HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

How to Measure Speech Recognition Performance in the Air Traffic Control Domain? The Word Error Rate is only half of the truth!

Hartmut Helmke, Shruthi Shetty, Matthias Kleinert, Oliver Ohneiser, Heiko Ehr,
German Aerospace Center (DLR)

Amrutha Prasad, Petr Motlicek,
Idiap Research Institute

Aneta Cerna,
Air Navigation Service Provider Czech Republic

Christian Windisch,
Austro Control (ACG)

SESAR JOINT UNDERTAKING

Founding Members
EUROPEAN UNION EUROCONTROL

Contents in Detail

1. *Buildings blocks of ASR applications for ATC*
2. *Requirements beyond Word Error Rates*
3. *Metric “Command Recognition Rates” etc.*
4. *Results*
5. *Conclusions*

From the Speech Signal to Benefits to the Air Traffic Controller (ATCo)

Assistant Based Speech Recognition

Surveillance Data,
Flight Plan Data
Weather Data

hello good morning
turkish five kilo juliet
reduce one sixty or greater
descent three thousand feet

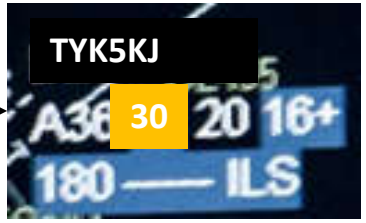
Aircraft: THY5KJ, DLH3ER
Commands:
THY5KJ DESCEND 3000 ft
THY5KJ REDUCE 180 kt …

Command Prediction

ATCO
Pilot

Voice To Text

Concept Recognition

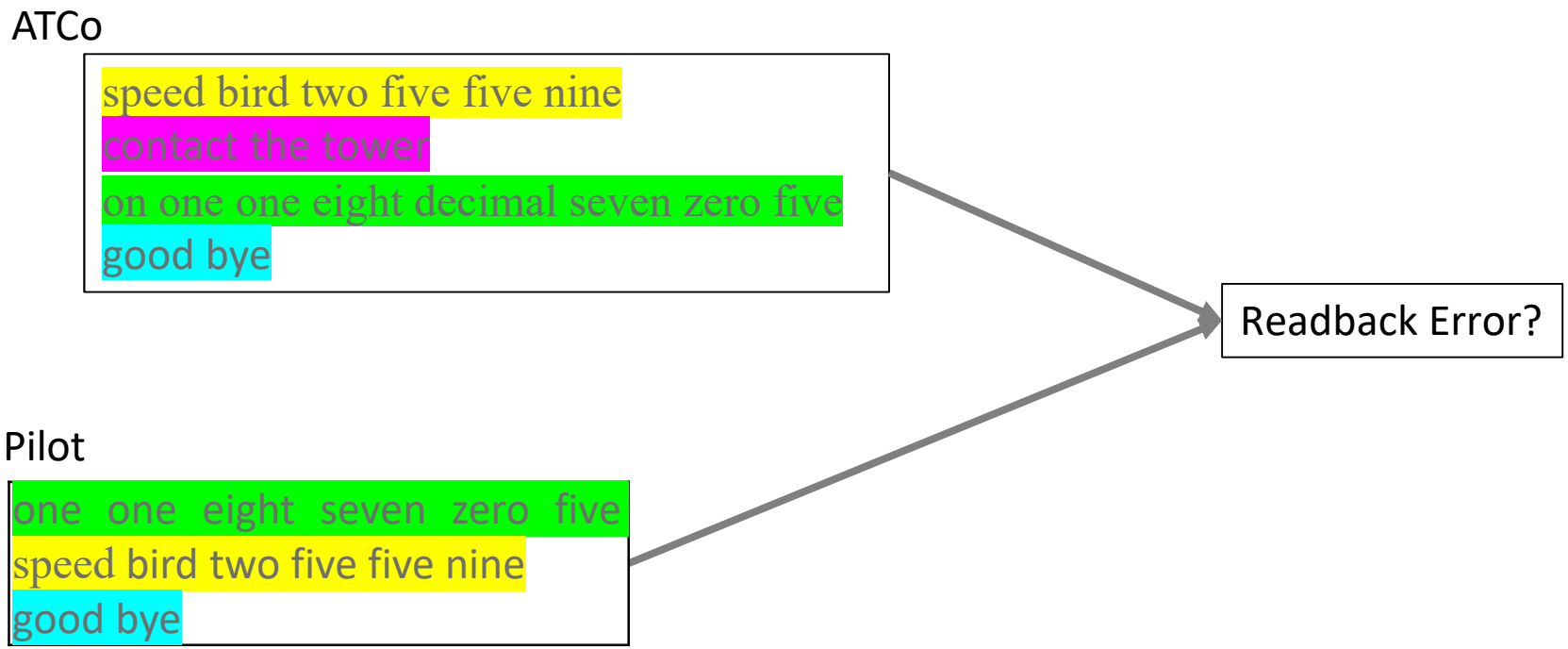


THY5KJ REDUCE 160 none OR_GREATER
THY5KJ DESCEND 3000 ft

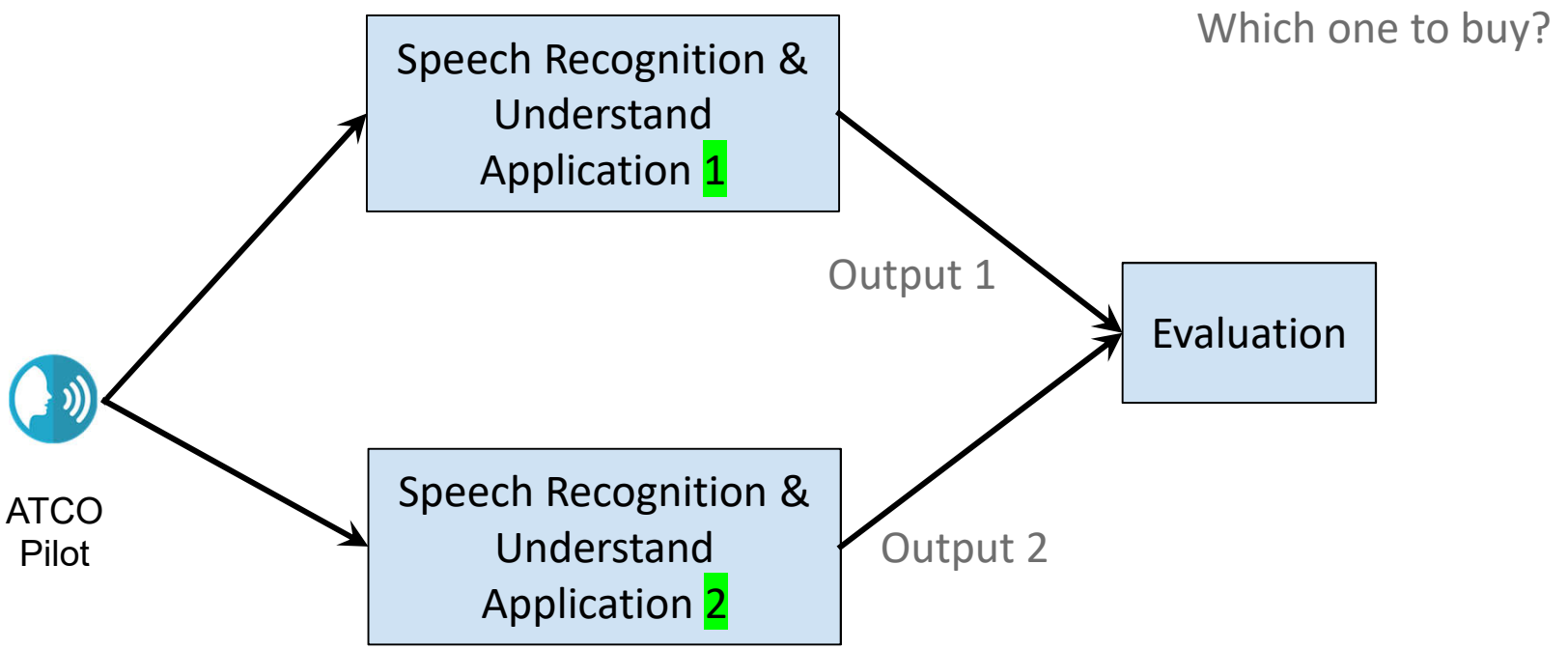
Readback error detection,
Workload prediction …

ATCO = Air Traffic Controller

Readback Error Detection (Real)



The Initial Question: Which one to buy?

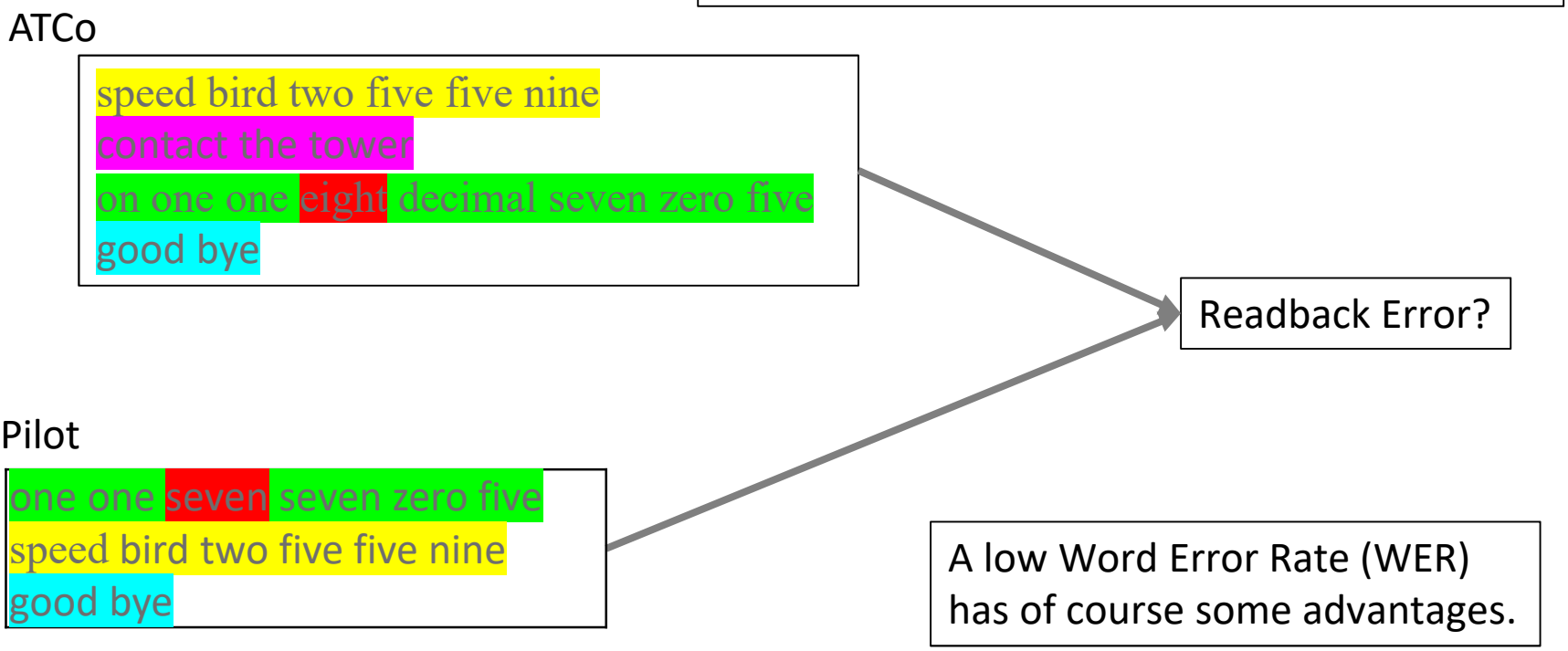


Contents in Detail

1. *Buildings blocks of ASR applications for ATC*
2. *Requirements beyond Word Error Rates*
3. *Command Recognition Rates etc.*
4. *Results*
5. *Conclusions*

Readback Error Detection (Real)

The user needs
- a high read back detection rate (accuracy)
- Low false alarm rate (false positives)



A low Word Error Rate (WER) has of course some advantages.

Readback Error Detection (Real)

Speed Recognition is NOT
Speech Understanding
Alan Turing 1952

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

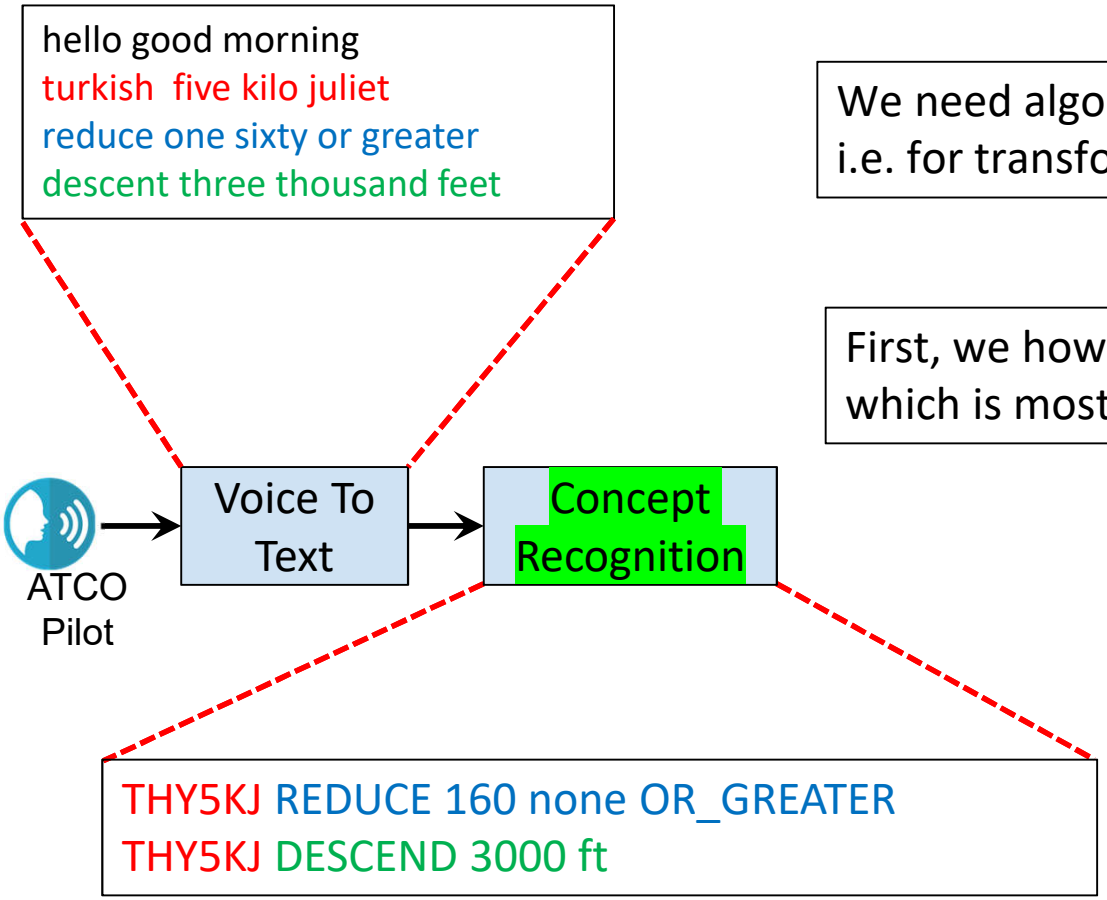
Readback Error?

Pilot

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

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

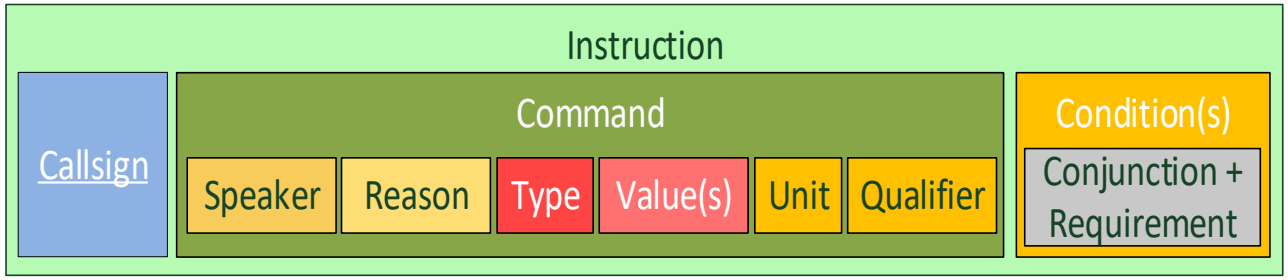
From the Speech Signal to Benefits to the Air Traffic Controller (ATCo)



We need algorithms for language understanding, i.e. for transformation from words to semantics

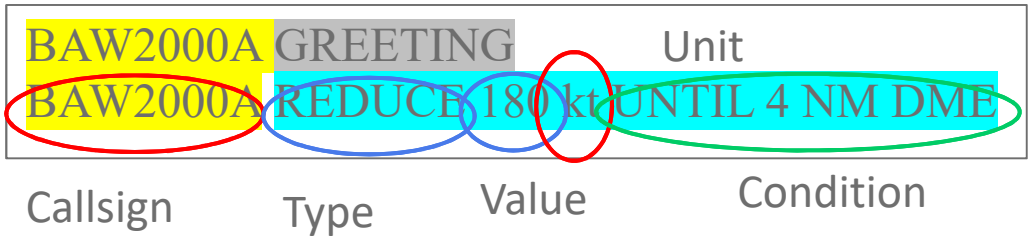
First, we however need to agree on the output, which is mostly done on word level.

From Words to ATC Concepts



good morning speed bird two zero zero zero alfa
reduce one eight zero knots until DME four miles

Speed Recognition is NOT
Speech Understanding
Alan Turing 1952



Contents in Detail

1. *Buildings blocks of ASR applications for ATC*
2. *Requirements beyond Word Error Rates*
3. *Command Recognition Rates etc.*
4. *Results*
5. *Conclusions*

From Words to ATC Concepts

BAW2000A REDUCE 180 kt UNTIL 4 NM DME

A command is correctly recognized, IFF both
the callsign,
the type,
the second type
the value,
the unit,
the qualifier,
the condition,
the speaker, (pilot, ATCO) and
the reason (command, readback, request, reporting)
are correct!!!
Otherwise it is a command recognition error or a rejection.

Command Recognition Errors

BAW2000A REDUCE 180 kt UNTIL 4 NM DME

BAW2000A REDUCE 170 kt UNTIL 4 NM DME (error)

NO_CALLSIGN REDUCE 170 kt UNTIL 4 NM DME (rejection)

BAW2000A NO_CONCEPT (rejection)

BAW2000A REDUCE 170 kt UNTIL 5 NM DME (one error)

BAW2000A REDUCE 180 kt UNTIL 5 NM DME (error)

BAW2000A REDUCE 180 kt OR_ABOVE UNTIL 4 NM DME (one error)

Transform annotation into one "big word" and calculate the Levenshtein distance.

Metrics

Transform annotation into one "big word” and calculate the Levenshtein distance.

Metric	Definition
Command Recognition Rate (RcR)	RcR = #matches / #gold

Metrics

Transform annotation into one "big word" and calculate the Levenshtein distance.

Metric	Definition
Command Recognition Rate (RcR)	RcR = #matches / #gold
Command Recognition Error Rate (ErR)	ErR = (subs + ins) / #gold
Command Rejection Rate (RjR)	RjR = del / #gold
Callsign Recognition Rate (CaR)	Same as RcR but only for callsigns without instructions
Callsign Recognition Error Rate (CaE)	Same as ErR, but only for callsigns without instructions
Callsign Rejection Rate (CaRj)	Same as RjR, but only for callsigns without instructions

Contents in Detail

1. *Buildings blocks of ASR applications for ATC*
2. *Requirements beyond Word Error Rates*
3. *Command Recognition Rates etc.*
4. *Results*
5. *Conclusions*

Results from Perfect Annotations
WER 0%

	#Cmd	#Utt	RcR	ErR	CaR
Ops Prague	6094	3038	98.5%	0.9%	99.8%
Lab Prague	6885	4211	99.2%	0.5%	99.7%
Ops Vienna	4417	2279	94.8%	4.0%	98.2%
Lab Vienna	6005	3562	95.3%	2.5%	96.4%

#Cmd: Number of given commands
#Utt: Number of utterances/ files,
an utterance contains between 1 and 7 commands

RcR: Command Recognition Rate
ErR: Command Recognition Error Rate

CaR: Callsign Recognition Rate

Results from ASR Output for Prague WER > 0%

	RcR	CaR	WER
Ops Prague, gold transcription	98.5%	99.8%	0.0%
Ops Prague, no callsign context for ASR	96.5%	98.7%	2.3%
Ops Prague, callsign context for ASR	96.6%	98.2%	2.8%
Ops Prague, bad speech model	76.8%	88.5%	13.5%

RcR: Command Recognition Rate
CaR: Callsign Recognition Rate
WER: Word Error Rate

Results from ASR Output for Vienna WER > 0%

	RcR	CaR	WER
Ops Prague, gold transcription	98.5%	99.8%	0.0%
Ops Prague, no callsign context for ASR	96.5%	98.7%	2.3%
Ops Prague, callsign context for ASR	96.6%	98.2%	2.8%
Ops Prague, bad speech model	76.8%	88.5%	13.5%

	RcR	CaR	WER
Ops Vienna, gold transcription	94.8%	98.2%	0.0%
Ops Vienna, no callsign context for ASR	89.9%	93.0%	5.1%
Ops Vienna, callsign context for ASR	88.6%	91.6%	6.7%
Ops Vienna, bad speech model	82.7%	87.8%	9.5%

RcR: Command Recognition Rate
CaR: Callsign Recognition Rate
WER: Word Error Rate

Contents in Detail

1. *Buildings blocks of ASR applications for ATC*
2. *Requirements beyond Word Error Rates*
3. *Command Recognition Rates etc.*
4. *Results*
5. *Conclusions*

Recommendations & Conclusions

- *Ontology from SESAR-2 16-04 solution updated for pilots*
- *Implementation of the ontology rules available, accuracy >95%*
- *Robust against errors from Speech-to-Text*
- *Metric of Command Recognition Rate not new, but the definition of the gold annotation itself*
- *WER gives first hints*

DLR

BRNO UNIVERSITY OF TECHNOLOGY

idiap RESEARCH INSTITUTE

austro CONTROL

isavia ANS Air Navigation Services

NATS

CROATIA CONTROL

Thank you very much for listening!

SESAR JOINT UNDERTAKING

Founding Members

EUROPEAN UNION EUROCONTROL