HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

Readback Error Detection by Automatic Speech Recognition to Increase ATM Safety

Hartmut Helmke, Matthias Kleinert, Shruthi Shetty, Oliver Ohneiser, Heiko Ehr (DLR),
Hörður Arilíusson, Teodor S. Simiganoschi (Isavia ANS),
Amrutha Prasad, Petr Motlicek (Idiap),
Karel Veselý, Karel Ondřej, Pavel Smrz (BUT),
Julia Harfmann (NATS), Christian Windisch (Austro Control)

14 authors from
6 partners

Founding Members
EUROPEAN UNION EUROCONTROL

Motivation

ATCo

good morning speed bird two zero zero zero alfa
reduce one eight zero knots until DME four miles
contact tower
on frequency one one eight decimal seven zero zero

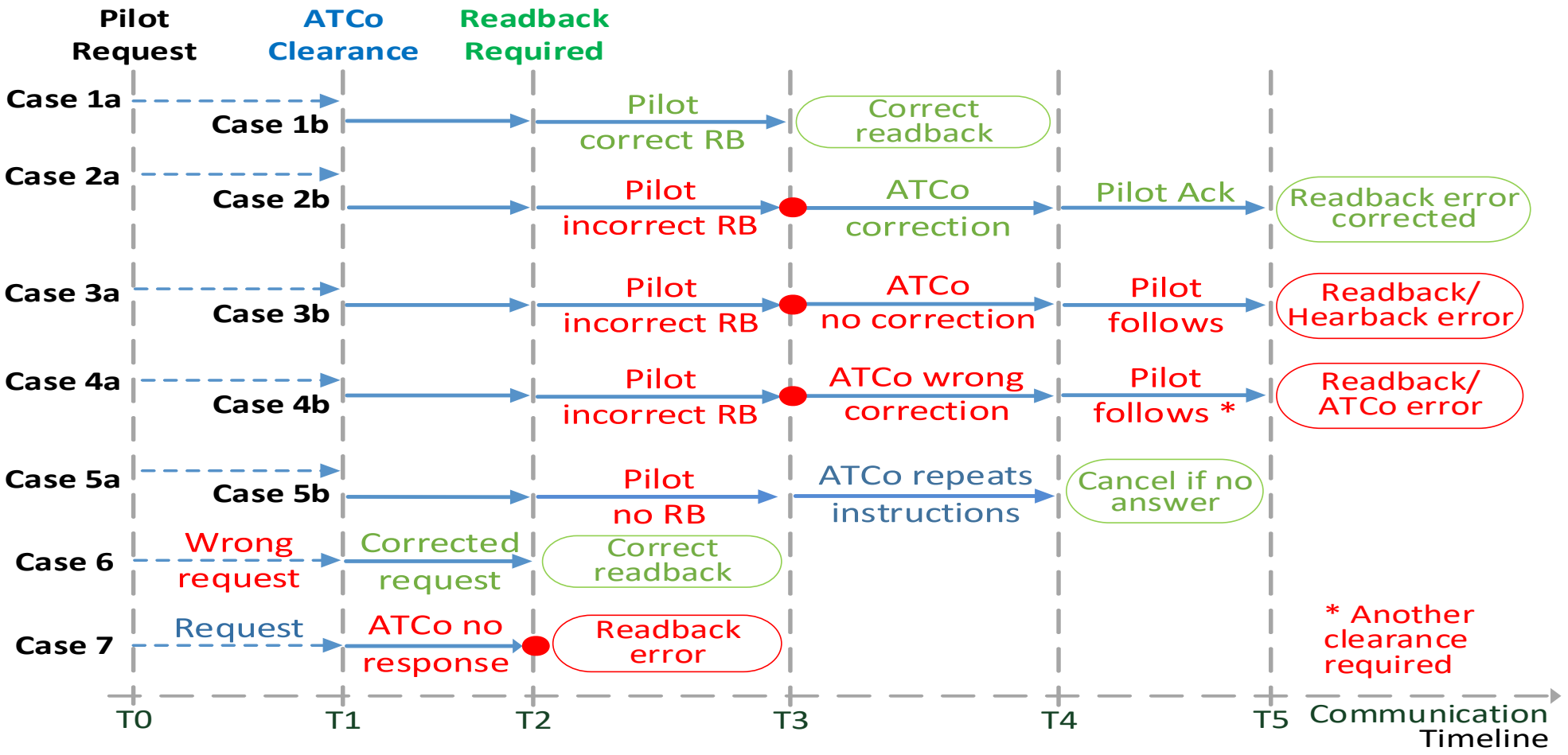
Pilot

one eighty to DME four
tower eighteen seven
speed bird two thousand alfa

Is this a readback error or a correct readback?

Contents in Detail

Overview of the Use Cases



Use Cases

	Sequence of spoken Words / Transcription
Pilot	reykjavik control faeroline five five requesting lower
ATCo	faeroline five five descend flight level three one zero
Pilot	descending three one zero faeroline five five
REDA	No Readback error

REDA = Readback Error Detection Assistant

	Sequence of spoken Words / Transcription
ATCo	faeroline five five further descend two five zero
Pilot	descending two one zero
REDA	Readback error

No ATCo reaction → **Hearback Error**

ATCo	faeroline five five **negative** descend two five zero
Pilot	descending two five zero
REDA	No readback error Readback error indication should disappear

Use Cases

ATCo	faeroline five five descend two five zero
Pilot	“no reply”
REDA	Missing readback indicator, which will be deleted when ATCo repeats the clearance

Use Cases with more than one Pilot

	Sequence of spoken Words / Transcription
ATCo	lufthansa two alfa four turn left heading three two zero
Pilot 1	two alfa four turning right three two zero
REDA	Readback error for DLH2A4,

Use Cases with more than one Pilot

	Sequence of spoken Words / Transcription
ATCo	lufthansa two alfa four turn left heading three two zero
Pilot 1	two alfa four turning right three two zero
REDA	Readback error for DLH2A4,
ATCo	speed bird one one descend flight level one two zero
REDA	Hearback error for DLH2A4, because ATCo talks to BAW11

Use Cases with more than one Pilot

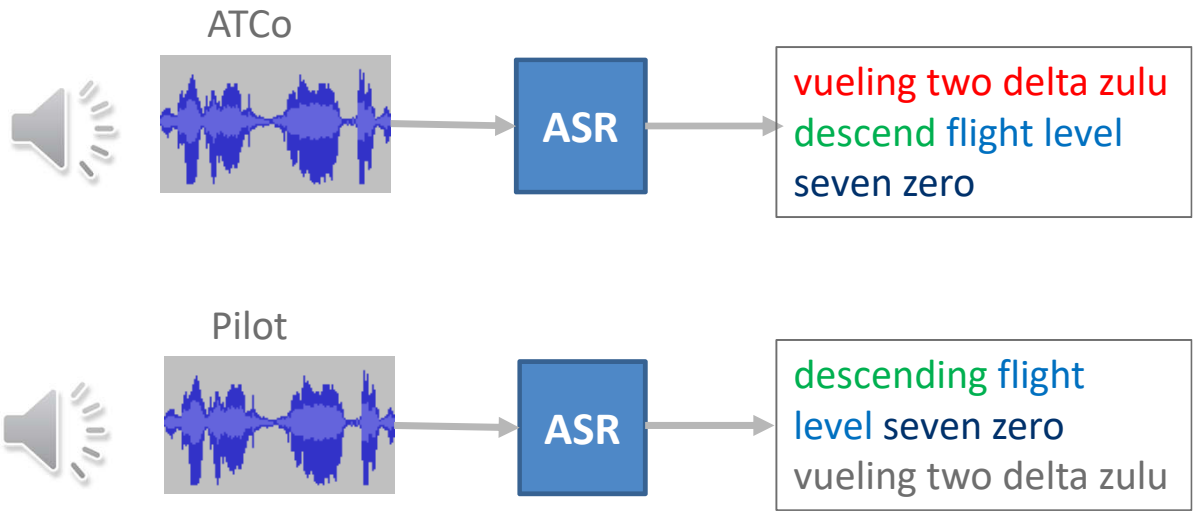
	Sequence of spoken Words / Transcription
ATCo	lufthansa two alfa four turn left heading three two zero
Pilot 1	two alfa four turning right three two zero
REDA	Readback error for DLH2A4,
ATCo	speed bird one one descend flight level one two zero
REDA	Hearback error for DLH2A4
Pilot 2	descending level one two zero speed bird one one
REDA	No readback / hearback error for BAW11, but still for DLH2A4

Use Cases with more than one Pilot

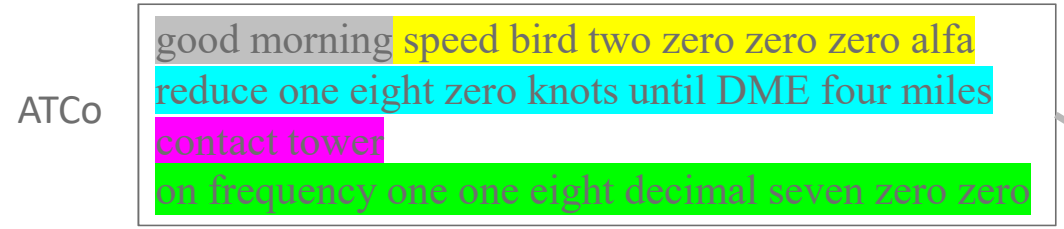
	Sequence of spoken Words / Transcription
ATCo	lufthansa two alfa four turn left heading three two zero
Pilot 1	two alfa four turning right three two zero
ATCo	speed bird one one descend flight level one two zero
REDA	Readback error for DLH2A4
Pilot 2	descending level one two zero speed bird one one
REDA	No readback error for BAW11, but still for DLH2A4
ATCo	lufthansa two alfa four negative turn left heading three two zero turn left
REDA	Readback error indicator now disappears also for DLH2A4
Pilot 1	two alfa four turning left three two zero

Contents in Detail

Recognition of ATCos and Pilots not Easy



Understanding even more Challenging

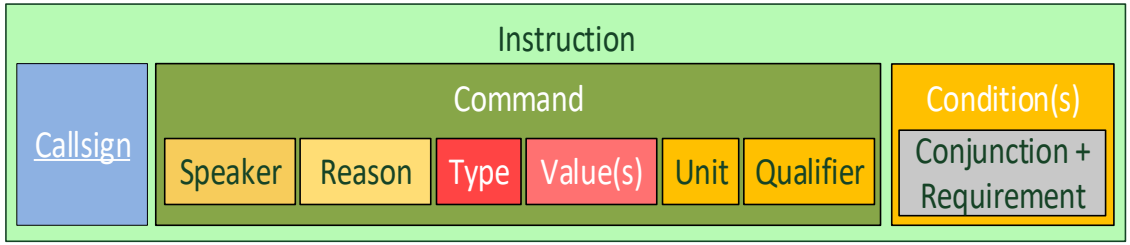


Speed Recognition is NOT
Speech Understanding
Alan Turing 1952

Readback Error?

- Word sequences are different
- Not each command needs a readback
- Sequence of commands can be different
- “nineteen” and “one one nine” are the same
- “thousand” and “zero zero zero” are the same

Ontology enables just String Matching



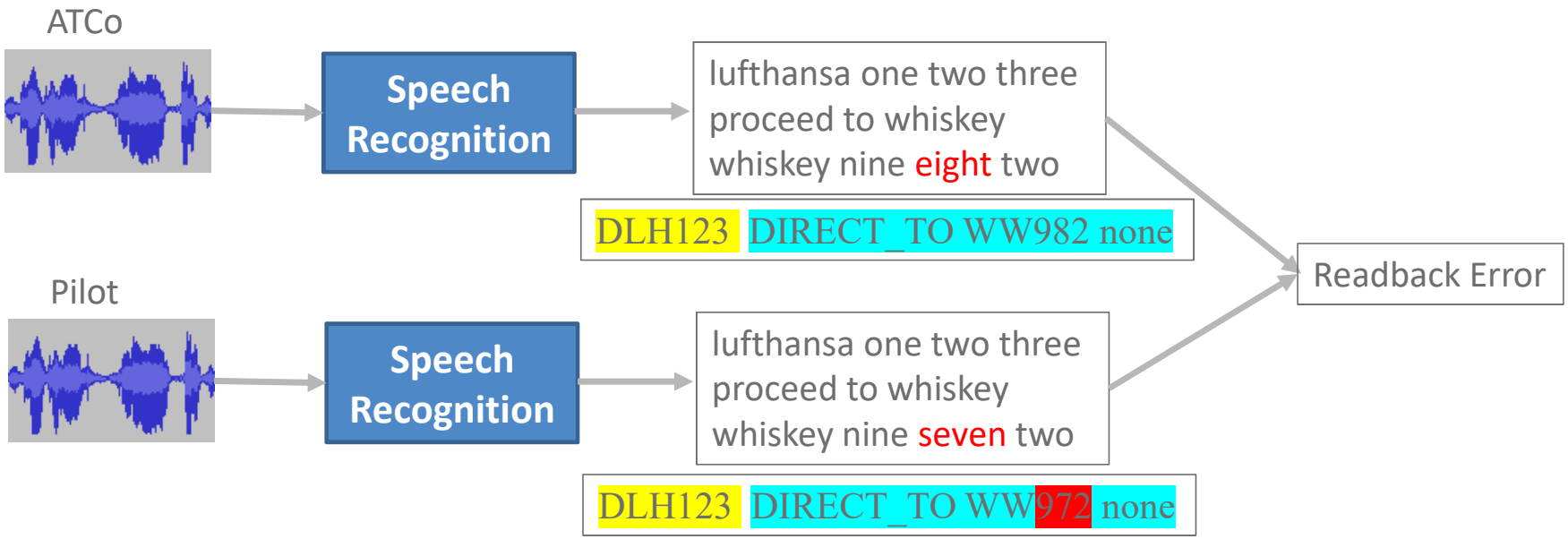
good morning speed bird two zero zero zero alfa
reduce one eight zero knots until DME four miles
contact tower
on frequency one one eight decimal seven zero zero

BAW2000A REDUCE 180 kt UNTIL 4 NM DME
BAW2000A CONTACT TOWER
BAW2000A CONTACT_FREQUENCY 118.700

one eighty to DME four
tower one eighteen seven
speed bird two thousand alfa

BAW2000A PILOT SPEED 180 none UNTIL 4 none DME
BAW2000A PILOT CONTACT TOWER
BAW2000A PILOT CONTACT_FREQUENCY 118.700

ASR Applications of HAAWAII
Readback Error Detection (simple)



No Common Opinion whether RBE or not

virgin five one four contact arrival one two four decimal nine eight zero

ATCo

VIR514 CONTACT ARRIVAL
VIR514 CONTACT_FREQUENCY 124.980

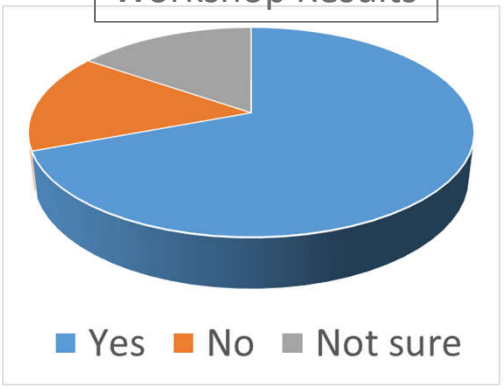
virgin five one four contact arrival one two four decimal nine zero zero bye

Pilot

VIR514 PILOT CONTACT ARRIVAL
VIR514 PILOT CONTACT_FREQUENCY 124.900
VIR514 PILOT FAREWELL

1. Readback Error?
   a. Yes, of course
   b. No, never
   c. Not sure, maybe

Workshop Results



No discussion on the phraseology and on rules of the annotation. That will be another talk.

HAAWAII-Stakeholder Workshop:
https://www.haawaii.de/wp/first-stakeholder-workshop-presentation-slides

Even more Discussion

leave gedern radar heading two seven zero degrees with present speed and that's copied speed bird nine six victor charlie

ATCo

BAW96VC HEADING 270 none WHEN PASSING GED
BAW96VC MAINTAIN PRESENT_SPEED

heading two seven zero from gedern with present speed

Pilot

NO_CALLSIGN HEADING 270 none WHEN PASSING GED
NO_CALLSIGN MAINTAIN PRESENT_SPEED

1. Readback Error?
a. Yes, of course
b. No, never
c. Not sure, maybe

HAAWAII-Stakeholder Workshop:
https://www.haawaii.de/wp/first-stakeholder-workshop-presentation-slides

Use Callsign in Readback or Not?

	ANSP1	ANSP2	ANSP3
ATCo utterance without callsign	15%	12%	8%
Pilot utterance without callsign	19%	10%	6%

- Callsign often missing
- Especially after immediate responses callsign is missing
- Good callsign extraction performance very important
- If ATCo is sloppy, pilot is also and vice versa

MALORCA ANSPs	Prague	Vienna
ATCo utterance without callsign, **Ops**	9.6%	5.4%
ATCo utterance without callsign, **Lab**	3.1%	1.2%

Significant difference between lab and ops room trials.

Readback without Units

air france one two uniform foxtrot speed one eight zero knots

ATCo

AFR12UF SPEED 180 kt

one eighty air france one two uniform foxtrot

Pilot

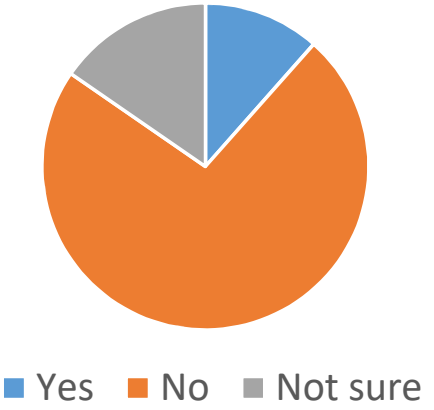
AFR12UF PILOT SPEED 180 none

1. Readback Error?
a. Yes, of course
b. No, never
c. Not sure, maybe

	ANSP1	ANSP2	ANSP3
ATCo utterance without unit	3%	20%	5%
Pilot utterance without unit	26%	42%	26%

Sloppy ATCo → sloppy pilot and vice versa

Workshop Results



HAAWAII-Stakeholder Workshop:
https://www.haawaii.de/wp/first-stakeholder-workshop-presentation-slides

Results from the Working Groups

Distinguish between

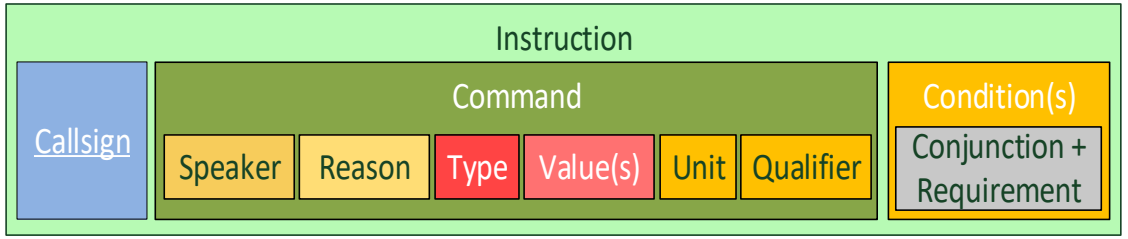
1. Readback errors **according to rules/manual**
2. Readback error that should be **brought to the controller's attention**
3. Readback error that should be **communicated with the pilot**

Contents in Detail

12:30 min

User Requirements

- Detection Rate > 50%:
  51 of 100 Readback Errors should be detected
- False Alarm Rate < 10%:
  From 100 alarms, ninety must be correct
- < 2% of the commands contain
  readback errors (seldom events)



A Readback Error (RBE) is correctly detected, if all ATC concept elements
- callsign, type, value, etc.

from **both** the ATCo utterance **and** pilot's readback are correctly recognized
and neither the ATCo's recognition nor the pilot's recognition is rejected (details follow).
Assuming independence, the recognition rate for the combined commands R_{both} is:

$$R_{Both} = R_{ATCo} * R_{Pilot}$$

R_{ATCo}: Command Recognition Rate for ATCo commands
R_{Pilot}: Command Recognition Rate for Pilot commands

Readback Error Detection Rate RD

$$RD = \frac{TP}{TP + FN} = \frac{RE * R_{both}}{RE * R_{both} + RE * (1 - R_{both})} = R_{both}$$

RD: Readback Error Detection Rate

TP: True Positive: Readback error is present which is **correctly** detected,
i.e., recognition correct and not rejected

FN: False Negative: Readback error is present, but is **falsely** classified as no readback error,
i.e., one recognition is wrong or one of the two recognitions is rejected

RE: Readback Error Rate (of pilot) (e.g. 2%)

R_{both}: ATCo and pilot command correctly recognized

The Readback Error Detection Rate is **independent of RE**, i.e. the readback error rate of the pilot.

MALORCA Prague: $\boldsymbol{R_{ATCo} = 92\%}$; $\boldsymbol{R_{Pilot}}$ = ?, Let assume 85% → R_{both} = 78%

Readback Error False Alarm Rate FA

$$FA = \frac{FP}{TP + FP} = \frac{(1 - RE) * E_{both}}{RE * R_{both} + (1 - RE) * E_{both}}$$

FA: Readback Error False Alarm Rate, also False Discovery Rate

TP: True Positive: Readback error is present which is **correctly** detected,
i.e., recognition correct and not rejected

FP: False Positive: No readback error is present, but is falsely classified as a readback error,
i.e., recognition wrong and no rejection for none or for both

RE: Readback Error Rate of pilot (e.g. 2%)

R_{both}: ATCo and pilot command correctly recognized

E_{both}: ATCo or pilot command are wrongly recognized and none of them rejected

The Readback Error Detection Rate **dependent on RE**, i.e. the readback error rate of the pilot.

Readback Error False Alarm Rate FA

$$FA = \frac{FP}{TP + FP} = \frac{(1 - RE) * E_{both}}{RE * R_{both} + (1 - RE) * E_{both}}$$

RE: Readback Error Rate of pilot (e.g. 2%)
R_{both}: 78%, R_{ATCo}: 92%, R_{Pilot}: 85%,
E_{both}: 2.6%, E_{ATCo}: 0.6%, E_{Pilot}: 2%,

$$E_{both} = E_{ATCo} + E_{Pilot} - E_{ATCo} * E_{Pilot}$$

$$62\% = \frac{(1 - 2\%) * (0.6\% + 2\% + 0.6\% * 2\%)}{2\% * 78\% + (1 - 2\%)(0.6\% + 2\% + 0.6\% * 2\%)}$$

Still 62% of false alarms.

1000 ATCO-pilot pairs, 20 readback errors,
15 correctly detected (TP); 5 undetected (FN); 24 False Alarms (FP); 956 no warning correct (TN)

Dependence of False Alarm Rate

R_{both} / E_{both}	0.1%	0.2%	0.3%	0.4%	0.5%	0.6%
98%	4.8%	9.1%	13.0%	16.7%	20.0%	23.1%
95%	4.9%	9.4%	13.4%	17.1%	20.5%	23.6%
90%	5.2%	9.8%	14.0%	17.9%	21.4%	24.6%
85%	5.5%	10.3%	14.7%	18.7%	22.4%	25.7%
80%	5.8%	10.9%	15.5%	19.7%	23.4%	26.9%
75%	6.1%	11.6%	16.4%	20.7%	24.6%	28.2%
70%	6.5%	12.3%	17.4%	21.9%	25.9%	29.6%
60%	7.6%	14.0%	19.7%	24.6%	29.0%	32.9%
50%	8.9%	16.4%	22.7%	28.2%	32.9%	37.0%
40%	10.9%	19.7%	26.9%	32.9%	38.0%	42.4%
20%	19.7%	32.9%	42.4%	49.5%	55.1%	59.5%
10%	32.9%	49.5%	59.5%	66.2%	71.0%	74.6%

→ Extremely good Command Recognition Error Rates are needed

→ Recognition Error Rates < 0.2% on Command Level

→Recognitions/Errors/Rejections

→Plausibility Measures

Contents in Detail

18 min

Plausibility Values on Word Level

climb flight level one one zero enzo one one zero

For a human challenging, for the ASR system not
"climb flight level one one zero one one zero"
Interpretation easy, because we have an ENZ110 in the air

Readback Error False Alarm Rate FA

ASR output: fraction five quebec juliett descend altitude four thousand feet be level by a that sir

Ground Truth: fraction five quebec juliett descend altitude four thousand feet be level by evata

```
thousand feet   be level   by    a that  sir
valu unit unkn  unkn unkn unkn unkn unkn unkn
                          ################
```

If words could not be classified, reduce plausibility of extraction.

Using Surveillance Data

one two eight tango x-ray contact radar one three four decimal three five five

Gold:
four two eight tango x-ray contact farnborough radar one three four decimal three five five

N428TX CONTACT RADAR
N428TX CONTACT_FREQUENCY 134.355

“november” is blocked out by late PTT.

From surveillance data you know, that a N428TX is in the air.

Using Surveillance Data for Manual Transcription

Callsign Recognition and Callsign Error Rate, when extracted from manual transcribed data

all rates in [%]	ANSP 1	ANSP 2	ANSP3
Surveillance data used	99 / 0.5	**99 / 0.5**	99 / 0.5
Surveillance data ignored	84 / 12	**75 / 15**	84 / 6

Dramatic improvement when surveillance data is available.

The whole command is considered: Command Recognition/Error Rate

all rates in [%]	ANSP1	ANSP2	ANSP3
Surveillance data used	99 / 0.6	99 / 1.2	100 / 0
Surveillance data ignored	83 / 5	79 / 9	84 / 10

Error Rate decreases, because callsign often not extracted, therefore command rejected

Benefits of Surveillance Data for Automatic Transcriptions

Callsign Recognition and Callsign Error Rate, when extracted from automatic transcribed data, after using six hours of training data

all rates in [%]	ANSP1	ANSP2	ANSP3
Surveillance data used	95 / 3	no	85 / 10
Surveillance data ignored	79 / 13	data	60 / 19

The whole command is considered: Command Recognition/Error Rate

all rates in [%]	ANSP1	ANSP2	ANSP3
Surveillance data used	86 / 5	no	69 / 12
Surveillance data ignored	72 / 10	data	49 / 23

The data just shows the trend, just one hour of data used for testing.

Contents in Detail

1. Use Cases
2. What is a readback error (and what not)?
3. Readback Error Detection Rates and False Alarm Rates
4. Plausibility Measures
5. **Conclusions**

24 min

Conclusions

- Readback Error Detection Assistant includes many use cases, e.g. missing readbacks
- No common agreement, concerning what is a readback error
  - According to the book
  - Interesting for ATCo
  - Interesting for the pilot
- False Alarm Rate < 10% requires Command Recognition Error Rate < 0.2% with Command Recognition Rate > 90%
- plausibility values or/and redundant REDA engines
- Usage of Surveillance Data increases Command Recognition Performance

Thank you very much for staying in the conference.

Hartmut Helmke, Project Leader, DLR
Hartmut.Helmke@dlr.de
www.haawaii.de

SESAR JOINT UNDERTAKING

Founding Members
EUROPEAN UNION EUROCONTROL