HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

Measuring Speech Recognition And Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates

Hon. Prof. Dr. Hartmut Helmke (DLR)

SESAR Innovation Days

Motivation

Comment from the reviewers:

A question that I got after reading this paper, one that I think should be stated/dealt with up-front, is

what accuracy do we need to safely introduce this into the ATM system?

Of course, 100% is desired, but we are not quite there yet I read. So what do we consider as a target accuracy before we can seriously think of introducing this new technology into our operations?

Answer:

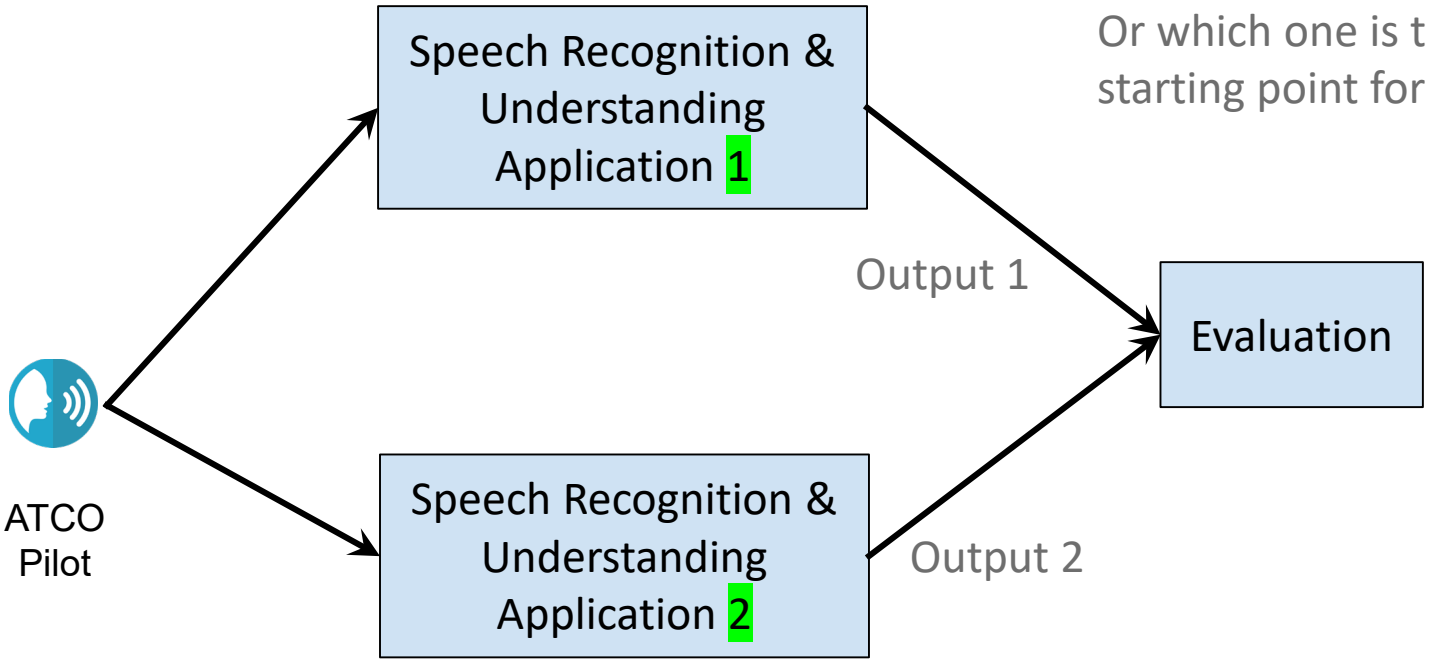
Yellow accuracy is enough, pink might be not enough.

Useless answer,

but currently there we are.

We have at least difficulties to compare two speech recognizers (on application level)!

Motivation



Which one to buy?
Or which one is the starting point for improvements?

HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

Measuring Speech Recognition And Understanding Performance in Air Traffic Control Domain Beyond Word Error Rates

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)

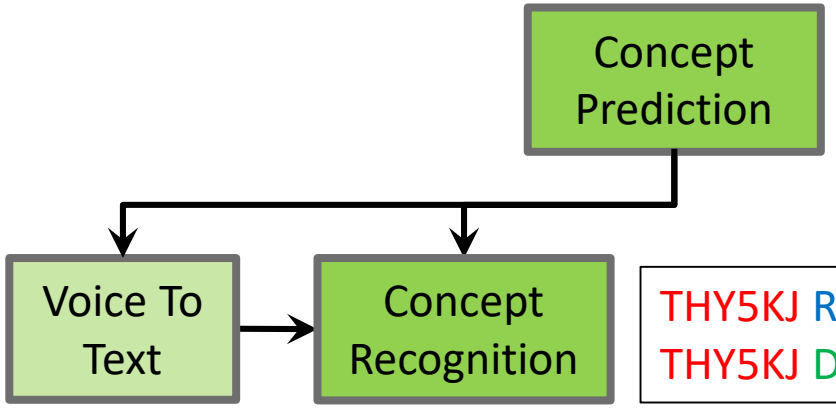
Contents in Detail

HAAWAII

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

ABSR: From Words to ATC Concepts

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



THY5KJ REDUCE 160 none OR_GREATER
THY5KJ DESCEND 3000 ft

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

Contents in Detail

Readback Error Detection (Real)

HAAWAII

Speech 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

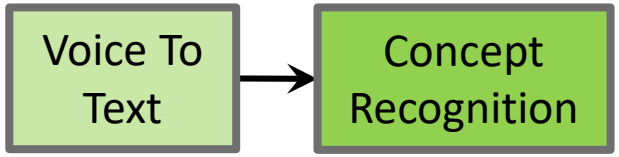
Readback Error?

Pilot

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

- Word sequences are different
- Not every command requires a readback
- Sequence of commands can be different
- “eighteen” and “one one eight” are the same
- “thousand” and “zero zero zero” are the same

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



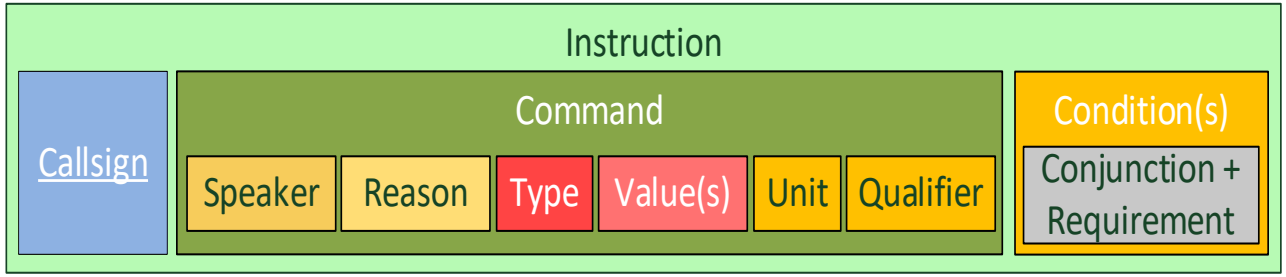
THY5KJ REDUCE 160 none OR_GREATER
THY5KJ DESCEND 3000 ft

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

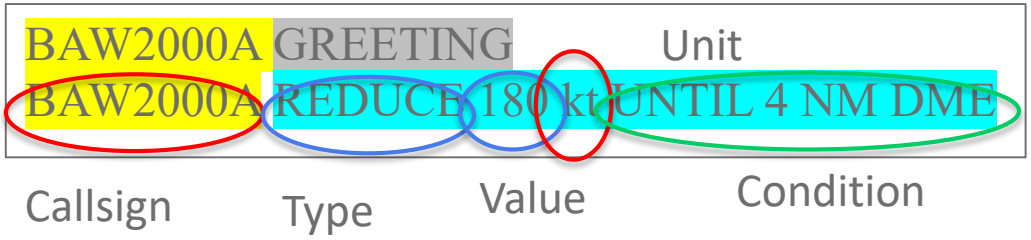
We need algorithms for language understanding, i.e. for transforming words to semantics

However, we first need to agree on the output format, which is mostly done on word level, but NOT on semantic level.

From Words to ATC Concepts



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



Contents in Detail

From Words to ATC Concepts

BAW2000A REDUCE 180 kt UNTIL 4 NM DME

A command is correctly recognized, IFF

- 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

Gold/correct annotation

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

Example

Gold Annotation	Command Extraction
AFR123 INIT_RESPONSE AFR123 TURN LEFT AUA1AB SPEED 140 kt DLH123_NO_CONCEPT	AFR123 DIRECT_TO OKG none AFR123 INIT_RESPONSE AFR123 TURN RIGHT AUA1AB NO_CONCEPT DLH123 NO_CONCEPT
Result:	
RcR = 2/4 = 50% (green)	ErR = 2 / 4 = 50% (purple)
RjR =1/4 = 25% (yellow)	

- Callsign recognition rate is 100%,
- Callsign recognition error rate is 0%
- Callsign rejection rate is 0%

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

Dependency of Readback Error False Alarm Rate on Command Recognition and Error 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%

Requirements:
Command Recognition Rate > 50% enough (for ATCo and pilot together)
Command Recognition Error Rate < 1%

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

HAAWAII

- 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 annotations itself is new
- WER (= word error rate) gives first hints

- Command Recognition Rates > 90% for Radar Label Maintenance (for the relevant command types)
- Command Recognition Error Rates < 1% for a REDA (= Readback Error Detection Assistant)

More information on

www.haawaii.de

Thank you very much for listening!

SESAR JOINT UNDERTAKING

Founding Members

EUROPEAN UNION EUROCONTROL