aircraft being called and the impact is negligible.</p>	<p>No difference to current operating method.</p>
<p>FHz#04 ASR erroneously identifies an aircraft from pilot's utterance—wrong aircraft is highlighted</p> <p>No Immediate Effect on Safety</p>	<p>-If pilot performs the radio call and erroneous flight is highlighted, the ATCO may focus on the highlighted aircraft and issue the clearance intended for the calling aircraft to the wrong flight. The difference to the current operating method is that while occasional callsign confusion may occur between similar callsigns, now the ASR system is enforcing the ATCO's perception of issuing the clearance to what is expected to be the correct flight. If the confusion is not clarified through read-back and hear-back procedure or with the assistance of the planning controller, issued clearance to the wrongly highlighted aircraft may result in an unintended trajectory change. From a safety perspective, this is not significantly different from the current operating method, when ATCO enters a clearance into the radar label for the wrong callsign.</p>	<p>If the confidence level of the callsign recognition is not sufficiently high, it is not highlighted. For lower confidence levels to highlight with different color to emphasize the uncertainty of correct recognition. If the erroneous recognition persists, ATCO switches off the ASR and continues working as in today's operations.</p>
<p>FHz#05 ASR erroneously identifies an aircraft from controller's utterance—wrong aircraft is highlighted.</p> <p>No Immediate Effect on Safety</p>	<p>If ATCO performs the radio call and erroneous flight is highlighted, it is assumed the impact is minor as the ATCO's attention is on the aircraft being called and the impact of erroneous highlight is negligible.</p>	<p>If the confidence level of the callsign recognition is not sufficiently high, it is not highlighted. For lower confidence levels to highlight with different color to emphasize the uncertainty of correct recognition.</p>
<p>FHz#06 ASR fails to identify a command from controller's utterance.</p> <p>No Immediate Effect on Safety</p>	<p>-ATCO utters an illegal/non-understandable command -ATCO utters a legal and understandable command, but ASR fails to recognize it The failure of ASR to identify a complete command force ATCO to manually make the input. In such cases the negative impact on ATCO workload and situational awareness is expected, as in the new operating method there is a default expectation by the ATCO that ASR is functional and assisting in inputting commands in the labels.</p>	<p>ATCO inputs command manually. If the failure of ASR to recognize commands persists, ATCO switches off the ASR and continues working as in today's operations.</p>

Table 2. Cont.

Functionality Hazard & Severity	Potential Causes & Operational Effect	Mitigations Protecting against Propagation of Effects
<p>FHz#07 ASR erroneously identifies a command from controller's utterance</p> <p>No Immediate Effect on Safety</p>	<p>-ATCO utters an illegal/non-understandable command -ATCO utters a legal and understandable command, but ASR recognizes the incorrect command, and wrong command is displayed in the CWP HMI.</p> <p>In cases where inputs are provided but are erroneous, the ATCO will have to recognize the error and change information already put into the system.</p> <p>Depending on the ATM system, some parts of correcting the clearance, route change, etc., may require manipulation of the flight plan route data to input the correction. In such cases the impact on ATCO workload and potential disruption to the ATCO workflow may be higher than in cases where only missing data need to be input to complete the clearance.</p>	<p>ATCO corrects ASR input manually for the intended callsign.</p> <p>If the failure of ASR to recognize commands correctly persists, ATCO switches off the ASR and continues working as in today's operations.</p>
<p>FHz#08 ASR recognizes an incorrect callsign, and the command is displayed for the incorrect flight in the CWP HMI</p> <p>No Immediate Effect on Safety</p>	<p>ATCO utters a legal and understandable command, but ASR recognizes an incorrect callsign, which is being considered by the system, and the command is displayed for the incorrect flight in the CWP HMI</p> <p>In cases where inputs are provided but are erroneous (i.e., command input for wrong aircraft), the ATCO will have to recognize the error and change information already put into the system.</p> <p>If the error is not recognized by the controller, the contacted pilot will nevertheless follow the clearance issued by controller on the frequency. The erroneous input in the label of another aircraft will soon be detected by clearance monitoring aids.</p>	<p>If ATCO recognizes the error, he/she rejects and either repeats the clearance or inputs it manually directly into the label of the correct aircraft radar label. If ATCO does not recognize the error, monitoring aids will detect the discrepancy between the flown trajectory and the command inserted in the label of the erroneous aircraft.</p>

Considering that the hazards identified did not directly impact safety, the requirements for design are derived from operational and functional needs. The following list of requirements was, therefore, developed as preventive mitigations for functional hazards. The hazards were further analysed via top-down and bottom-up techniques with the support of fault trees to identify all possible causes for the hazards to occur, and to limit the propagation of the effects of hazards. The details of the derivation for the hazard are presented in Appendix A: Top-down analysis and Appendix B: Bottom-up analysis. The full list of the requirements can be found in the CORDIS portal of the European Commission [27]. To facilitate understanding of the requirements, they are listed here divided into two subcategories: those related to callsign recognition (as a mitigation for hazards FHz#02, FHz#03, FHz#04, FHz#05 and FHz#08) and those related to command recognition (mitigation to hazards FHz#01, FHz#02, FHz#06, FHz#07 and FHz#08):

4.1.1. Safety and Performance Requirements Concerning Callsign

ASR should send a recognized callsign to the cooperating ATC system, no later than one second after the ATCO has ended the radio transmission.

- For 99.9% of the ATCO utterances (except callsign), the system shall be able to give the output in less than 2 s after the ATCO ended the radio transmission.
- If the confidence level of the callsign recognition is not sufficiently high, it shall not be highlighted. Confidence level corresponds to a plausibility value derived by ASR. If the plausibility value is below a given threshold, the callsign is set to 'not recognized'.
- The HMI shall highlight the track label or part of it (or the track symbol) after recognizing the corresponding callsign.

4.1.2. Safety and Performance Requirements Concerning Commands

- The ASR shall recognize commands of different command categories (such as descend, reduce, heading).
- The Command Recognition Rate of ASR for ATCOs should be higher than 85%.
- The Command Recognition Error Rate of ASR should be less than 2.5% for ATCOs.
- The Command Recognition Error Rate of ASR should be less than 5% for pilots.

- The HMI should present the recognized (and validated) command types together with the command values in the radar label.
- The HMI shall enable manual correction/update of automatically proposed command value/type.
- The ASR system shall have no significant differences in the recognition rates of different command types, if the command types are often used (e.g., more than 1% of the time).

The set of proposed requirements satisfying “failure approach” enables achieving sufficient assurances for safety in the design of the system. The next step of the assessment consisted of demonstrating the evidence for the mitigation of each hazard and achievability of the safety requirements in the validation activities: real time, (human-in-the-loop simulations described in Sections 3.3 and 3.4) conducted in operationally representative environment. The evidence for each hazard was collected via objective metrics and subjective feedback from the ATCOs as shown in Table 3.

Table 3. Specific metrics and measures for collecting evidence from demonstration activities.

Hazard	Objective Metrics	Subjective Feedback on the Statement
FHz#01 Significant delay in ASR command recognition and/or display (all use cases).	Timeliness (processing time)	Applicable to all hazards
FHz#02: ASR fails to identify an aircraft from pilot’s utterance—no aircraft is highlighted (use case 1)	Pilot’s callsign recognition rate (no callsign highlighted)	The accuracy of the information provided by the ASR system is adequate for the accomplishment of operations.
FHz#04: ASR erroneously identifies an aircraft from pilot’s utterance—wrong aircraft is highlighted (use case 1).	Pilot’s callsign recognition error rate	Command Recognition Error Rate stays in the acceptable limits.
FHz#03: ASR fails to identify an aircraft from controller’s utterance—no aircraft is highlighted (use case 2).	Controller’s callsign recognition rate (no callsign highlighted) Controller’s callsign recognition error rate (wrong callsign highlighted)	The number and/or severity of errors resulting from the introduction of the ASR system is within tolerable limits, considering error type and operational impact.
FHz#05: ASR erroneously identifies an aircraft from controller’s utterance—wrong aircraft is highlighted (use case 2).	Controller’s callsign recognition rate (no callsign highlighted) Controller’s callsign recognition error rate (wrong callsign highlighted)	The level of ATCO’s situational awareness is not reduced with the introduction of the ASR system (ATCO is able to perceive and interpret task relevant information and anticipate future events/actions).
FHz#06: ASR fails to identify a command from controller’s utterance (use case 3, 4).	Controller’s command recognition rate	The level of ATCO’s workload is maintained or decreased with the introduction of the ASR system.
FHz#07: ASR erroneously identifies a command from controller’s utterance (use case 3 and 4)	Controller’s command recognition error rate	The number and/or severity of errors resulting from the introduction of the ASR system is within tolerable limits, considering error type and operational impact.
FHz#08: ASR recognizes an incorrect callsign, and the command is displayed for the incorrect flight in the CWP HMI (use case 3, 4)	Controller’s callsign recognition error rate	

4.2. Results of the Validation Activities

The results of the two real-time human-in-the-loop simulations described in Sections 3.3 and 3.4 are presented in the following two subsections.

4.2.1. Validation Activity “Callsign Highlighting”

Evidence Based on the Objective Metrics

Table 4 presents the total number of callsigns present in the simulation audio logs and the number of callsigns that were correctly detected by the ASR. The percentage of correctly detected callsigns is higher for ATCOs than for flight crew in both cases as the algorithm is optimized for the ATCO locutions.

Table 4. Callsign recognition rates for ATCOs and Flight Crew.

Analysis Type	ATCO			Flight Crew		
	N° of Callsigns	N° Callsigns Detected	Percentage	N° of Callsigns	N° Callsigns Detected	Percentage
RTS recordings	859	721	84%	457	687	67%
Operational recordings	143	127	87%	158	77	49%

Regarding the comparison between simulation and operational recordings, the percentage for ATCOs are similar but the percentage for flight crew is better in the simulation. This was already expected as the quality of the recording (signal-to-noise ratio) is better in the simulation and the accent (mother tongue) of the simulation pilots is unique (Spanish), while the one from the operational recordings is very diverse with 29 airlines from 18 different countries.

No callsign was wrongly recognized as only complete callsigns were detected. Feedback from ATCOs indicated that they would like to have higher recognition rates even if some callsigns were incorrectly detected and highlighted. The error allowance is something to be further investigated in follow-up research.

Table 5 presents the number of commands that were present/detected and the callsign + command that were correctly detected for each analysis. Only commands that fall within the five categories, for which the prototype was optimized, are presented. There were several other commands that ATCOs used during the simulation such as squawk change, standard terminal arrival route (STAR) assignment, and information (traffic information, barometric pressure setting). In the first column “Only Command Recognition”, only the type of the command to classify whether the command was detected or not are considered. In the last three columns “Callsign + Command Recognition”, the command type plus the callsign must be correctly extracted to count as a detected command. The results show that the lower callsign recognition rate resulting from the very low callsign error rate also results in a lower command recognition rate, if both callsign and command type must be correct.

Table 5. Detected command types, when considering only command type.

Analysis Type	Only Command Recognition			Callsign + Command Recognition		
	Commands	Detected Commands	%	Commands	Detected Callsign + Commands	%
RTS recordings	695	619	89%	695	523	75%
Operational recordings	182	167	92%	182	146	80%

The performance for operational data is better than for RTS recordings. There is a 3% difference between the RTS and the operational recognition percentages regarding command recognition, and a 5% difference in callsign and command recognition percentages. During the exercise, the participating ATCOs were encouraged to test the recognition system. They thus issued longer, and more complex authorizations than usually issued in operational environments. This together with the fact that the ASR prototype was trained and optimized using operational communications explains the difference between both percentages.

If not only the command type, but also the information contained in the command are considered (i.e., values, units, qualifiers), the extraction performance is lower, as shown in Table 6.

Table 6. Detected commands, i.e., command, when callsign plus command information, i.e., values, units, qualifiers are considered.

Analysis Type	Complete Command Recognition			Callsign + Complete Command Recognition		
	Commands	Detected Commands	%	Commands	Detected Commands	%
RTS recordings	695	498	72%	695	416	60%

Evidence Based on Subjective Feedback

Accuracy of ASR was collected through tailor-made questionnaires, debriefings, and data logs. The ATCO feedback was that the tool that needed improvement in the recognition rates to be able to effectively support them in the execution of their tasks. ATCOs indicated that they would prefer some occasional false positive callsign recognized if that would mean higher recognition rates.

Timeliness was collected through tailor-made questionnaires, debriefings, and data logs. Data logs indicated that when the callsigns were located at the end of an utterance the radar track and command information was presented in 0.9 s, but when it was located at the beginning of a sentence it took up to 3.0 s. Controllers subjective feedback indicated that timeliness was rated as adequate for the callsigns at the end of the utterance but inadequate when the callsign was at the beginning of the utterance.

ATCOs' situational awareness, measured with SASHA [28], slightly improved with the use of ASR (score 4.0 in the reference questionnaire and 4.4 in the solution scenario). During the debriefings, ATCOs stated that situational awareness was improved but they considered that the ASR recognition rate was not high enough to allow them to completely confide and exploit the tool. They consider that higher callsign recognition rates would further improve their situational awareness.

ATCO workload was collected through Nasa-TLX [29] questionnaire, tailor-made questionnaires, and debriefings. The Nasa-TLX scored 9.1 (out of 20) for the baseline scenario and 7.9 (out of 20) for the solution scenario questionnaire. The tailor-made questionnaire and debriefings indicated that workload slightly decreased in the solution scenario.

4.2.2. Validation Activity 2: Radar Label Maintenance

Evidence Based on the Objective Metrics

The validation activities were performed between September 2022 and November 2022 as described in Section 3.4.

Table 7 provides the speech recognition and understanding performance taken from [5]. A word error rate (WER) of 3.1% was achieved, i.e., only every 33rd word was wrongly recognized. This is extremely good considering that humans usually achieve a WER of 4 to 11%, depending on the noise level and the option to listen more than once [30].

Table 7. Performance at the semantic level quantified as recognition and error rates.

Level of Evaluation	WER	Cmd-Recog-Rate	Cmd-Error-Rate	Csgn-Recog-Rate	Csgn-Error-Rate
Full Command	3.1%	92.1%	2.8%	97.8%	0.6%
Only Label		92.5%	2.4%		

These results are based on 118,800 manually transcribed words resulting in 17,100 commands from 8850 utterances. The word error rate of the used speech recognizer is 3.1%.

Out of all the given commands, 92.1% were recognized and 2.8% were wrongly recognized. The difference to 100% means rejections, i.e., nothing was recognized for this command. A command is only correctly recognized, if the callsign of the command, the command type (descend, reduce, heading. . .), the values, the unit, the qualifier (left, right, etc.) and the conditions are all correct. Therefore, the callsign recognition rate (column "Csgn-Recog-Rate") is always better than the recognition rate of the total command. A

callsign error rate of 0.6% in the last column corresponds to only one of 165 callsigns being wrongly recognized. The last row “Only Label” shows the results when we do not consider all 17,100 recognized commands, but only the 12,600 commands which are also shown in the radar label cells; e.g., a QNH or a squawk command are not shown in the radar label.

A recognition rate of 92.5% means that 7.5% of the given commands were not correctly pre-filled by the ASR support functionality, i.e., the remaining 7.5% of the commands need to be manually input by the ATCOs [5]. Only 50% of them were manually corrected or inputted, respectively. Out of the 6400 commands given in the solution runs, 219, i.e., 3.4% remained incorrect in the radar labels [5]. One can then pose the question whether safety has now decreased, when pre-filling of radar label cells is supported by speech recognition? It can be argued that no, the contrary is the case. Helmke et al. [5] also showed that in the baseline runs without ASR support, 617 of the given 6320 radar cell label relevant commands were not correct or remained missing in the radar label cells, i.e., 11.6% versus 3.4%.

Evidence Based on the Subjective Metrics

Table 8 shows the mean values of the normalized answer differences of the 12 ATCOs after having compensated for sequences effects. The answers were scaled into the interval [1..10]; 1 meaning very good performance and 10 meaning bad performance. Negative values in column “Diff” mean that the ATCO judged the safety aspect relevant to this question higher with ASR than without ASR. The *p*-value is the statistical significance of a performed *t*-test. The cells are shaded in green for $0\% \leq p\text{-value} < 5\%$, in light green for $5\% \leq p\text{-value} < 10\%$, and in yellow for no real evidence (absolute *p*-value $\geq 10\%$). There were no single cases which would have provided evidence that working without ASR is safer than with ASR, i.e., $-10\% \leq p\text{-value} < 0\%$ was not measured; see [5] for more details.

Table 8. Subjective feedback of ATCOs to safety-related questions.

Question	Diff	<i>p</i> -Value
Stress annoyed	−0.16	34%
Peak workload	−0.32	9.9%
Single aircraft	0.04	−41%
Risk to forget	−0.64	0.7%
Conflict resolution	−0.26	24%
Consequences	0.30	−21%
Reliable	−0.24	30%
Confidence	−1.59	1.1%
Complexity	−1.98	2.0×10^{-4}
User Acceptance	−1.01	6.3%
Total	−0.56	0.4%

It should be noted that the performance achieved was higher than the minimum performance as defined by the safety/performance requirements [27]; the command recognition rate was expected to be higher than 85%, whereas 92.5% was achieved. The command Recognition Error Rate of ASR showed slightly better performance with 2.4% against 2.5% set by the requirements.

The callsign recognition rate and error rate, although not quantified by the requirements, showed high performance. The subjective data based on the SHAPE Automation Trust Index (SATI) [28] questionnaire confirmed that the level and quality of information provided by the system (as displayed in the radar labels) was acceptable with an average score of 8.8 on a scale from 1 to 10, i.e., 10 indicating the best rating option. With the support of ASR, the value was 0.8 units better than without ASR support.

Timeliness of the information provided by the ASR.

The design recommendation set an expectation of 99.9% for the ATCO utterances, except the callsign, being available in less than two seconds after the ATCO has released the push-to-talk button.

The ATCO subjective feedback demonstrated that the timeliness of ASR output in the aircraft radar labels was considered adequate with an average score of 8.5 on a scale from 1 to 10, i.e., 10 indicating the best rating option.

Number and type (nature) of human errors.

ATCOs confirmed that ASR did not increase the potential for human errors with an average score of 3 on a scale from 1 to 10, i.e., 10 indicating the worst rating option.

An objective analysis actually confirms that the number of errors in the radar label cells, i.e., missing input, is much less if ATCOs are supported by ASR compared to entering everything manually with a mouse ($\alpha < 10^{-7}\%$).

Level of ATCO's situational awareness.

ATCOs confirmed that their situational awareness is maintained at an acceptable level with ASR with an average score of 8.9 on a scale from 1 to 10, i.e., 10 indicating the best rating option.

ATCO's workload.

The secondary validation objective regarding objective workload measurement showed a statistically significant decrease (p -value = 0.3%) in the workload when ATCOs are supported by ASR. The ATCO-self-rated Instantaneous Self-Assessment (ISA) [31] score confirmed this with the same statistical significance (p -value = 3.1%, see Table IX in [5]). ATCOs confirmed that ASR supported them in maintaining the workload at an acceptable level with an average score of 7.9 on a scale from 1 to 10, i.e., 10 indicating the best rating option. More meaningful, however, is the clicking time. In all baseline runs together, the ATCOs need 12,700 s for maintaining the radar label contents. In the solution runs with ASR support, only 405 s were needed, i.e., an improvement by a factor 31.

4.3. Limitations

The safety assessment performed as part of the SESAR2020 PJ.10-W2-96.2 ASR research and development activities covers specific use cases in specific ATC environments and the extrapolation of the results of the safety assessment may, therefore, not be applicable to all operational environments and other ASR applications. The safety analysis described in this paper focuses on the generic application of ASR technology in the pre-industrial phase. Thus, the research results achieved may not be fully transferable to live operations and any local implementation requires further investigation to satisfy the safety requirements as defined in relevant regulation and/or the local competent authority. The results achieved in these validations did not contain a long-term assessment of ASR functionality and potential impacts to safety thereof. Likewise, the impact of external factors such as background noise, various ATCO accents, and radio transmission quality and interference were not assessed beyond their possible occurrence in the research environment. Implementation of ASR capabilities may introduce the ATM system to new cyber security vulnerabilities which would need to be evaluated through a local cyber security assessment. Further research may be required for the implementation of ASR in different operational environments with different traffic demands and complexity characteristics in ATC facilities applying different ATM platforms. If ASR is used to supplement other safety tools present in an ATM system (e.g., conformance monitoring, conflict detection) as suggested by some studies [17–19], the safety considerations of ASR may require a detailed assessment of the various system components' interactions as opposed to the comparison between current and new operating methods focused solely on availability of ASR as presented in this paper. Furthermore, the acceptance of ASR by ATCOs—and subsequent impact on workload and situational awareness—may vary between different organizational cultures and a holistic assessment of ASR suitability to a specific environment would be required.

5. Discussion of Results

Overall, two different approaches for callsign extraction from spoken utterances have been validated. The first one emphasized a very low error rate. The prize is a lower callsign recognition rate. As a result, ATCOs reported preference for higher recognition rates with

an occasional false positive call sign. This was addressed by the second approach trying a compromise between low error rate and recognition rate. The analysis shows that it is up to the user, i.e., ATCO, to decide what is preferred on a daily basis. An error rate of 0% will not be possible, if a reasonable recognition rate is needed. Human actors do not achieve an error rate of 0% either.

In general, ATCOs confirmed that they are able to perform their ATC tasks when working with ASR support. The positive results achieved in situational awareness and workload measurements in both validation exercises indicate the potential for further benefits in ATCO performance in an operational environment. Timeliness of ASR output in the first validation—radiotelephony utterances with aircraft call sign at the end only—was found acceptable. In the second validation the timeliness was found to be adequate. ATCOs from both validations confirmed that the application of ASR did not introduce any additional risks for errors.

A recognition rate of 92.5% is still far away from 100%. It means, however, that the time spent by the ATCOs manually updating the radar labels with clearance information could be reduced by a factor of 31 from 12,700 s, i.e., 25% of the total simulation time, without ASR support down to 405 s with ASR support. These numbers are based on a very heavy traffic scenario, in which ATCO plus simulation pilots blocked the frequency for 70% of the time.

The safety and hazard analysis of this research work has shown that no severe hazards exist, when using ASR applications as call sign highlighting or pre-filling radar labels in an operational environment. In heavy traffic scenarios, 3.4% of the given commands are not correctly entered into the radar label cells, when ASR support is available. Without ASR support the missing command rate increases to 11.6%. ASR support does not decrease safety, but rather increases safety, when ATCO and ASR work as a team.

Research, therefore, should not concentrate on increasing command recognition rate from 92.5% to 95% or to 99.9%, which is not expected to happen in the near future. Research needs to focus on attention guidance, which gives hints to the ATCO, when something might be wrong or missing in the radar labels. This can be done by integration of ASR with other assistant systems already available in the ops room. Comparing downlinked mode-S data with radar label cell entries can even further reduce the number of erroneous label value entries.

6. Conclusions

In this article, we focused on demonstrating the safety of ASR application in ATC operational environments.

The safety assessment showed that the eight ASR functional hazards have no significant effect on overall ATM safety. Mitigations were derived from operational needs, to ensure acceptable ATCO performance without degrading ATCO's task execution. A potential decrease in situational awareness or increase in workload in the case of insufficient ASR performance were already present in the current operating method, but can be further mitigated through the use of clearance monitoring tools to prevent the escalation of these events to safety relevant occurrences.

The requirements developed as part of the safety assessment for the application of ASR technology in ATM were achieved in operational environments reflecting real-life ATC centers for en-route and approach control. The technical system, in terms of accuracy and timeliness, outperformed expectations required by the design and associated targets. The subjective feedback of ATCOs from two different validation setups was encouraging and confirmed that ASR application not only generated benefits, but also showed to be feasible for implementation in currently deployed ATM systems.

Author Contributions: Conceptualization, E.P.-C., J.D., P.H. and H.H.; data curation, H.H., R.G.L. and O.O.; formal analysis, E.P.-C., J.D. and H.H.; funding acquisition, P.H.; investigation, E.P.-C., J.D. and H.H.; methodology, E.P.-C. and J.D.; project administration, P.H.; resources, H.H. and O.O.; software, E.P.-C. and H.H.; supervision, P.H. and H.H.; validation, E.P.-C.; J.D., R.G.L. and P.H. visualization, E.P.-C. and O.O.; writing—original draft, E.P.-C., J.D. and H.H.; writing—review and editing: J.D., P.H., R.G.L. and O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by SESAR Joint Undertaking under grant agreement No. 874464 under European Union’s Horizon 2020 research and innovation programme.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of this study, in the collection, analyses, or interpretation of the data, in the writing of the manuscript, and in the decision to publish the results.

Appendix A. Top-Down Analysis

The hazards were further analyzed by top-down technique with the support of fault trees to identify all possible causes of the hazards to occur, and to limit the propagation of the effects of hazards. The details of the derivation for the hazard are presented in this appendix.

Cause ID (in Fault Tree)	Cause	Detailed Description	Mitigation/Safety Requirement
FHz#01 Significant delay in ASR callsign/command recognition and/or display	-ASR provides delayed output	One or more of the ASR components is not performing as expected and causing the delay in the ASR output display.	The ASR system should provide the functionality to be switched off and switched on when necessary. ASR should send a recognized callsign to the cooperating ATC system when the controller ends the radio transmission within a maximum of 1.0 s. For 99.9% of the ATCO utterances except callsign itself, the system shall be able to give the output in less than two seconds after the ATCO has released the push-to-talk button. The ASR system shall have no significant differences in the recognition rates of different command types, if the command types are not very seldom use (e.g., less than 1% of the time).
FHz#02 ASR fails to identify an aircraft from pilot’s utterance—no aircraft is highlighted	-Pilot utters a non-understandable callsign -Pilot utters a legal and understandable callsign, but ASR fails to recognize it	Pilot performs the radio call and the flight is not highlighted in the CWP HMI.	The HMI shall highlight the Track Label or part of it (or the track symbol) after recognizing the corresponding callsign. If the confidence level of the callsign recognition is not sufficiently high, it shall not be highlighted.
FHz#03 ASR fails to identify an aircraft from controller’s utterance—no aircraft is highlighted	-ATCO utters an illegal/non-understandable callsign -ATCO utters a legal and understandable callsign, but ASR fails to recognize it	ATCO performs the radio call and the flight is not highlighted in the CWP HMI.	The HMI shall highlight the Track Label or part of it (or the track symbol) after recognizing the corresponding callsign. If the confidence level of the callsign recognition is not sufficiently high, it shall not be highlighted.
FHz#04 ASR erroneously identifies an aircraft from pilot’s utterance—wrong aircraft is highlighted	-Pilot utters a non-understandable callsign Pilot utters a legal and understandable callsign, but ASR recognizes an existing wrong callsign -Pilot utters a legal and understandable callsign, but ASR recognizes a wrong callsign not matching to callsigns considered by the system	ATCO focuses on the highlighted aircraft and issues the clearance intended for the calling aircraft to the wrong flight. If the ATCO issues the clearance to the wrongly highlighted aircraft, it may result in an unintended trajectory change.	The HMI shall highlight the Track Label or part of it (or the track symbol) after recognizing the corresponding callsign. The Command Recognition Error Rate of ASR should be less than 5% for pilots. The Command Recognition Rate of ASR for pilots should be higher than 75%. If the confidence level of the callsign recognition is not sufficiently high, it shall not be highlighted.

Cause ID (in Fault Tree)	Cause	Detailed Description	Mitigation/Safety Requirement
FHz#05 ASR erroneously identifies an aircraft from controller's utterance—wrong aircraft is highlighted	-ATCO utters an illegal/non-understandable callsign -ATCO utters a legal and understandable callsign, but -ASR recognizes an existing wrong callsign -ATCO utters a legal and understandable callsign, but -ASR recognizes a wrong callsign not matching to callsigns considered by the system	ATCO may get confused and issue a wrong clearance. If the ATCO issues the clearance to the wrongly highlighted aircraft, it may result in an unintended trajectory change.	ATCOs will use standard phraseology as per ICAO Doc.4444 [32]. The ASR shall recognize commands of different command categories (such as descend, reduce, heading). The Command Recognition Error Rate of ASR should be less than 2.5% for ATCOs. The Command Recognition Rate of ASR of ATCOs should be higher than 85%. If the confidence level of the callsign recognition is not sufficiently high, it shall not be highlighted.
FHz#06 ASR fails to identify a command from controller's utterance	-ATCO utters an illegal/non-understandable command ATCO utters a legal and understandable command, but ASR fails to recognize it	ATCO manually makes the input resulting in workflow disruptions and workload increase and situational awareness reduction.	The HMI shall enable manual correction/update of automatically proposed command value/type. The HMI should present the recognized (and validated) command types together with the command values in the radar label. The ASR system should provide the functionality to be switched off and switched on when necessary. The Command Recognition Rate of ASR of ATCOs should be higher than 85%.
FHz#07 ASR erroneously identifies a command from controller's utterance	-ATCO utters an illegal/non-understandable command -ATCO utters a legal and understandable command, but ASR recognizes an incorrect callsign, which is being considered by the system, and the command is displayed for the incorrect flight in the CWP HMI -ATCO utters a legal and understandable command, but ASR recognizes the incorrect command, and wrong command is displayed in the CWP HMI	ATCO will have to change information already input into the system. Depending on the ATM system, some parts of correcting the clearance, route change, etc., may require manipulation of the FPL route data to input the correction.	The HMI shall enable manual correction/update of automatically proposed command value/type. The HMI associated with ASR shall enable the ATCO to reject recognized command values for pre-filling radar label values by clicking on a rejection button. The ASR system should provide the functionality to be switched off and switched on when necessary. The Command Recognition Error Rate of ASR should be less than 2.5% for ATCOs.
FHz#08 ASR recognizes an incorrect callsign, and the command is displayed for the incorrect flight in the CWP HMI	-ATCO utters an illegal/non-understandable callsign followed by a command ATCO utters a legal and understandable callsign and a command, but ASR recognizes the incorrect callsign	ATCO utters a legal and understandable command, but ASR recognizes an incorrect callsign, which is being considered by the system, and the command is displayed for the incorrect flight in the CWP HMI.	The Command Recognition Error Rate of ASR should be less than 2.5% for ATCOs. The Command Recognition Rate of ASR of ATCOs should be higher than 85%. ATCO is supported by the clearance monitoring aids as in today's operations.

Appendix B. Bottom-Up Analysis

In view of complementing the fault tree findings, the bottom-up analysis of the failure modes of the ASR system elements and element interfaces and of their effects was conducted to determine potential common cause failures and to allow a more in-depth causal analysis of certain parts of the functional system design. The details of the derivation for the hazard of the bottom-up analysis are presented in this appendix.

Technical System Element	Failure Mode	Effects	Mitigation/Safety Requirement
Command Prediction	<p>Fails to forecast possible future controller commands.</p> <p>Failure to receive external data required for forecast of future controller commands (external data can be radar data, flight plan data, weather data, airspace data, and also historic data of those types).</p>	The speech recognizer relies on the input of the predicted commands. Commands which are not predicted (normally) cannot be recognized. So, if command prediction accuracy is worse than recognition accuracy itself, the command prediction functionality might have no benefits for the recognition engine any more.	If ASR is used, the Command Prediction Error Rate should not be higher than 10% and also not be higher than 50% of the opposite command recognition rate (i.e., 100% minus the command recognition rate), without using a plausibility checker.
Recognize Voice Words	<p>Fails to analyze the voice flow and to transform into a text string.</p> <p>Does not receive the Voice Flow.</p>	No callsign or command is recognized by ASR and displayed on the CWP HMI.	If ASR does not provide an input, ATCO proceeds as in current operations (manual input and with no highlight of the callsign).
Apply Ontology and Logical check	<p>Fails to analyze the text string and to transform into a set of predefined commands to discard incoherent commands.</p>	The ASR output is erroneous and incoherent.	The ASR shall recognize commands of different command categories.

References

- Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. Reducing Controller Workload with Automatic Speech Recognition. In Proceedings of the 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 September 2016. [CrossRef]
- European Commission. Commission Implementing Regulation (EU) 2017/373 of 1 March 2017 Laying down Common Requirements for Providers of Air Traffic Management/Air Navigation Services and Other Air Traffic Management Network Functions and Their Oversight Repealing Regulation (EC) No 482/2008, Implementing Regulations (EU) No 1034/2011, (EU) No 1035/2011 and (EU) 2016/1377 and Amending Regulation (EU) No 677/2011. 2017. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0373> (accessed on 23 October 2023).
- SESAR. SESAR Safety Reference Materials Ed 4.1. 2019. Available online: [https://www.sesarju.eu/sites/default/files/documents/transversal/SESAR2020%20Safety%20Reference%20Material%20Ed%2000_04_01_1%20\(1_0\).pdf](https://www.sesarju.eu/sites/default/files/documents/transversal/SESAR2020%20Safety%20Reference%20Material%20Ed%2000_04_01_1%20(1_0).pdf) (accessed on 23 October 2023).
- García, R.; Albarrán, J.; Fabio, A.; Celorrio, F.; Pinto de Oliveira, C.; Bárcena, C. Automatic Flight Callsign Identification on a Controller Working Position: Real-Time Simulation and Analysis of Operational Recordings. *Aerospace* **2023**, *10*, 433. [CrossRef]
- Helmke, H.; Kleinert, M.; Ahrenhold, N.; Ehr, H.; Mühlhausen, T.; Ohneiser, O.; Klamert, L.; Motlicek, P.; Prasad, A.; Zuluaga-Gómez, J.; et al. Automatic Speech Recognition and Understanding for Radar Label Maintenance Support Increases Safety and Reduces Air Traffic Controllers' Workload. In Proceedings of the 15th USA/Europe Air Traffic Management Research and Development Seminar, ATM 2023, Savannah, GA, USA, 5–9 June 2023.
- Kleinert, M.; Helmke, H.; Moos, S.; Hlousek, P.; Windisch, C.; Ohneiser, O.; Ehr, H.; Labreuil, A. Reducing Controller Workload by Automatic Speech Recognition Assisted Radar Label Maintenance. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
- European Space Agency. Technology Readiness Levels Handbook for Space Applications. September 2008. TEC-SHS/5551/MG/ap. Available online: https://connectivity.esa.int/sites/default/files/TRL_Handbook.pdf (accessed on 23 October 2023).
- Santorini, R.; SESAR Digital Academy—Innovation in Airspace Utilization, 29 April 2021. SESAR Joint Undertaking | Automated Speech Recognition for Air Traffic Control. Available online: <https://www.sesarju.eu/node/3823> (accessed on 6 October 2023).
- Zuluaga-Gomez, J.; Nigmatulina, I.; Prasad, A.; Motlicek, P.; Khalil, D.; Madikeri, S.; Tart, A.; Szoke, I.; Lenders, V.; Rigault, M.; et al. Lessons Learned in Transcribing 5000 h of Air Traffic Control Communications for Robust Automatic Speech Understanding. *Aerospace* **2023**, *10*, 898. [CrossRef]
- Khalil, D.; Prasad, A.; Motlicek, P.; Zuluaga-Gomez, J.; Nigmatulina, I.; Madikeri, S.; Schuepbach, C. An Automatic Speaker Clustering Pipeline for the Air Traffic Communication Domain. *Aerospace* **2023**, *10*, 876. [CrossRef]
- Zuluaga-Gomez, J.; Prasad, A.; Nigmatulina, I.; Motlicek, P.; Kleinert, M. A Virtual Simulation-Pilot Agent for Training of Air Traffic Controllers. *Aerospace* **2023**, *10*, 490. [CrossRef]

12. Kleinert, M.; Helmke, H.; Shetty, S.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Harfmann, J. Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning. In Proceedings of the 40th Digital Avionics Systems Conference (DASC), Hybrid Conference, San Antonio, TX, USA, 3–7 October 2021.
13. Chen, S.; Kopald, H.D.; Chong, R.; Wei, Y.; Levonian, Z. Read back error detection using automatic speech recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
14. Lin, Y.; Deng, L.; Chen, Z.; Wu, X.; Zhang, J.; Yang, B. A Real-Time ATC Safety Monitoring Framework Using a Deep Learning Approach. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4572–4581. [[CrossRef](#)]
15. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. Increasing ATM Efficiency with Assistant Based Speech Recognition. In Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar (ATM2017), Seattle, WA, USA, 26–30 June 2017.
16. Kleinert, M.; Ohneiser, O.; Helmke, H.; Shetty, S.; Ehr, H.; Maier, M.; Schacht, S.; Wiese, H. Safety Aspects of Supporting Apron Controllers with Automatic Speech Recognition and Understanding Integrated into an Advanced Surface Movement Guidance and Control System. *Aerospace* **2023**, *10*, 596. [[CrossRef](#)]
17. Karlsson, J. Automatic Speech Recognition in Air Traffic Control: A Human Factors Perspective. In *NASA, Langley Research Center, Joint University Program for Air Transportation Research, 1989–1990*; NASA: Washington, DC, USA, 1990; pp. 9–13.
18. Lin, Y.; Ruan, M.; Cai, K.; Li, D.; Zeng, Z.; Li, F.; Yang, B. Identifying and managing risks of AI-driven operations: A case study of automatic speech recognition for improving air traffic safety. *Chin. J. Aeronaut.* **2023**, *36*, 366–386. [[CrossRef](#)]
19. Zhou, S.; Guo, D.; Hu, Y.; Lin, Y.; Yang, B. Data-driven traffic dynamic understanding and safety monitoring applications. In Proceedings of the 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology, Dali, China, 12–14 October 2022.
20. European Union Aviation Safety Agency. *EASA Artificial Intelligence Roadmap 2.0; Human-Centric Approach to AI in Aviation*; European Union Aviation Safety Agency: Cologne, Germany, 2023; Available online: <https://www.easa.europa.eu/ai> (accessed on 23 October 2023).
21. European Union Aviation Safety Agency. *EASA Concept Paper: Guidance for Level 1 & 2 Machine Learning Applications—Proposed Issue 02*, Cologne, Germany, 2023. Available online: <https://www.easa.europa.eu/en/downloads/137631/en> (accessed on 23 October 2023).
22. SESAR. *Guidance to Apply SESAR Safety Reference Material*, Ed. 3.1. 2018. Available online: <https://www.sesarju.eu/sites/default/files/documents/transversal/SESAR%202020%20-%20Guidance%20to%20Apply%20the%20SESAR%20%20Safety%20Reference%20Material.pdf> (accessed on 23 October 2023).
23. EUROCONTROL. *Safety Assessment Methodology Ed2.2*; EUROCONTROL: Brussels, Belgium, 2006.
24. Insignia. Available online: <https://insignia.enaire.es> (accessed on 28 February 2023).
25. SESAR. D4.1.100—PJ.10-W2-96 ASR-TRL6 Final TVALR—Part I. V 01.00.00; SESAR Joint Undertaking, Brussels, Belgium, May 2023. Available online: <https://cordis.europa.eu/project/id/874464/results> (accessed on 23 October 2023).
26. European Organization for Civil Aviation Equipment. *EUROCAE ED-125, Process for Specifying risk Classification Scheme and Deriving Safety Objectives in ATM*; EUROCAE: Malakoff, France, 2010.
27. SESAR. D4.1.020—PJ.10-W2-96 ASR-TRL6 Final TS/IRS—Part I. V 01.00.00; SESAR Joint Undertaking, Brussels, Belgium, May 2023. Available online: <https://cordis.europa.eu/project/id/874464/results> (accessed on 23 October 2023).
28. Dehn, D.M. Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. *Air Traffic Control Q.* **2008**, *16*, 127–146. [[CrossRef](#)]
29. Hart, S. NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA, USA, 16–20 October 2006; pp. 904–908.
30. Stolcke, A.; Droppo, J. Comparing Human and Machine Errors in Conversational Speech Transcription. In Proceedings of the Proc. Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 137–141. Available online: https://www.isca-speech.org/archive/interspeech_2017/stolcke17_interspeech.html (accessed on 23 October 2023).
31. Jordan, C.S.; Brennen, S.D. *Instantaneous Self-Assessment of Workload Technique (ISA)*; Defence Research Agency: Portsmouth, UK, 1992.
32. ICAO. *Procedures for Air Navigation Services (PANS)—Air Traffic Management Doc 4444*, 16th ed.; ICAO: Montreal, QC, Canada, 2016.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.