HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

Machine learning and AI

Technical details on automatic speech recognition applied in ATC

First results in HAAWAII project

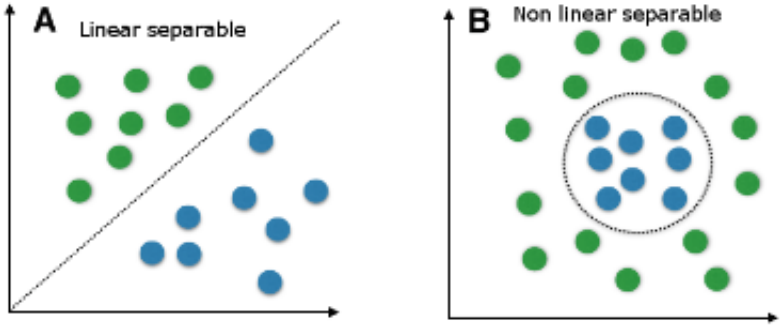
Petr Motlicek, Pavel Smrz
Idiap Research Institute, Martigny, Switzerland
Brno University of Technology, Czech Republic

SESAR JOINT UNDERTAKING

co-funding Members

Introduction - ML & AI in current technologies

- Machine learning is only one part of AI
- HAAWAII relies on data-driven approach (although in an ideal case - a mix of large amount of data with expert knowledge)



- Our focus is ATC - understand the ATC communication (both channels)
  - General objective is to decrease a workload, decrease a probability of potential incidents due to miscommunication
  - Automatic speech recognition should be seen as a part of the solution

Content

(1) Importance of data
(2) What is automatic speech recognition (ASR)
(3) Development of ASRs
    (a) Rapid domain adaptation
    (b) learning w/o teacher - semi-supervised approach
    (c) Boosting on “unseen/rare words”, dynamic insertion
(4) Natural language understanding
(5) Performance on HAAWAII data

1. Importance of data

- Similar to any other machine learning approach - data is a key factor (in addition to expert knowledge) in building robust ASR
- The best commercial systems (e.g. GAFA) are trained on 100k's hours of speech
  - However, as these services are mostly free, end-users are "a part of services"
  - Their adaptation to target domain requires significant expert knowledge
- ASR can be developed using out-of-domain data (e.g. broadcast), but the robustness will not be high
  - Therefore, domain data is important to exploit
- Possibility to train with or w/o teacher
- Error-rates ~= 0 is practically impossible to reach (e.g. human error in verifying data)

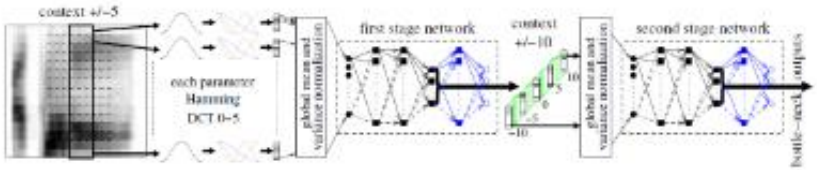
2. What is automatic speech recognition (ASR)

- Speech-to-text transformation, without extracting any meaning (called ASR)
- It also includes other parts:
  - Speaker diarization/segmentation
  - Paralinguistic pathological speech
  - Voice activity detection, or speech-non-speech detection?

2. Cont.

Automatic Speech Recognizer

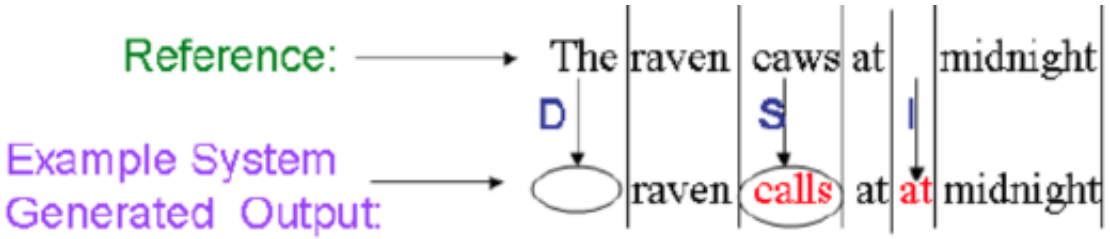

```
{ "response":
  { "task_id": "c2596dfc-057c-4615-83a2-c3586bccbc18",
    "task_type": "asr",
    "segments": [
      { "metadata":
        { "non_speech_ratio": "0",
          "type": "speech"
        },
        "nbests": [
          { "words": [
            ["text": "Do", "start": "0", "end": "1290", "confidence": "0.78"},
            ["text": "you have", "start": "1290", "end": "1630", "confidence": "0.92"],
            ["text": "a", "start": "1630", "end": "1700", "confidence": "1.00"],
            ["text": "suspect", "start": "1700", "end": "2610", "confidence": "1.00"],
            ["text": ".", "start": "2610", "end": "2620", "confidence": "1.00"]
            ],
            "confidence": "0.94",
            "transcription": "Do you have a suspect."
          }
        ]
      }
    ],
    "global_confidence": "0.94"
  }
}
```

2. Cont.

ASR evaluations:

$$\text{WER} = 100 \cdot \frac{S + D + I}{N}\% \qquad \text{Accuracy} = 100 - \text{WER}\%$$



$$\text{WER} = \frac{(1+1+1)}{5} = \frac{3}{5}$$

N - #words in reference transcripts, S, D, I - substitutions, deletions, insertions

Dynamic programming applied

But we need more: WER is not directly related to to higher level information extraction - we use also other metric

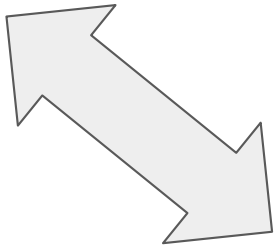
2. Cont.

```
Per utt details:
2020-08-01__14-57-06-59
speed bird six six golf alfa reduce speed two two zero knots or less
2020-08-01__14-57-10-46_P
one sixty knots to four dme kenya one hundred
2020-08-01__14-57-42-43
speed bird two five five nine contact the tower on one one eight decimal seven zero five good bye
2020-08-01__14-57-47-34_P
one one eight seven zero five good bye speed bird two five five nine
2020-08-01__14-58-25-45_P
descend flight level seven zero fly heading three six zero degrees > speed roger > bird six speed bird five one five
2020-08-01__14-58-32-73_P
director hello kenya one hundred descending level eight zero bravo seven eight eight information foxtrot direct biggin > <gbg>
2020-08-01__14-58-38-67
kenya one hundred heathrow > easy two hello hold at > it biggin it'll be > <gbg> a > serbia five minute > min no delay
2020-08-01__14-58-42-64_P
hold > london at biggin > <gbg> kenya > now one hundred
2020-08-01__14-59-22-99
speed bird five one five turn right heading zero nine zero degrees
2020-08-01__14-59-26-46_P
speed zero > bird nine zero degrees speed bird five one five

WER: 16.8 (ins 4, del 6, sub 12 / 131)
SER: 70.0

Insertions:
six     1
two     1
no      1
speed   1

Deletions:
on      1       1
it'll   1       1
at      1       2
biggin  1       3
one     2       12

Substitutions:
roger > bird    1       1
heathrow > easy 1       1
be > <gbg>      1       1
a > serbia      1       1
minute > min    1       1
biggin > <gbg>  2       3
at > it 1       2
hold > london   1       2
```

3. Development of ASR

Kaldi toolkit

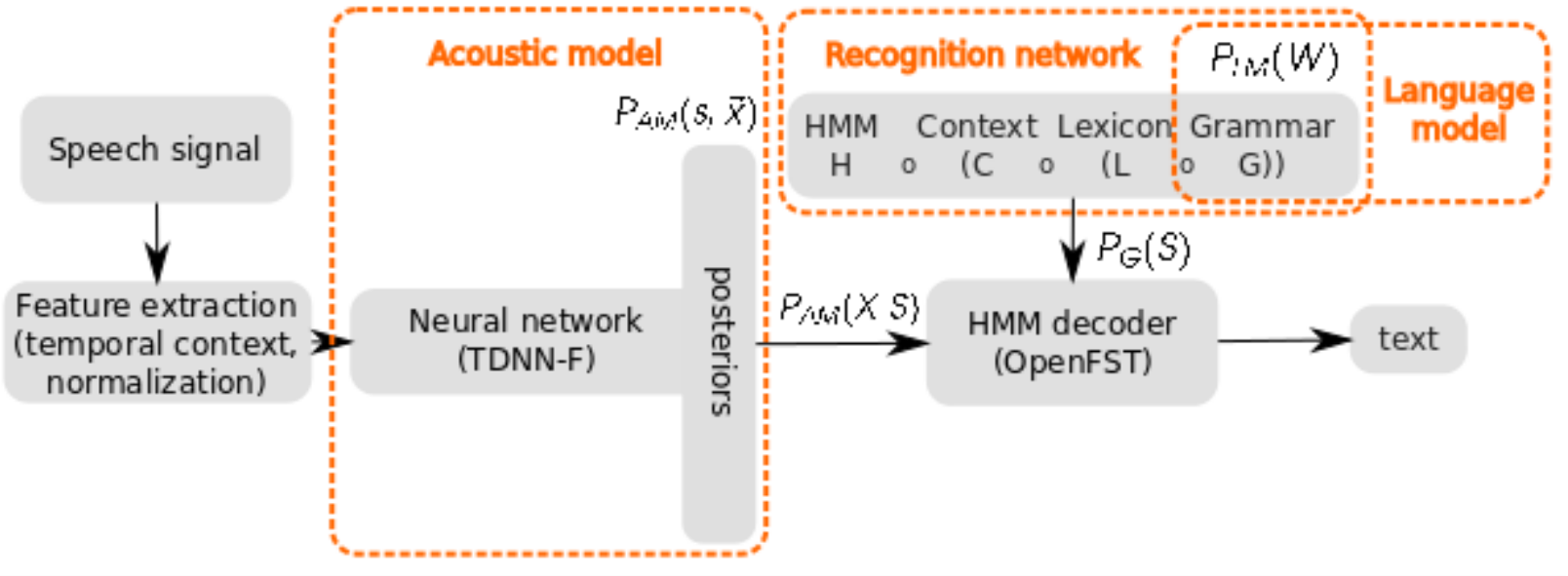
A toolkit for speech recognition research



(According to legend, Kaldi was the Ethiopian goatherd who discovered the coffee plant).

3. Cont.

Current KALDI - ASR state-of-the-art system



3. Cont. (rapid development of ASR)

Day	System for new language	WER%	Relative improvement
2	Baseline (Monolingual, TDNN-Factorized, LF-MMI)	~41	TBD
5	+ Lexicon augmentation	~35	TBD
7	+ Use of target data (30h)	~30%	TBD
8	+ Learning without teacher (600h) (*Single system*)	~25%	TBD
9	+ Customized LM	~20%	TBD
10	+ Boosting of specific words	~15%	TBD

- Amount of improvement for various domains may not be the same

3. Cont.

Incremental learning without teacher - semi-supervised training

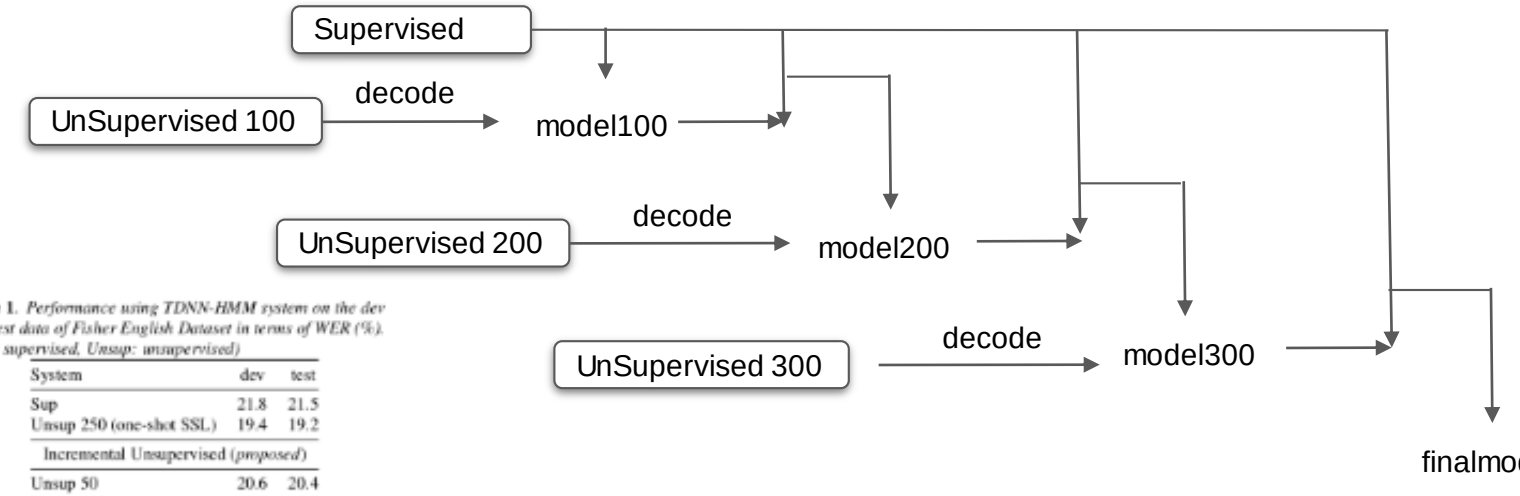
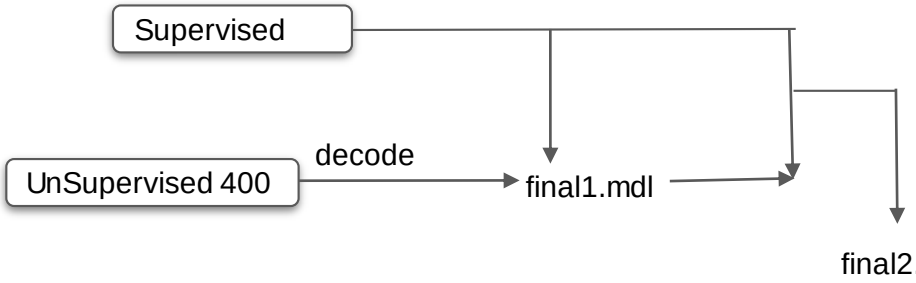
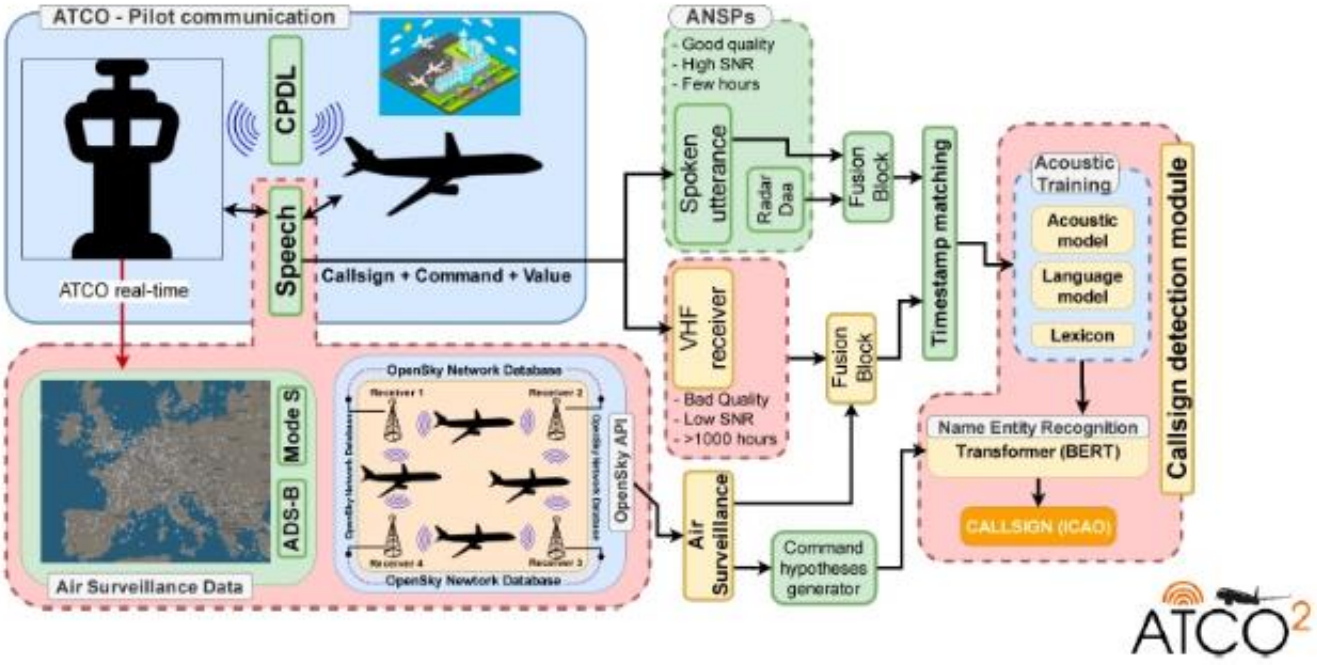


Table 1. *Performance using TDNN-HMM system on the dev and test data of Fisher English Dataset in terms of WER (%). (Sup: supervised, Unsup: unsupervised)*

System	dev	test
Sup	21.8	21.5
Unsup 250 (one-shot SSL)	19.4	19.2
Incremental Unsupervised (*proposed*)		
Unsup 50	20.6	20.4
Unsup 100	19.7	19.2
Unsup 150	**19.1**	**19.0**
Unsup 200	**18.8**	**18.5**
Unsup 250	**18.6**	**18.3**
Oracle 250	17.7	17.5

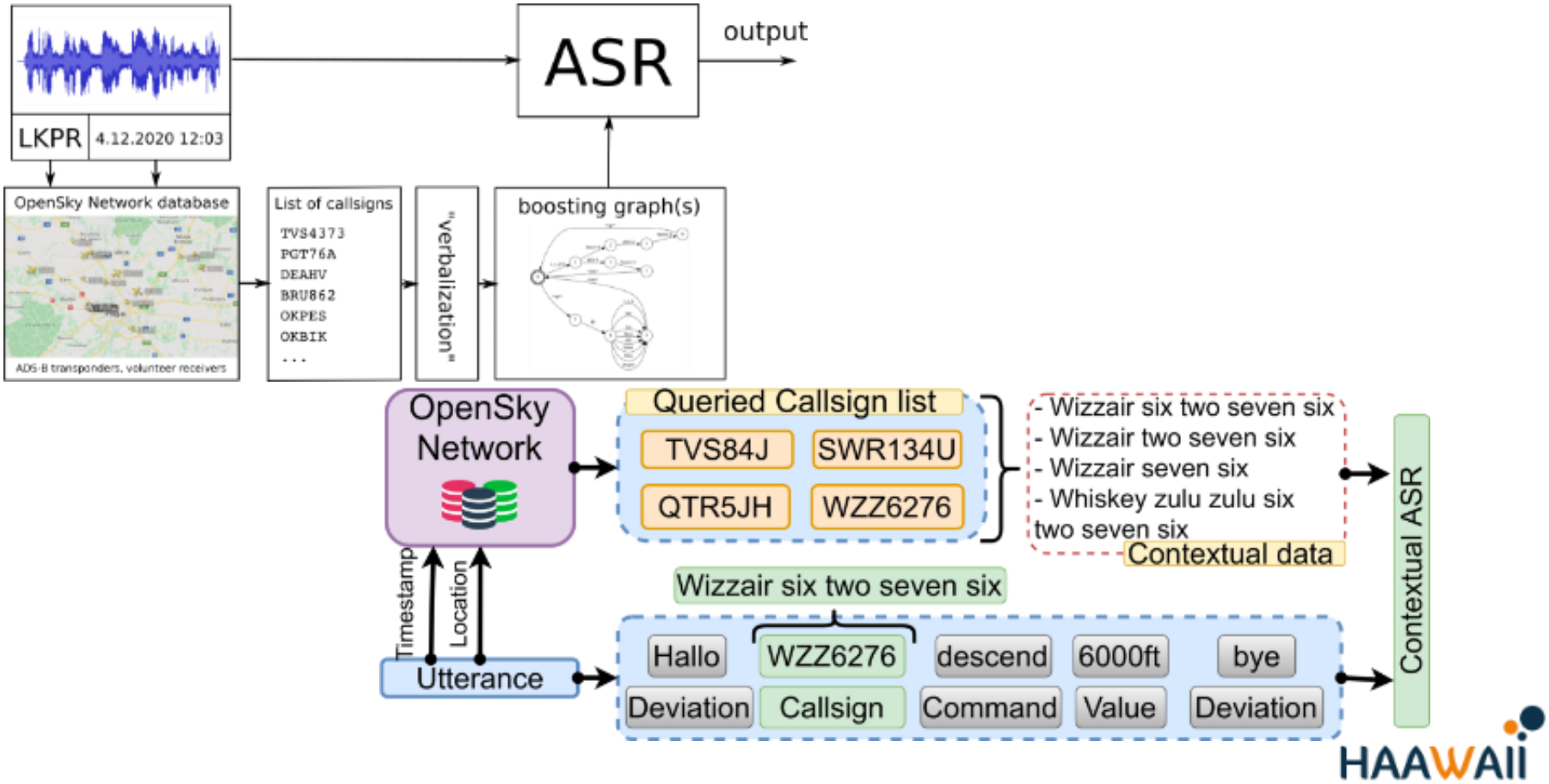


3. Cont. - current version of the ASR



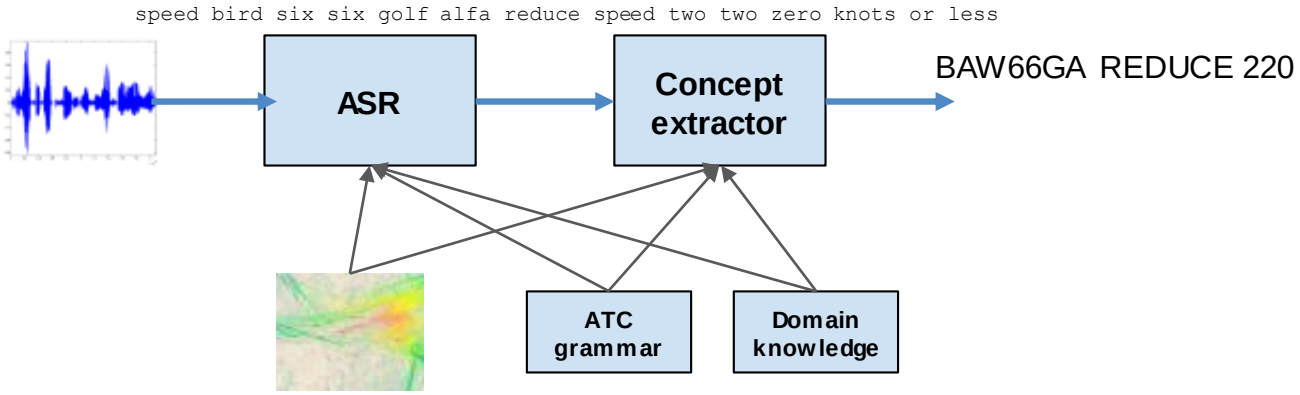
3. Cont.

Boosting of important words: callsign boosting



4. Natural language understanding

- Recognition is only a part of understanding
- ASR → NLU pipeline:
  - How to assess ASR impacting NLU
  - Impact of ASR errors on NLU
- NLU i ATC: concept extraction

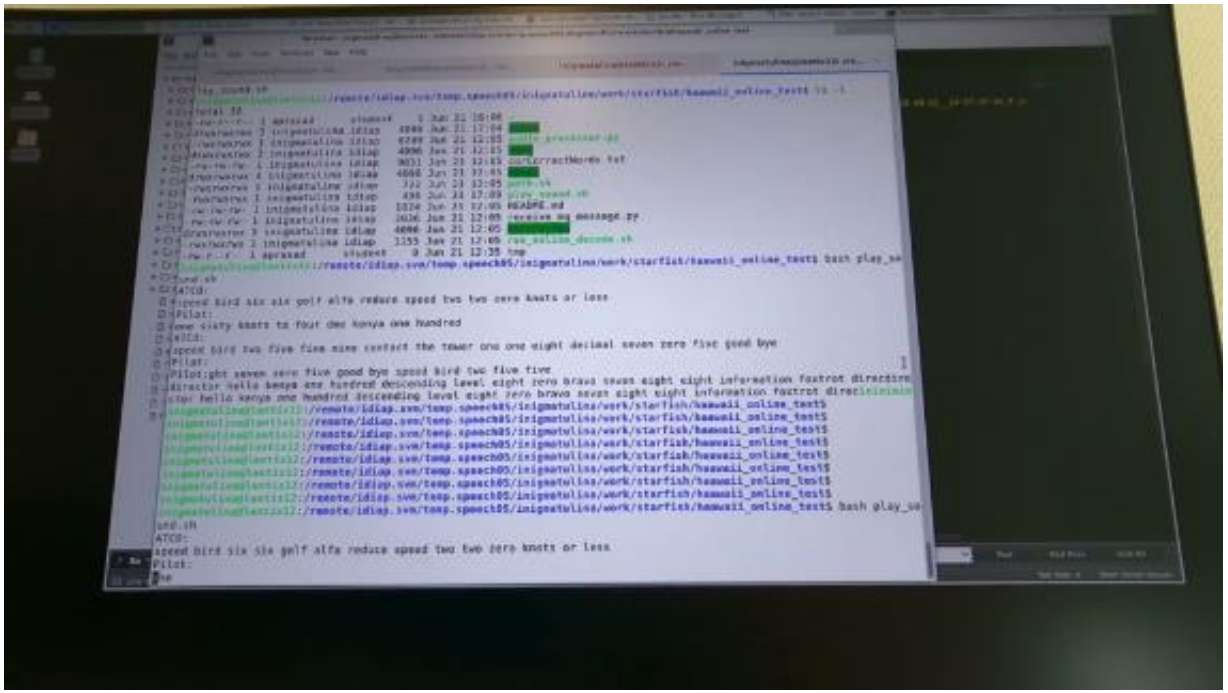


5. Performance on HAAWAII

ASR system	Results [WER]	Command accuracy
Preliminary results (beg. Of the project)	~30%	??
Pilot	11%	??
ATCO	8%	??

5. Cont.

Video: real-time ASR engine applied on HAAWAII data



Cnclusions

Special session: ASR in ATC

- Special session on the use of automatic speech recognition in ATM will be organised
- INTERSPECH 2021 conference, 30/8 - 3/9/2021 in Brno, Czech Republic
  - Hybrid organisation mode
- https://www.interspeech2021.org
- Interspeech is the major international conference focusing on speech processing R&D
  - Around 2'000+ participants, around 10+ tracks with talks and posters
- https://www.interspeech2021.org/call-for-special-sessions-and-challenges

HAAWAII
Highly Advanced Air Traffic Controller Workstation with Artificial Intelligence Integration

Thanks

This project has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 884287.