ASSISTANT BASED SPEECH RECOGNITION – ANOTHER PAIR OF EYES FOR THE ARRIVAL MANAGER

Hejar Gürlük, Hartmut Helmke, Matthias Wies, Heiko Ehr, Matthias Kleinert, Thorsten Mühlhausen, Kathleen Muth, Oliver Ohneiser
German Aerospace Center, Lilienthalplatz 7, 38108 Braunschweig, Germany

Abstract

Nowadays Arrival Managers (AMAN) are available to produce efficient inbound traffic sequences and to create guidance advisories for optimized approaches. Information about deviations from the planned sequence is exchanged between controller and pilot via radio communication. The AMAN is only able to derive these deviations from the radar data. Using radar data as single input sensor, however, results in adaptation delays of 30 seconds and more – and even worse, the controllers' intent is still missing.

The AcListant® AMAN (Active Listening Assistant) [1] has shown for the Dusseldorf Approach Area how to avoid this sensor delay by analyzing the controller-pilot-communication and using the gained information as an additional sensor. An Assistant Based Speech Recognition system (ABSR) is embedded in an AMAN, which provides a dynamic minimized world model to the speech recognizer.

Validation trials were performed from February to March 2015 with seven male and four female air traffic controllers from Dusseldorf, Frankfurt, Munich, Vienna, and Prague. Depending on the accepted rejection rate of the speech recognizer, recognition rates between 90% and 95% were achieved, whereas without ABSR only rates between 58% and 83% were possible. Furthermore ABSR significantly reduces the deviation between the controllers' plan and the plan of the AMAN and, at the same time, significantly reduces the controllers workload.

Introduction

Conversation is a core element of society concerning its further development since centuries. Hence, a significant part of human collaboration is coordinated via voice, especially when complex contexts or meta-concepts are considered. By tracing communication, new actors can get an idea of the actual and planned situations and interpret the actions, so that they can easily integrate themselves into this environment. Listening actors can follow the conversation and contribute to problem solving.

Nowadays, people get more and more supported by technical systems like assistant or decision support systems which can be found in nearly every working and leisure environment. Latest applications, such as those by Apple (Siri®) [2], and Google (Voice Search®) [3] use Automatic Speech Recognition (ASR) as input interface for a direct communication between human and machine to trigger a defined action, as for instance a query. Speech recognition is used in most medium-sized cars [4]. However, all these systems still require improvements regarding their recognition rate.

In an Air Traffic Control (ATC) working environment, communication between the involved parties is the most important mean to control the air traffic. Controlling aircraft in the vicinity of an airport is an example of such a working environment in which two working groups communicate, i.e. pilots and controllers. All pilots in the same sector are supported by a dedicated controller (team). They use a unique frequency for communication within this sector. This enables a party line effect, i.e. all actors – excluding today's assistant systems – can create a common mental model of the current situation and of future actions.

Today ATC communication is still split into two different worlds: one in which humans communicate via radio links, and another in which machines communicate via computer networks. These two worlds are connected by a human machine interface used by humans to inform the machines and vice versa. Intents and plans of both humans and machines are the basis for these two worlds.

As controllers are responsible for air traffic control, they sometimes implement plans deviating from those of the automation. If these deviations occur in situations with high workload, the controllers do not have time to inform the assistant system about their strategies and intentions. In these cases, the automation may suggest advisories contrary to the intent of the controller, because the support system is not aware of agreements between human operators. Even worse, the operators have additional effort to inform the support systems about their communication. This situation may persist until the assistant system realizes the deviation, e.g. through the analysis of radar data. Hence, the system requires attention from the controllers, exactly when the controllers would urgently need the support of the system due to high workload.

To overcome this situation and to enable air traffic management (ATM) systems to follow the conversation between controller and pilots, ASR is an important element of future ATM assistant systems. The ability to listen has to be implemented into the assistant system. This allows following the conversation and to synchronize the intents and strategies of human and machine world. Crucial for user acceptance is the quality of ASR, especially recognition time and rate.

The integration of Automatic Speech Recognition (ASR) into ATM systems has been attempted since (at least) the early 90s. We briefly review this prior work in the next section. In section “The AcListant® Project“ we briefly describe the project which aims to integrate ASR with an assistant system, so that acceptable error rates are achieved. Section “Validation Trials” describes the performed validation at German Aerospace Center in Braunschweig. DLR’s arrival manager 4D-CARMA (4D Cooperative ARrival Manager) [5] and, the speech recognizer from Saarland University (UdS) were combined. Section “Validation Results" presents the results from trials with approach controllers from Austrian, Czech, and German Air Navigation Service Providers (ANSPs). The last section describes further steps and summarizes the results.

Background

Human-machine interaction systems have known a significant improvement in their performance in the last decade, leading to more sophisticated human-machine applications. Voice-enabled systems in particular are increasingly deployed and used in many different areas. The most popular use case among these is the Automatic Speech Recognition deployed on most mobile phones. In fact, ASR systems are becoming a significant component in hand-free systems and a cornerstone for tomorrow's applications, such as smart homes [2].

The ATM world, following and deploying the advances of today's research and technology, is increasingly developing ASR-based applications to provide more sophisticated assistant systems. ASR is a potential extension of many existing systems where speech is the primary mode of communication, such as Arrival Managers (AMAN), Surface Managers (SMAN), and Departure Managers (DMAN). First commercial implementations of an AMAN have been operational at European hubs (Frankfurt, Paris) since the early 90s. Today their application is often limited to the coordination of traffic streams between different working positions (e.g. sector and approach controllers) [6]. Implementations in Europe are e.g. OPTAMOS [7], OSYRIS [8], 4D Planner [9], [10], MAESTRO [11]. An extension of their application to support the controllers by advisories in order to implement fuel and noise efficient approaches (e.g. DLR’s AMAN 4D-CARMA [5], [12]) currently fails due to insufficient reliability of the advisories. The support quality of such systems highly depends on knowledge of the development of the situation in the airspace.

Although ASR performance improved significantly in the last decade, it is far from being a solved problem, in particular for large vocabulary applications. Limited vocabulary (in-domain) applications, however, are being more successfully deployed. Moreover, the existence of prior information about the task at hand and the expected spoken sentences summarily referred to as context can significantly improve performance. The first attempts to integrate ASR in ATM systems goes back to Hamel et al. [13] who described the application of speech technology in ATC training simulators in the

early 90s, however with limited success. Schäfer [14] used an ASR system to replace pseudo pilots in a simulation environment. He used a dynamic cognitive controller model. He, however, does not use an assistant system to dynamically generate the context so that assistant system and ASR improve each other. Dunkelberger et al. [15] described an intent monitoring system which combines ASR and reasoning techniques to increase recognition performance: In a first step, a speech recognizer analyses the speech signal and transforms it into an N-best list of hypotheses. The second step uses context information to reduce the N-best list. Our approach described in more detail in [16], however, uses context to directly reduce the recognition search space, rather than only rejecting the resulting hypotheses. The latter approach would only reduce error rate without increasing recognition rate.

These early attempts opened the gate to more concrete and successful applications of ASR in ATC training simulators, e.g. FAA [17] and DFS [18] for advanced training technologies, speech-to-text application for controller workload assessment [19] prevention of runway incursion problems due to clearances for closed or blocked runways by MITRE [20].

Although data link might replace voice communication in ATC environment, voice communication and data link with their different advantages will coexist for a long time at least in General Aviation. Here voice communication will remain the central means of coordination. The agreements coordinated by voice, automatically have to be integrated into SWIM (System Wide Information Management) based on reliable speech recognition. The same accounts for different types of on-board equipment of the airliners. This supports the coexistence of varying levels of automation in different tightly-coupled subsystems. Furthermore, careful transitions between different levels of automation are more easily possible.

Our developed ASR system AcListant® directly builds on the pilot study conducted by Shore et al. [21], [22]. The goal of his study was to provide a proof of concept for integrating situational context information into ASR for ATC task. Reported results strongly indicate that incorporating context information significantly reduces recognition error rates [12]. Shore, however, did not consider the problem of dynamically deriving the situational context.

The AcListant® Project

Based on the results of Shore, Saarland University, and DLR started the venture capital funded AcListant® project (Active Listening Assistant) in February 2013. Its aim was to provide the arrival manager with an additional pair of eyes, i.e. an additional sensor which enables to detect deviations of the controller from the AMAN plan earlier than this is possible by observing only the radar data. This enables common situation awareness without lack of information on the part of the automation and without discrepancies between voice communications and data link information.

Operational Concept

The AcListant® Workplace makes explicit use of dynamic context information by the ASR (automatic speech recognition) system. This context information is derived by an assistant system, in this case an AMAN. The assistant system informs the ASR of expected and possible air traffic controller advisories, i.e. it continuously creates context information. Consequently these speech hypotheses help the speech recognizer to detect speech commands with improved reliability. Based on the extracted information from the controller-pilot communications, the assistant system can more quickly adapt its own model, i.e. its knowledge concerning possible future system states.

In the next section we will go into detail regarding the implementation of AcListant's operational concept.

Technical Implementation

The presence of information about the target task can significantly improve the ASR performance. More particularly, radar information, flight plan data, aircraft status information etc. available during ATC tasks can be used to improve the speech recognition quality. Our Arrival Manager 4D-CARMA derives context information from above mentioned input data. The context is input into the Hypotheses Generator dynamically generating a set covering the possible

commands which can be spoken in the current situation. These predictions can then be used to improve the speech recognition performance by reducing the search space of the possible utterances.

We call this approach *Assistant Based Speech Recognition (ABSR)* [16]. Speech recognition and assistant system improve each other. On the one hand the context of the assistant system reduces the search space. This increases recognition rate and recognition speed. On the other hand the assistant sensor gets an additional input sensor, i.e. another pair of eyes. In an emergency situation the controller may e.g. change the initial sequence and give a direct waypoint or heading command to the emergency aircraft. Without speech recognition the assistant system, e.g. an AMAN, will not immediately detect the new situation. Either the controller informs the AMAN via additional mouse or keyboard commands, which causes additional workload, or the AMAN has to observe the radar data until it is obvious, that the controller will deviate from the planned sequence and trajectories. In this case, however, the AMAN only knows that the controller deviates, but he/she does still not know what the controller really wants.

With ABSR, the system immediately knows that the controller e.g. gives a "direct-to-waypoint" command. The system already knows that, before the pilot has confirmed the command via read-back or the aircraft has started the direct approach. The AMAN can update the trajectory of the emergency aircraft, and even more important, the AMAN can immediately update the whole sequence and, therefore, also can update the trajectories of the other aircraft. This is the support, the controller needs in an unusual situation!

Controller Display with speech recognition output

The prototypic air traffic controller display RadarVision (Figure 1) visualizes air traffic situation and planning data from a database, which stores among others planned touchdown sequences, aircraft arrival times, trajectories and advisories, calculated by the arrival management system. The situation data display is a conventional radar screen and visualizes aircraft positions with additional information in labels on a two-dimensional airspace map. Alphanumeric data in the flight data block consists of callsign, weight category, current altitude, speed, aircraft type, and computed distance-to-go. By highlighting an aircraft the actual heading as well as last clearances on altitude and speed are visible. Furthermore the calculated future trajectory is shown as a yellow solid line.

As depicted in Figure 1 the prototypic controller HMI RadarVision shows a radar screen with AMAN data (timeline with sequences and planned touchdown times, yellow trajectory, values for distance-to-threshold and a centerline separation range).

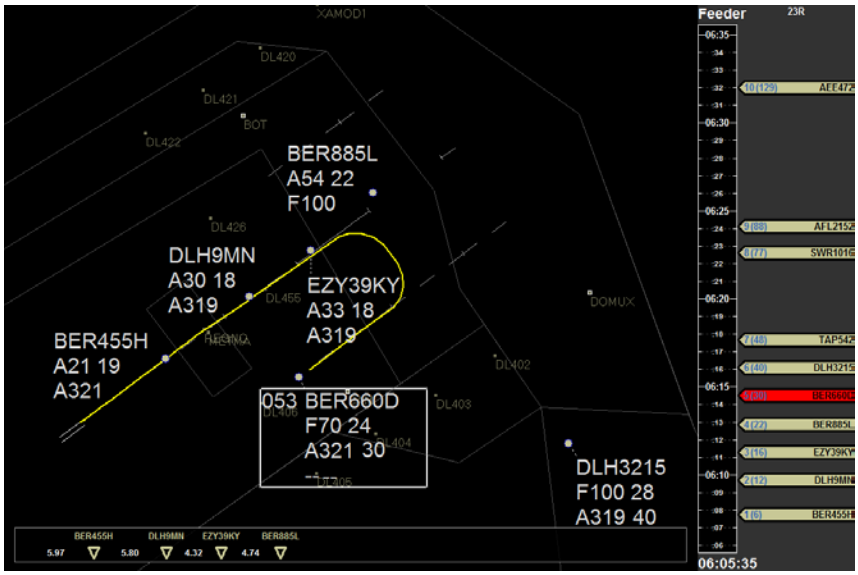

Figure 1: Prototypic Controller HMI RadarVision

The Düsseldorf airspace in AcListant® configuration on the radar screen consists of runway 23R, significant waypoints of the Aeronautical Information Publication (AIP), such as Initial and Final Approach Fixes (IAF/FAF), marked centerline and downwinds. The Centerline Separation Range at the bottom of the display in Figure 1 points out distances between aircraft on the real and virtual centerline in nautical miles. Distances are calculated for each aircraft on its final leg. Furthermore, the hypothetic distance on an elongated centerline is also computed for aircraft which are already on base leg or have a straight-in approach. The right side of the display shows a downwards moving timeline with assigned touchdown target times for each aircraft.

The Speech Recognition Log (SRL, Figure 2) lists the recognition output of the ABSR system (Ohneiser et al. [23]). Each utterance may consist of multiple commands with a callsign, a type, and a value each. Different colors indicate plausibility of the ABSR output with respect to the current situation. Green rows in the SRL symbolize plausible commands which are directly transferred to the AMAN. Blue entries are also plausible, but they are rejected due to the current context, whereas yellow rows show commands not plausible (e.g. Reduce seven zero knots or a direct and a heading command for the same callsign). Current commands of the last five seconds are highlighted in a bigger font with lighter background colors and a frame showing all commands corresponding to the last recognized callsign. A purple mouse-over highlighting not shown here is also possible. This is connected to highlighting all corresponding aircraft information in the radar, timeline, and on auxiliary screens.

The plausibility of commands depends on predefined value limits that are reasonable in the air traffic domain respectively in the approach area of aircraft. Those ranges are 150 to 300 knots for speed, 2000 to 6000 feet respectively 50 to 400 flight levels, 10 to 360 degrees for headings, and a rate of descent between 1000 and 3500 feet per minute. Other commands do not have numbers as a value but the name of a transition, a waypoint, or a runway in case of a cleared ILS command. If the value is reasonable with respect to the above mentioned requirements, the combination of value, command type, and the current aircraft state is investigated. An aircraft flying in FL 80 cannot “descend” but only “climb” to FL100. Furthermore, an explicit speed change from 250 to 200 knots would only be possible with a “reduce” command. All recognized commands violating those and similar rules are marked as invalid, i.e. they get a yellow color.

Figure 2: Speech Recognition Log

The speech recognition log constantly provides direct visual feedback to the controller. He occasionally may decide to get information about how well he/she was recognized by ABSR and have a look at the SRL. However, there is no obligation to use the output stack, so that actual work attitudes are not actively influenced. Nevertheless, the controller’s check can lead to modified voice awareness in a cognitive response. Even better recognition results could be achieved with the speaker articulating more clearly and strictly sticking to the ICAO aircraft radio regulations. The controller could be motivated by getting better automatic decision support when “causing” higher recognition rates.

It needs to be mentioned that the whole automatic speech recognition process is running in the background and, therefore, does not require any additional work of the controller. The radiotelephony channel is used anyway without a need for further typing, clicking, or other habit changes. A possible side effect could be an increased acceptance of electronic controller support systems induced by more actual and accurate information especially in air traffic situations with high-workload. The AMAN would dynamically calculate data and suggestions to be visualized more close to the real controller’s intent in case of low word error rates and predominantly correctly recognized controller commands. By experiencing benefits and better support of the

AMAN, the active encouragement to use support systems could even lead to an improved controller's behavior and interaction with the speech recognition system as well as implicitly more unambiguous communication with pilots.

Validation Trials

The final validation trials were performed from February to March 2015 at the Air Traffic Validation Center at the DLR premises in Braunschweig, Germany with seven male and four female air traffic controllers from Dusseldorf, Frankfurt, Munich, Vienna and Prague. Pre-Validation trials were performed in October 2014 with three controllers from Dusseldorf and Prague. These results are reported in [16].

Research Objectives

Two main research objectives were addressed by the validation trials: First, the functional benefits of an AMAN with/without additional input compared to the conventional working method using only radar screen, R/T and paper strips. Second, the reduction of controller workload concerning the electronic flight strip documentation when using an additional input modality, i.e. mouse input or speech recognition.

In both cases the first step is the analysis of the baseline, i.e. the situation without any controller assistant system (Run B). To answer the first question, a run with a standard AMAN without additional input (Run C, see Table 1) and a run with advanced AMAN with additional input created by ABSR support (Run D and Run G) are conducted in addition to Run C. To answer the second question, additional runs with an advanced AMAN, either with a manual input device (mouse and keyboard, Run F) or with ABSR support (Run D and G) are performed. Manual resp. ABSR inputs are used as additional information for the AMAN's planning cycle.

Experimental Setup

The experimental setup (see Figure 3) consisted of one controller working position and two pseudo pilot stations, which were linked and controlled via a supervisor station, using the simulation software tool NARSIM (NLR ATC research simulator).

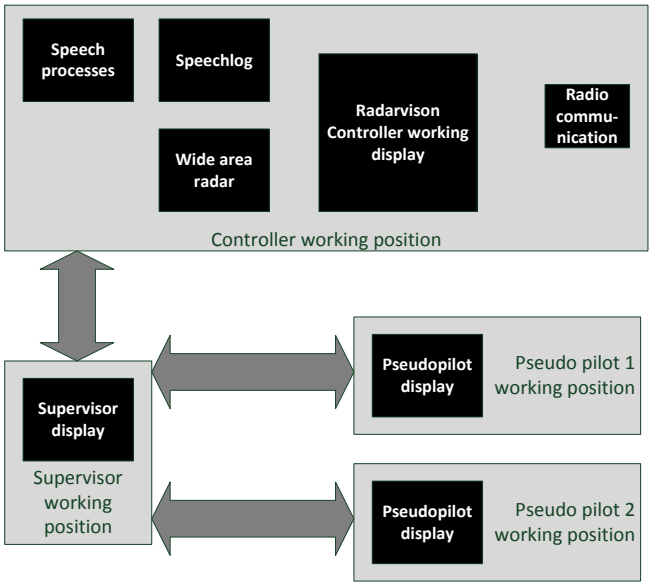

Figure 3: Experimental Setup

The controller working position was equipped with an advanced radar screen. Flight information was handled by paper flight strips, by mouse input into the label, or by speech recognition input.

The pseudo pilot stations were set up to control a maximum of ten aircraft each. They were equipped with a radar screen to improve the situational awareness of the pseudo pilots and a Narsim-specific interface to control each single aircraft. The communication between controller and pilots was established via a voice-over-IP system, which simulates the features of the standard radio system.

Two scenarios were set-up: an emergency and runway closure scenario (see Table 1). In the first one, a sick passenger is simulated and for the second one a temporary runway closure of ten minutes length is implemented. To reduce the learning effects, the callsigns are changed for each of the runs with the same scenario.

In each scenario we offered different levels of support to controllers (see Table 1). The baseline configuration ("Baseline 1") includes a state-of-the-art radar screen with speed vector information (predicted aircraft position in e.g. 90 seconds). In the AMAN configuration ("AMAN Only") we additionally offered touchdown sequences, as well as four-dimensional trajectories and distance-to-go

information. The controller had to input additional information via mouse or speech recognition to reach the highest AMAN functionality. By clicking on an aircraft label the controller can input the command with its command values, which encompass heading, speed or altitude, resp. waypoints for "direct-to" commands, controller positions for "handover" commands, runways for "cleared ILS" commands or other holding and path stretching possibilities. All these commands are also recognized by the assistant based speech recognizer.

The six simulation configurations were designed as follows:

Table 1: Overview of the Simulation Runs

Emergency Scenario	Runway Closure Scenario
Baseline 1 (Run B) No AMAN support paper flight strips	**Baseline 2 (Run E)** Mouse Input No AMAN support Electronic flight strips
"AMAN only" (Run C) AMAN support Paper flight strips	**AMAN + mouse (Run F)** AMAN support with mouse input Electronic Flight Strips
AMAN + ABSR (Run D) AMAN support with ABSR input Paper flight strips	**AMAN + ABSR (Run G)** AMAN support with ABSR input Electronic Flight Strips

Note: Training Run = Run A

Experimental Procedure

Each test person participated in six simulation runs, as well as a preceding training run, in which the controllers familiarized themselves with the different simulation configurations. Each simulation run lasted approximately 50 minutes. The validation itself took place on two half-days with three runs per day. The sequence of the simulation runs was counter-balanced across participants (random order) to avoid any sequence effects by the order of simulation runs.

During each simulation run, every five minutes a computer-based instantaneous self-assessment (ISA) [24] query was applied. Whenever a configuration contained an AMAN, also the assessment of the AMAN performance, as described beneath in section "AMAN Performance Scale", followed. After each run, computerized post-run questionnaires had to be filled out to assess user acceptance, workload (NASA TLX [24]) and situational awareness (SART [25]) when working with different configurations. The validation ended on the second day with a final debriefing.

Metrics

During the simulation various data were recorded in order to examine metrics. These metrics can be assigned to three types of data:

I. Traffic data

- Trajectory Conformance
- Sequence Stability

II. Radio communication data

- Word error rate (WER)
- Command error rate (CmdER)
- Command recognition rate (ReR)

A more detailed description of the assessment of speech recognizer metrics can be found in Helmke et al. [16].

III. Behavioral data

- 1. Workload: ISA and NASA TLX
- 2. User acceptance:
    - AMAN Performance Scale
    - Analysis of electronic flight strip data Entry (Mouse Input vs. ABSR Input)
- 3. Situational Awareness: SART

Validation Results

In this section we present the results we obtained during the final validation trials in February and March 2015 as described in section "Validation Trials".

Traffic Data

We first present the objective results which we got from the measurements in the recorded AMAN data base before presenting subjective results from controller questionnaires. Although the initial starting configuration of the runs B, C and D respectively E, F and G were the same, the resulting traffic situation were always different. In order to obtain comparable results for the quality of the AMAN planning, we needed to perform passive shadow mode trials. For that we used radar data and controller commands of seven controllers that have been recorded during the final validation trials in February 2015. Two simulation runs were performed for each set of data. In both runs the AMAN had to update its planning permanently based on the available information.

In the first of these runs (replays) the AMAN only received information about the current radar situation and had to update its plan accordingly. This was used as a baseline scenario. To evaluate the benefits of ABSR for the Arrival Manager we rerun the same scenario (replay) again, but this time the AMAN also received the recorded commands given by the controller as additional input. By doing so we were able to receive results that were only based on the different amount of information the AMAN had and not on different decisions made by the controller.

Sequence Stability

We defined the landing sequence as the order in which aircraft actually touch down and consider subsequences of successive aircraft of the landing sequence of size M. For a landing sequence of size N we can consider N-M+1subsequences. For each of these subsequences the time is determined until the AMAN is able to predict the touchdown order reliable. In our case a subsequence were considered reliable when every aircraft of the subsequence, deviates not more than one position from its actual touchdown position. We measure the time in seconds until the landing of the last airplane in the subsequence. This derived measurement gives a hint concerning the stability of the AMAN.

Table 2 shows the results of evaluating the planning stability. Depending on the controller and the sequence size the improvement varies between 37 seconds and 4 minutes. On average we have an improvement of roughly two minutes considering subsequences of 6 aircraft.

Table 2 : Sequence Stability

Sequence Size	**3**	**4**	**5**	**6**
C1 – with ABSR	801 sec	810 sec	817 sec	741 sec
C1 – without ABSR	693 sec	696 sec	665 sec	600 sec
Improvement	108 sec	114 sec	152 sec	141 sec
C2 – with ABSR	819 sec	825 sec	799 sec	724 sec
C2 – without ABSR	743 sec	736 sec	716 sec	651 sec
Improvement	76 sec	89 sec	82 sec	73 sec
C3 – with ABSR	790 sec	798 sec	801 sec	729 sec
C3 – without ABSR	752 sec	748 sec	721 sec	617 sec
Improvement	37 sec	50 sec	80 sec	112 sec
C4 – with ABSR	804 sec	808 sec	785 sec	706 sec
C4 – without ABSR	742 sec	721 sec	702 sec	634 sec
Improvement	62 sec	87 sec	83 sec	72 sec
C5 – with ABSR	747 sec	699 sec	598 sec	464 sec
C5 – without ABSR	703 sec	613 sec	505 sec	380 sec
Improvement	44 sec	86 sec	93 sec	83 sec
C6 – with ABSR	732 sec	681 sec	696 sec	505 sec
C6 – without ABSR	693 sec	589 sec	475 sec	260 sec
Improvement	39 sec	92 sec	221 sec	245 sec
C7 – with ABSR	802 sec	780 sec	761 sec	680 sec
C7 – without ABSR	713 sec	699 sec	679 sec	606 sec
Improvement	88 sec	81 sec	83 sec	74 sec
Avg. Improvement	65 sec	86 sec	113 sec	114 sec

We compared the effects of "ABSR vs. Without ABSR Input" on AMAN sequence stability for the different sequence sizes with a pair wise t-test analysis. The most significant effects were found for the sequence size 3 ($t(7) = -6.36$, $p < .001$, $d = -2.23$)) and sequence size 4 ($t(7) = -11.96$, $p < .001$, $d = 1.43$)), indicating that ABSR effectively supports the AMAN with a better planning functionality than without ABSR. Also significant effects for sequence size 5 (($t(7) = -5.56$, $p < .01$, $d = 1.23$)) and sequence size 6 (($t(7) = -4.80$, $p < .01$, $d = 0.84$)) were found.

Trajectory Conformance

4D-CARMA determines for each aircraft if the radar data is conform to the actual planned trajectory. The conformance monitoring considers lateral deviations (> 0.5 NM), vertical deviations (> 500 ft.), and temporal deviations (> 10 seconds). Based on these deviations each aircraft gets the status conform or non-conform. We calculate for each aircraft the total time the aircraft is in status non-conform (NConfT) and how often the status changes from conform to non-conform (NConfCnt). This measurement indicates how long resp. how often the internal picture of the controller differs from that of the machine.

Table 3 shows the improved conformance between the planned trajectory and the actual radar data. In runs without ABSR the overall percentage of NonConfT varies between 10% and 18%, whereas in runs with ABSR it only varies in between 3% and 10%. Also the number of times an aircraft switches in the state non-conform has been reduced significantly. This was also confirmed in another pair wise t-test analysis comparing the total time of all aircraft (NConfT) in status non-conform ($t(7) = 8.94$, $p < .001$, $d = 3.43$)) and the number of status changes from conform to non-conform ($t(7) = 10.95$, $p < .001$, $d = 3.58$)).

Table 3: Trajectory Conformance

		NConfT (all AC)	NConfT %	NConfCnt-Average
C1	*with ABSR*	601 sec	4.2%	2
	without ABSR	2545 sec	17.9%	10
C2	*with ABSR*	600 sec	4.1%	3
	without ABSR	2255 sec	15.4%	8
C3	*with ABSR*	745 sec	5.1%	3
	without ABSR	2364 sec	16.2%	9
C4	*with ABSR*	603 sec	4.1%	2
	without ABSR	1901 sec	13.0%	7
C5	*with ABSR*	560 sec	4.1%	2
	without ABSR	2355 sec	17.0%	9
C6	*with ABSR*	1803 sec	9.9%	6
	without ABSR	2624 sec	14.4%	11
C7	*with ABSR*	561 sec	3.7%	2
	without ABSR	1532 sec	10.0%	6

Radio Communication Data

ASR systems generally use the Word Error Rate (WER) metric for evaluation. This metric is defined as the distance between the recognized word sequence and the sequence of words which were actually spoken, referred to as the gold standard (see pp. 362-364 in [26]). WER is defined as a derivation of Levenshtein distance [27]:

$$WER(s) = \frac{ins(s) + del(s) + sub(s)}{W(s)} \quad (1)$$

Here, *ins(s)* is the number of word insertions (words never spoken), *del(s)* is the number of deletions (words missed by ASR), *sub(s)* is the number of substituted words, and *W(s)* is the number of words actually said. In ATM, however, the WER is not descriptive enough as metric. One would rather prefer a metric assessing the rate of correctly recognized concepts. It is not important that ASR correctly recognizes “Good morning Air France one two tree descend level one three zero”, but that the concept “AFR123 DESCEND FL130” is correctly extracted. The command error rate (CER) quantifies this metric. In Eq. (1) *ins(s)*, therefore, designs the number of commands insertions.

We distinguish between (1) recognition rate, which measure the number of commands correctly recognized and rejected by plausibility checks, (2) error, which considers the number of commands wrongly recognized and not rejected and (3) the rejection rate, which considers the number of commands being rejected independent of correctly or wrongly rejected. Table 4 summarizes the achieved results during the trials.

Table 4: Rates Observed During Trials

Recognition Rate	Rejection Rate	Error Rate
91.6%	8.4%	3.0%

When the controller gives additional information to the pilot (e.g. expect delays due to runway closure, you are number five in sequence, runway reopen at 8 hundred, reduce speed two three zero knots), which deviates from phraseology the speech recognizer

often recognizes more commands than really given by the controller, i.e. the number ins(s) in Eq. (1) is high. This is the reason why the sum of the percentages is greater than 100%.

Depending on the accepted rejection rate of the speech recognizer we got command error rates between 2% and 5% resulting in command recognition rates between 90% and 95%. These recognition rates were, however, only achieved with assistant based speech recognition, i.e. an AMAN dynamically generates context information to increase the recognition rate. Without context generation the recognition rate was between 50 and 80%. Since, the main focus of our data analysis in this paper lies on behavioral data of air traffic controllers we refer to a more thorough speech recognizer related data analysis in [16].

Behavioural Data

In this part of the result section we present various subjective data that we assessed of the air traffic controllers.

AMAN Performance Scale

Tailored to one of the validation objectives the AMAN Performance Scale was developed to continuously evaluate the subjective opinion on the AMAN performance and was thus only used in the configurations with AMAN support. It is a five point scale (see Figure 4 below) ranging from “strongly agree” (indicating 100% conformity or grade “A”) to “strongly disagree” (0% conformity or grade “E”) . The controllers were asked every five minutes during the runs to rate subsequently their degree of agreement with the AMAN sequence and the AMAN trajectory by tapping on either the grade or description which suited best.

Grade	A	B	C	D	E
Feeling	strongly agree	agree	neutral	Disagree	strongly disagree
Conformity [%]	100%	75%	50%	25%	0%

Figure 4: AMAN Performance Scale

AMAN Sequence

A comparison of the results for the support levels AMAN and AMAN with ABSR showed a higher agreement with the proposed sequence in the condition AMAN with ABSR (see Figure 5 below). Note, that a lower score (e.g. grade A = 1, grade B = 2) represents a higher agreement. Interestingly, AMAN planning that was based solely on radar data ("Run C") was rated better than when the ATCO entered flight data ("Run F") with the mouse. Here the AMAN at least obtained additional information, which had to be entered by mouse suggesting an improved AMAN planning. Although the descriptive analysis reveals better ratings for an AMAN sequence updated with ABSR when compared to AMAN without any support multiple pair wise t-tests did not reveal significant effects of an improved AMAN Sequence.

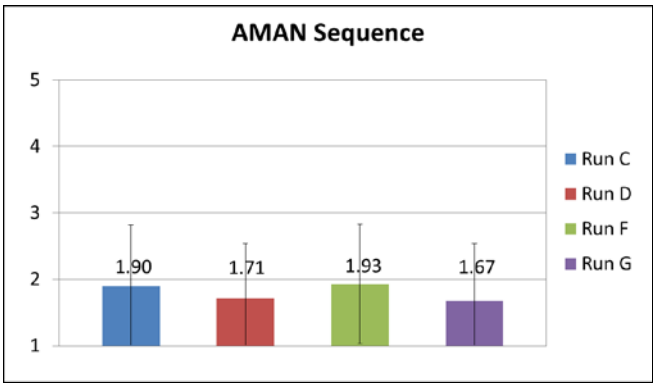

Figure 5: AMAN Sequence

AMAN Trajectory

The ATCOs rated also the trajectory conformance by comparing the AMAN trajectories with their planned trajectories as follows: "Run D" was rated best ($t(58) = 2.27$, $p < .05$, $d = 0.41$)), indicating that an ABSR support increased the level of consistency between AMAN and ATCO plan (see Figure 6). At worst, the trajectory conformance for "Run C" was rated, in which the AMAN calculates its trajectory only on the basis of radar data. Pairwise t-tests did also reveal significant effects for the support levels AMAN with mouse (Run F, ($t(58) = 2.27$, $p < .05$, $d = 0.41$)) and AMAN with ABSR (Run D ($t(58) = 2.27$, $p < .05$, $d = 0.41$) , Run G ($t(58) = 2.02$, $p < .05$, d = 0.33)) when contrasted with the baseline run C.

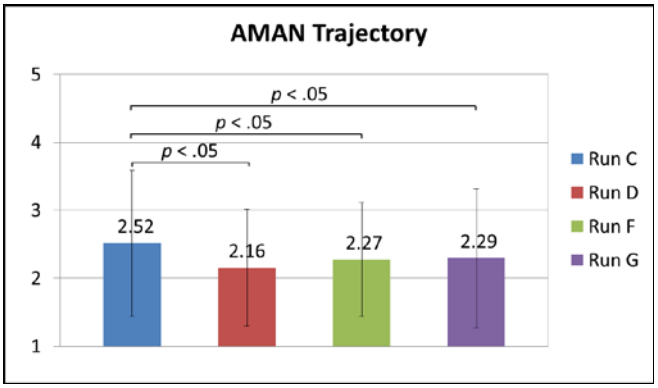

Figure 6: AMAN Trajectory

Workload

The workload of the controllers was measured with the Instantaneous self-assessment (ISA) method and NASA TLX. ISA is an on-line subjective measure of workload, i.e. the workload is recorded every five minutes during the simulation and not afterwards. It is a five point scale which ranges from "1= underutilized" to "5 = excessive" with intermediate rating points. The results as depicted in Figure 7 show that the workload increased when the controllers had to input the additional information via mouse (Run E and F). The workload was again significantly reduced to the normal baseline level when this input was made by the speech recognizer (Run G).

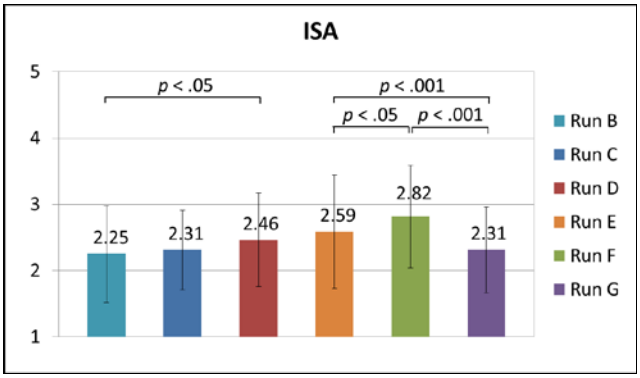

Figure 7: ISA workload

This could be confirmed by post-hoc pair-wise t-tests (Run F vs. Run G (t(73) = 4.98, p < .001, d = 0.73 and Run E vs. Run G, t(81) = 3.47, p < .001, d = 0.45)) which revealed highly significant effects, i.e. ABSR inputs are producing significantly less workload than when using a mouse as an additional input modality. Interestingly, the level of perceived workload between the conventional method of operation (paper flight strips and no AMAN, ISA Score = 2.25) and new method of operation (AMAN-ABSR and use of electronic flight strips, ISA Score = 2.31) is nearly the same and thus negligible.

Post-run assessments of workload were carried out with the NASA TLX. The results in Figure 8 represent the overall workload score assessed for each simulation run. The analysis of the NASA TLX showed the lowest workload score for the baseline (Run B) which corresponds well with the ISA results (see Figure 7). Participants evaluated Run E (mouse input, electronic flight strips, no AMAN support) to be the highest although the difference failed to reach significance after we calculated analysis of variance (ANOVAs). A pairwise t-test however showed a significant difference between run B and run E (baseline vs. mouse-input, ($t(10) = -5.06$, $p < .05$, $d = 0.78$)) but since the scenario types (emergency vs runway closure) were different in terms of traffic complexity both runs, the comparability is difficult.

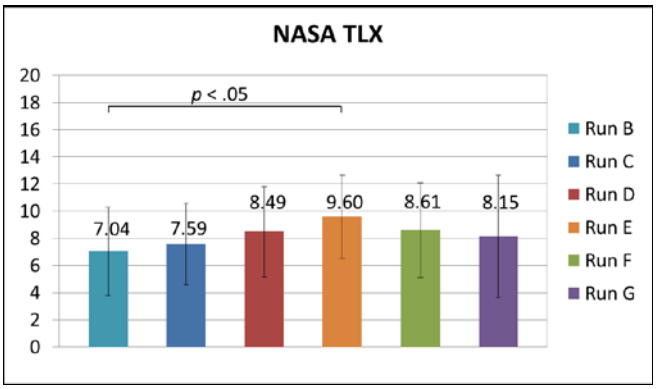

Figure 8: NASA TLX workload scores

Situational Awareness

The assessment of the air traffic controllers' situational awareness in dependence on the configuration was carried out with the 3D SART. This is a narrowed down version of SART covering three dimensions of situational awareness (demand and supply of attentional resources and understanding of situation). The 3D SART results revealed (see Figure 9), that situational awareness was evaluated as to be the lowest in Run E (electronic flight strips with

mouse input, without AMAN without speech) and highest in Run G (with electronic flight strips with AMAN, with voice recognition). However, inferential statistics did not show significant effects of the configuration on situational awareness.

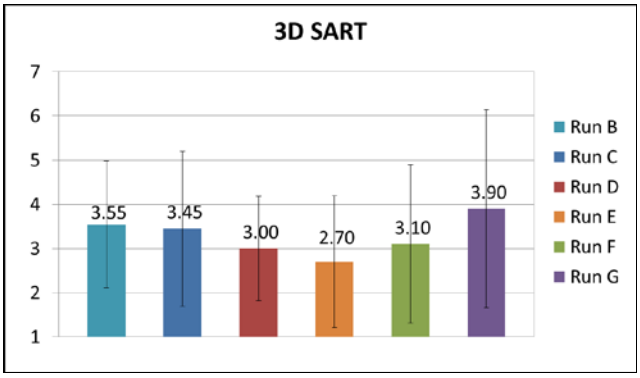

Figure 9: 3D SART situational awareness results

General User Acceptance

A tailored questionnaire was developed in order to assess the user acceptance. One important aspect of the user acceptance assessment was the ATCOs opinion on the effectiveness of support of the various workstation configurations. The results show that the speech recognizer (Run D and G) provides a significantly higher support, with configuration "AMAN + ABSR" and electronic flight strips yielded the highest ranking. A highly significant effect was found in the comparison of Run E vs. Run G ($t(10) = 5.16$, $p < .001$, $d = 2.30$). The configurations with mouse input received the lowest acceptance scores.

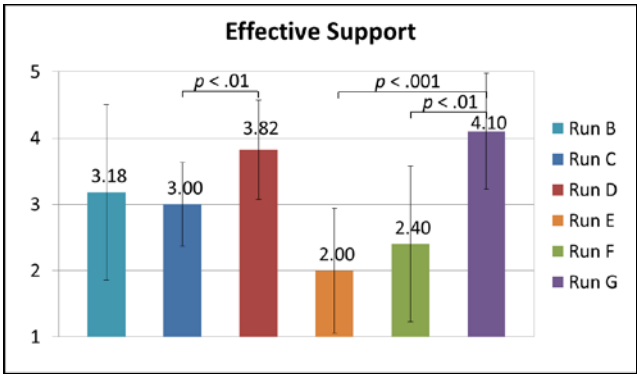

Figure 10: Post-Run Questionnaire Assessing Effectiveness of Controller Support Dependent of Configuration

In both runs with paper flight strips (Run C and D) as well as in runs with electronic flight strips (Run E, F and G), the speech recognizer supported the ATCOs significantly more effective in their work than other configuration did.

Electronic Flight Strips Data Entry

In this section we compare how often the real command given via voice to the pilots was stored in the data base. It was stored in the data base, if ABSR recognizes it correctly (and was not rejected) and in the case we use mouse input, how often the controller inputs the given (spoken) command also via mouse into the label. Table 5 shows the results.

Table 5: Accuracy of mouse and ABSR input

	Speech	Mouse
Total given commands	189.2	160.2
Rate of correct commands	91.9%	77.6%
ErrorRate	2.2%	10.7%
Rejection rate	8.1%	22.5%

We were surprised that only 77.6% of the given commands were inputted also via mouse into the label. This is even more surprising because we did not require inputting all given commands. If the controller gives an ILS clearance together with a heading or with a descend command only the ILS clearance was required. If the controller repeats a clearance we of course only expected one mouse input and so on. Therefore, the number of total given commands (average per scenario) is much smaller when mouse is used. The controller does not only forget to input a command, but also 10.7% of these commands were never given to the pilot.

The used HMI for mouse input was not optimized for that task and we also did not consider the effect, that the controller gives a command (e.g. heading 320), the pilot reads back a slightly different command (e.g. heading 310) and the controller did not correct, but just inputs the readback to avoid further frequency usage. Nevertheless the measured effect is significant and requires further analyzes. With assistant based speech recognition we now have the tools to really analyze these effects with low effort.

Conclusions and Outlook

With our validation setup consisting of six simulation runs for eleven different air traffic controllers, we assessed both, the workload reduction by means of ABSR as well as the increased support for the controller.

Results of this study revealed that ABSR positively influences the controller workload, especially in terms of electronic flight strip documentation. Another outcome was the functionality benefits of an ABSR based AMAN in terms of a higher degree of trajectory conformance and sequence stability when compared to the performance of an AMAN merely based on radar updates.

In summary, the results of this study point out that ABSR supports the controller better and more effectively than without any sensor or even when updated with mouse. Results regarding the latter showed that with mouse input roughly 78% of the given commands were correct, whereas with ABSR this was true in 92% of the cases.

Another interesting finding of the study was, that the analysis of the workload measures (ISA and NASA TLX) revealed that even with a more complex traffic scenario like a runway closure the amount of perceived workload was almost the same (see Figure 7 and Figure 8) when the controller worked conventionally (paper flight strips, R/T and situation data display). This indicates that the new working method based on ABSR is possibly paving the way for the "digital revolution" in ATM.

References

[1] AcListant homepage: www.AcListant.de

[2] SRI International "Siri-based virtual personal assistant technology" http://www.sri.com/engage/ventures/siri

[3] Schalkwyk, J., D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, B. Strope, 2010, Google search by voice: A case study, in Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, pp. 61–90. Springer.

[4] Hamerich, S.W., 2007, Towards advanced speech driven navigation systems for cars, in "Intelligent Environments," 2007. IE 07. 3rd IET International Conference, Sept. 2007, pp. 247-250.

[5] Helmke, H., R. Hann, M. Uebbing-Rumke, D. Müller, D. Wittkowski, 2009, Time-based arrival management for dual threshold operation and continuous descent approaches, 8th USA/Europe ATM R&D Seminar, 29. Jun. - 2. Jul. 2009, Napa, California (USA), 2009.

[6] Hasevoets, N., P. Conroy, 2010, AMAN Status Review 2010, Eurocontrol, Edition number 0.1, 17 December, 2010.

[7] Avibit "AMAN OPTAMOS," http://www.avibit.com/Solutions/OPTAMOS.htm.

[8] Barco "AMAN OSYRIS," http://www.barco.com/en/product/1229.

[9] DFS, "AMAN 4D Planner," http://www.dfs.de/dfs/internet2008/module/worldwide_solutions/deutsch/worldwide_solutions/download/4d_planner.pdf.

[10] Gerling, G. D. Seidel, 2002, "Project 4D-Planner," Scient. Seminar, Braunschweig.

[11] Egis-Avia, "AMAN MAESTRO," http://www.egis-avia.com/products/ ATC-Systems

[12] Helmke, H., H. Ehr, M. Kleinert, F. Faubel, D. Klakow, 2013, Increased Acceptance of Controller Assistance by Automatic Speech Recognition, in: Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013), Chicago, IL, USA, 2013.

[13] Hamel, C., D. Kotick, M. Layton, 1989, Microcomputer System Integration for Air Control Training, Special Report SR89-01, Naval Training Systems Center, Orlando, FL., USA, 1989.

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

[15] Dunkelberger, K.,R. Eckert, 1995, Magnavox Intent Monitoring System for ATC Applications, Magnavox.

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

[17] FAA, 2012, 2012 National Aviation Research Plan (NARP), March 2012.

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

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

[20] Chen, S.,Hunter Kopald, 2015, The Closed Runway Operation Prevention Device: Applying Automatic Speech Recognition Technology for Aviation Safety, in: Eleventh USA/ Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[21] Shore, T., F. Faubel, H. Helmke, D. Klakow, 2012, Knowledge-Based Word Lattice Rescoring in a Dynamic Context, Interspeech 2012, Sep. 2012, Portland, Oregon.

[22] Shore, T., 2011, Knowledge-based word lattice re-scoring in a dynamic context, master thesis, Saarland University (UdS), 2011.

[23] Ohneiser, O. , H. Helmke, H. Ehr, H. Gürlük, M. Hössl, T. Mühlhausen, Y. Oualil, M. Schulder, A. Schmidt, A. Khan, D. Klakow, 2014, Air Traffic Controller Support by Speech Recognition,in: "Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, Advances in Human Aspects of Transportation: Part II," N. Stanton, S. Landry, G. Di Bucchianico, and A. Vallicelli, Eds. Krakow, Poland: CRC Press, 2014, ISBN: 978-1-4951-2098-5, pp. 492-503.

[24] Hart, S.G., L.E. Staveland, 1988, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), Human mental workload (pp. 139–183). Amsterdam: North-Holland.

[25] Taylor, R.M., 1990, Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478) pp3/1 – 3/17, Neuilly Sur Seine, France: NATO-AGARD.

[26] Jurafsky, D., J. H. Martin, 2008, Speech and Language Processing: An In-troduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edition. Englewood Cliffs, NJ, USA: Prentice-Hall, 9th Feb. 2008.

[27] Levenshtein, V. I., 1966, Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics, Doklady 10.8, Feb. 1966.

Acknowledgements

The work was conducted in the AcListant® project, which is supported by DLR Technology Marketing and Helmholtz Validation Fund. We also like to thank Jörg Buxbaum, Rocco Bandello from DFS (Deutsche Flugsicherung GmbH) and Jiri Janda (Air Navigation Services of the Czech Republic) for their valuable inputs when preparing the trials.

Email Addresses

mail to: Hejar.Guerluek@dlr.de
Hartmut.Helmke@dlr.de
Matthias.Wies@dlr.de
Heiko.Ehr@dlr.de
Matthias.Kleinert@dlr.de
Thorsten.Muehlhausen@dlr.de
Kathleen.Muth@dlr.de
Oliver.Ohneiser@dlr.de