Introduction

Two different working groups discussed safety aspects of SELF-MADE-ATC.

Two working groups (8 members each) met — WG1 with five attendees in Braunschweig, WG2 fully on site.

Motivation

- The assistant systems contain models, which are AI-based. Not the whole system is AI-based. Safety nets including the human at least in the near future will still be rule-based.
- ASRU has a recognition rate on command level of 90%+ and an error rate of 1%. The rest is “I do not know”.

This working group focusses on updates of an already existing and certified application. Considered aspects are e.g.:

- **Deployment Speed vs. Validation:** Can an updated model be rolled out overnight or does it require an extended validation period (e.g., eight months of testing)?
- **Operational Trust:** Under what specific circumstances and metrics can we safely trust a newly trained model?
- **Human-in-the-Loop & Control:** Should supervisors or ATCOs have dedicated mechanisms to provide direct feedback on the recognition quality or to even shut down the ASRU system?
- Do we need the MONITOR and/or the EVALUATION component of the SELF-MADE-ATC toolset? Which improvements are needed? Can you imagine that your organization will test this?

Summary:

It is accepted that ASRU does not recognize 100% and it is also accepted that ASRU has an error rate of greater 1%. The consequence is that full automation based on ASRU inputs is not possible. The ATCO stays responsible. If the ATCO has the feeling that ASRU usage increases the workload or decreases safety, (s)he needs the possibility to shut down the ASRU system. When an ASRU system is certified once, regular updates after sufficient testing are acceptable.

Results of the Working Groups

Working Group 1

The working group agreed on

- We accept that the systems will not achieve 100% of accuracy,
- We accept 1% of wrong identification and use them operationally; the controller and the other assistant systems are responsible that these errors do not result in safety critical situations
- Continuous updates of models to increase accuracy are appreciated. Reasons for this could be infrastructure changes, AIRAC cycle updates, new call signs, i.e. new airline designators (air berlin does not exist anymore, airline itarrow is new). In those cases, an automated training for models without human interaction is preferred.
- Training with voice recordings from the ops room for the new data are not always available. In those cases, training with synthetic voice might be an option with, however, limited performance for the new words

Mitigation Strategies to cope with errors of the ASRU system:

- The ATCO is still responsible; a confidence level on ASRU output can reduce ATCO's workload;
- A possibility for a user-friendly roll back of ATCO inputs into the ATM system initiated by ASRU must be available. In the extreme case, a shut-down of the whole ASRU system must be available.
- Feedback loop to model (human controller correction).
- Benefits (decrease of workload) provide incentive for ATCOs to stick to phraseology => accuracy will increase over time and therefore even safety will increase since ATCOs will conform better to ICAO phraseology.

Working Group 2

The second group agreed on the following results:

- An Option to switch off the ASRU support is definitely needed due to AIRAC update
- The shut-down might only affect some CWPs' (Controller Working Position) operational work, but not the work at other stations

- Suddenly shutting down might be bad in high traffic if I have more traffic DUE to the assistant system at the moment
- 28 days AIRAC cycle analogue for deployment of models
- The first seven days the confidence threshold should be higher, i.e., more ASRU output is shown as „I am not so sure“; after 7 days, there can be a decision reset the threshold (and/or to use the 7 day operational recordings for deploying a new model; because the model before might have relied on synthetic data „only“ for the additions in AIRAC); extra effort for intermediate new model must be low
- A simple „report issue“ button in DLR’s SpeechLog would be acceptable; but who is reporting? Some ATCOs/supervisors might have time, but not all→system should REDUCE workload, not increase; Therefore, the question was discussed whether a technician or the supervisor can do this job. No final answer could be agreed on.
- On-going discussion: Is it acceptable to generate further test data, i.e., some controllers uttering something for a few minutes and see the model output before deployment? More convenient would be synthesized tests. Which option has the highest effect, i.e. ASRU performance, is still on open question for research. Therefore, a test installation of SELF-MADE-ATC could help.