MALORCA

Machine Learning of Speech Recognition Models for Controller Assistance

MALORCA 2nd Stakeholder Meeting

Proof-of-Concept Trials

Hon. Prof. Hartmut Helmke

Vienna, 20th to 21st of February 2018

SESAR JOINT UNDERTAKING

Founding Members

EUROPEAN UNION EUROCONTROL

Abstract

Proof-of-concept of MALORCA project is split into two technical (T1, T2) and two operational (O1, O2) activities.

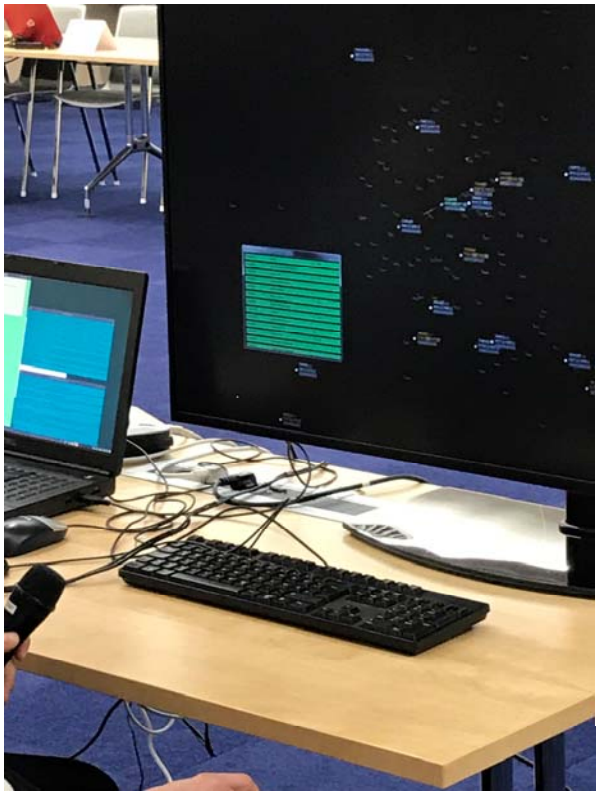
- T1 is a workshop with technical experts to evaluate the ABSR prototype implementation against the technical requirements.
- T2 is an offline evaluation to quantify the improvements of the ABSR system with respect to the amount of available training data.
- O1 involves controllers who concentrate only on the different outputs of a baseline ABSR system and on an ABSR system trained with all the available MALORCA training data.
- O2 puts the trained ABSR system in a simulation environment with a replay of historic radar data and controller voice recordings from real Prague and Vienna in- and out-bound traffic. ABSR is used here to support the controllers in maintaining radar labels.
- Debriefing including questionnaire

Contents

- T1: workshop to evaluate the ABSR against the requirements.
- O2: controller uses ABSR system for radar label maintenance
- O1: Baseline and trained system:
  36 different situations with always 2 different outputs
  Which one is better?
- T2: Simulate the effect of monthly adding more and more training data

T1 – Technical Validation 1

Test location and platform: Prague, live-mode, provided by DLR
Date: 23-24.1.2018



- Let the ATCos play with live-mode
- To verify ABSR level of readiness with respect to original ANSP requirements for ABSR
- To be familiar with current status of ABSR that has been reached prior to O1 and O2
- To be able to identify the starting point for upcoming projects (PJ16-04,..)

Test Scenario 1

- Participants: project partners' representatives
- Reference: Generic ASR User requirements - **D1-2 SRS**

Test Scenario 2

- Participants: project partners 'representatives
- Reference: Technical requirements - **D1-2a SRS Annex**

O2 – Operational Validation

Voice utterances were presented to controller together with the online recognition results

Audio,
radar data
Recognized commands

Is recognition quality sufficient to support radar label maintenance?

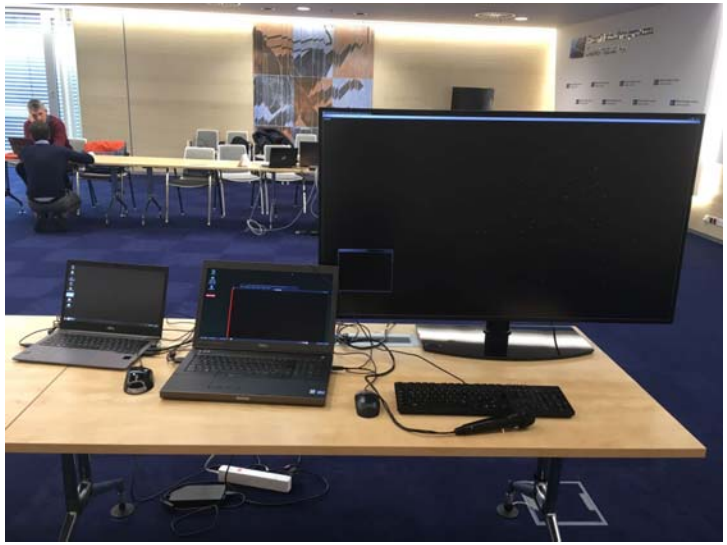
© DLR

O2 – Operational Validation Results

Prague:			
Number of Commands	Number of ABSR Errors/Rejections	corrected by controller	detected by controller
396	36	31	36

Vienna:			
Number of Commands	Number of ABSR Errors/Rejections	corrected by controller	detected by controller
610	80	79	80

T1 / O2 Some Findings



- Reaction time not real-time and not on-line
  Missing pilot read back
- Incorrectly assigned commands:
  e.g. “Intercept localizer“ (all words correctly recognized, but output to user CLEARED_ILS Commands
  reason only 1 time in transcribed training data of > 5000 commands in total
- STOP_DESCEND not modelled
  reason: never occurred in transcribed training data
- MAINTAIN_ALTITUDE recognized, but sometimes not shown in HMI
  64 times recognized in untranscribed training data
  reason: Was not expected by command predictor in test mode
- Noise of audio data :
  Real data shows the difference in “signal to noise” ratio (17 VIE, 22 PRG), which correlates to ASR performance, clearly visible in T2 as well.

Commands supported for Prague Approach

DESCEND
CLIMB
REDUCE
TURN_LEFT_HEADING
TURN_RIGHT_HEADING
TRANSITION
DIRECT_TO
CLEARED_ILS
HANDOVER
HANDOVER_FREQUENCY
INTERCEPT_LOCALIZER
QNH
INFORMATION
CLEARED_NDB

CLEARED_RNAV
INIT_RESPONSE
INTERCEPT_GLIDEPATH
STOP_DESCEND
STOP_CLIMB
ALTITUDE
MAINTAIN_ALTITUDE
REDUCE_OR_BELOW
REDUCE_NOT_BELOW
SPEED
SPEED_OR_ABOVE
SPEED_OR_BELOW

MAINTAIN_SPEED
MAINTAIN_SPEED_OR_BELOW
MAINTAIN_SPEED_OR_ABOVE
MAINTAIN_HEADING
HEADING
REDUCE_MIN_APPROACH_SPEED
STOP_ALTITUDE
SPEED_OWN
NO_CONCEPT
SQUAWK
REPORT_ESTABLISHED

Blue, < 10 training samples

Findings in T1 / O2

DESCEND
CLIMB
REDUCE
TURN_LEFT_HEADING
TURN_RIGHT_HEADING
TRANSITION
DIRECT_TO (*, some WP)
CLEARED_ILS
HANDOVER
HANDOVER_FREQUENCY
INTERCEPT_LOCALIZER (*)
QNH
INFORMATION
CLEARED_NDB

CLEARED_RNAV
INIT_RESPONSE
INTERCEPT_GLIDEPATH
STOP_DESCEND (*)
STOP_CLIMB
ALTITUDE
MAINTAIN_ALTITUDE (*)
REDUCE_OR_BELOW
REDUCE_NOT_BELOW
SPEED
SPEED_OR_ABOVE
SPEED_OR_BELOW

MAINTAIN_SPEED
MAINTAIN_SPEED_OR_BELOW
MAINTAIN_SPEED_OR_ABOVE
MAINTAIN_HEADING
HEADING
REDUCE_MIN_APPROACH_SPEED
STOP_ALTITUDE
SPEED_OWN
NO_CONCEPT
SQUAWK
REPORT_ESTABLISHED

Blue, < 10 training samples

Contents

- T1: workshop to evaluate the ABSR against the requirements.
- O2: controller uses ABSR system for radar label maintenance
- O1: Baseline and trained system:
  Which one is better?
- T2: Simulate the effect of monthly adding more and more training data

Proof-of-Concept Trials, January 2018 (3)

- O1 involves controllers who concentrate only on the different outputs of a baseline ABSR system and on an ABSR system trained with all the available MALORCA training data.

Baseline ABSR

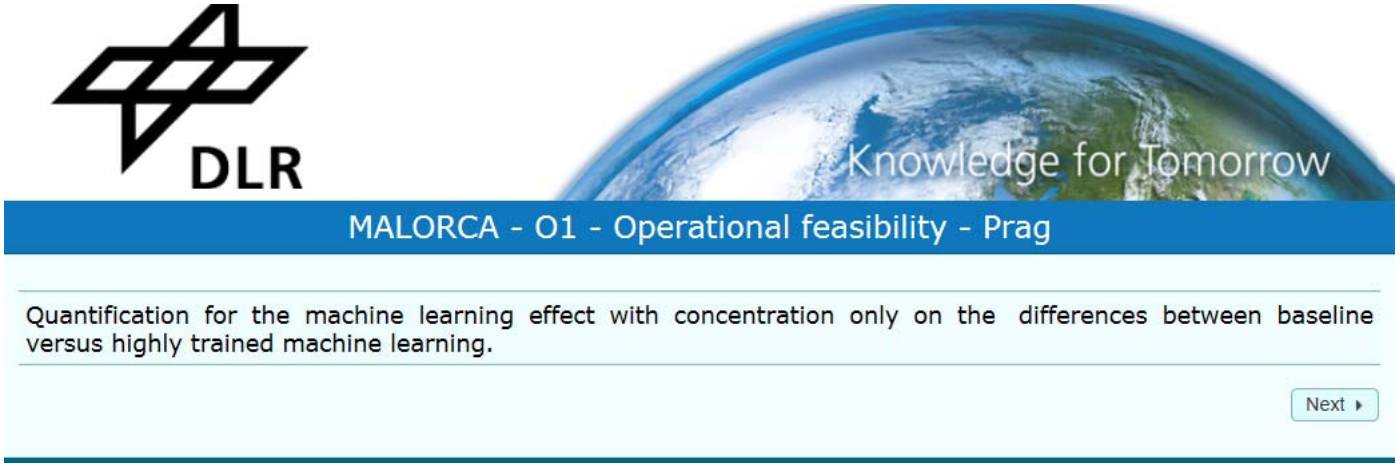
Trained ABSR

- Which one is better?
- Side by Side Experiment (SxS)

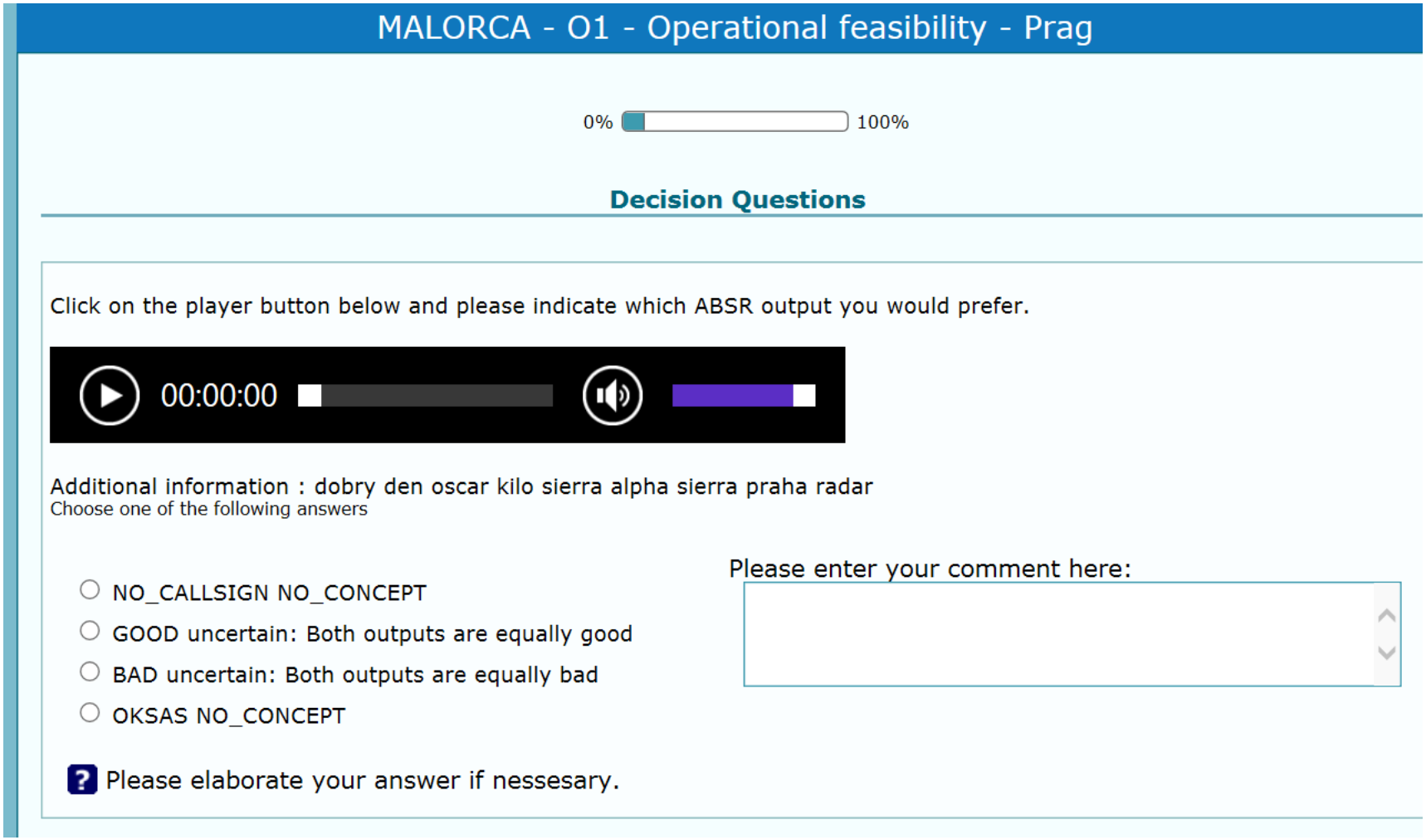
Side by Side(O2)

- 17 controllers from Vienna with transcription
- → 35 examples with differences
- 12 controllers from Prague with transcription
- → 36 examples with differences

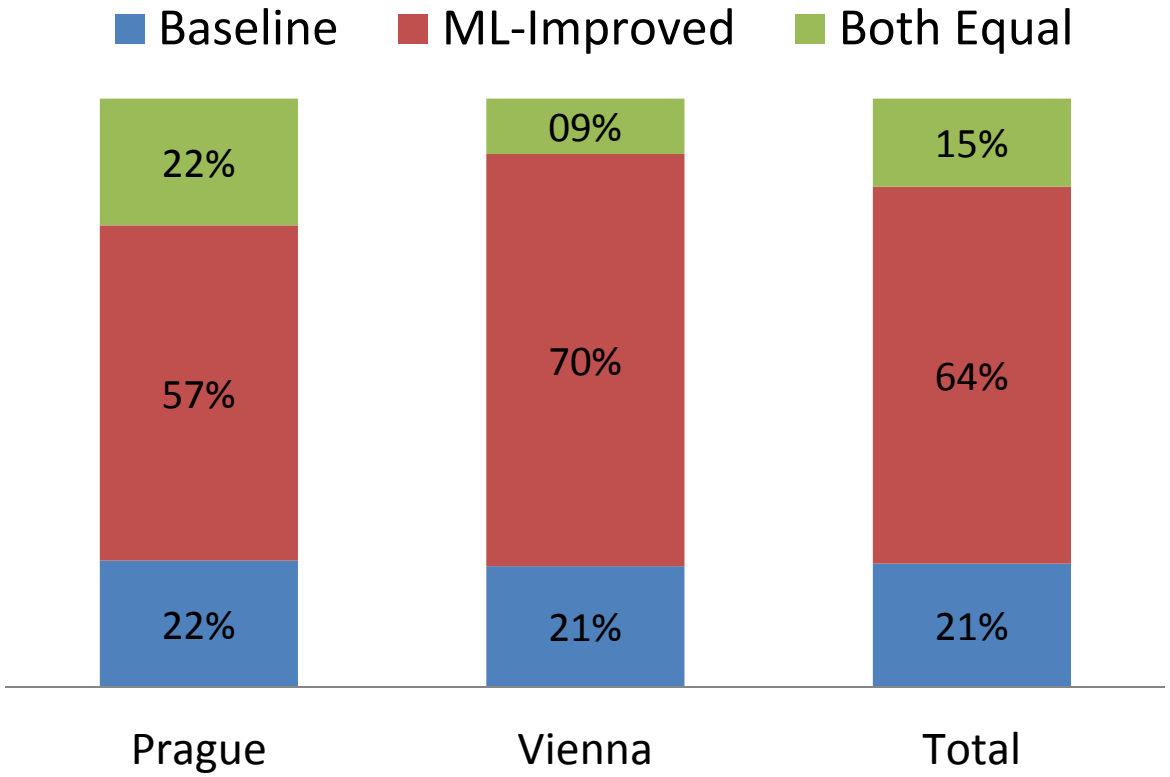
Prague utterances:
http://survey.fl.dlr.de/index.php/646844/lang-en
Vienna utterances:
http://survey.fl.dlr.de/index.php/224155/lang-en



Side by Side


MALORCA - O1 - Operational feasibility - Prag
0%
100%
Decision Questions
Click on the player button below and please indicate which ABSR output you would prefer.
00:00:00
Additional information : dobry den oscar kilo sierra alpha sierra praha radar
Choose one of the following answers
NO_CALLSIGN NO_CONCEPT
GOOD uncertain: Both outputs are equally good
BAD uncertain: Both outputs are equally bad
OKSAS NO_CONCEPT
Please enter your comment here:
Please elaborate your answer if nessesary.

Results of Side by Side Experiment



Dec. 2017 version Prague
Recognition Rate:
Baseline: 77.7%
Improved: 87.4%

Dec. 2017 version Vienna
Recognition Rate:
Baseline: 64.9%
Improved: 72.1%

Trained ABSR-System outperforms Basic Recognizer in 2/3 of the cases.

Results of Side by Side Experiment (4)

Example:

Recognizer 1:

- NO_CALLSIGN TURN_RIGHT_HEADING 310
- NO_CALLSIGN CLEARED_ILS 34
- NO_CALLSIGN REPORT_ESTABLISHED

Recognizer 2:

- NO_CALLSIGN NO_CONCEPT

Prague utterances:
http://survey.fl.dlr.de/index.php/646844/lang-en
Vienna utterances:
http://survey.fl.dlr.de/index.php/224155/lang-en

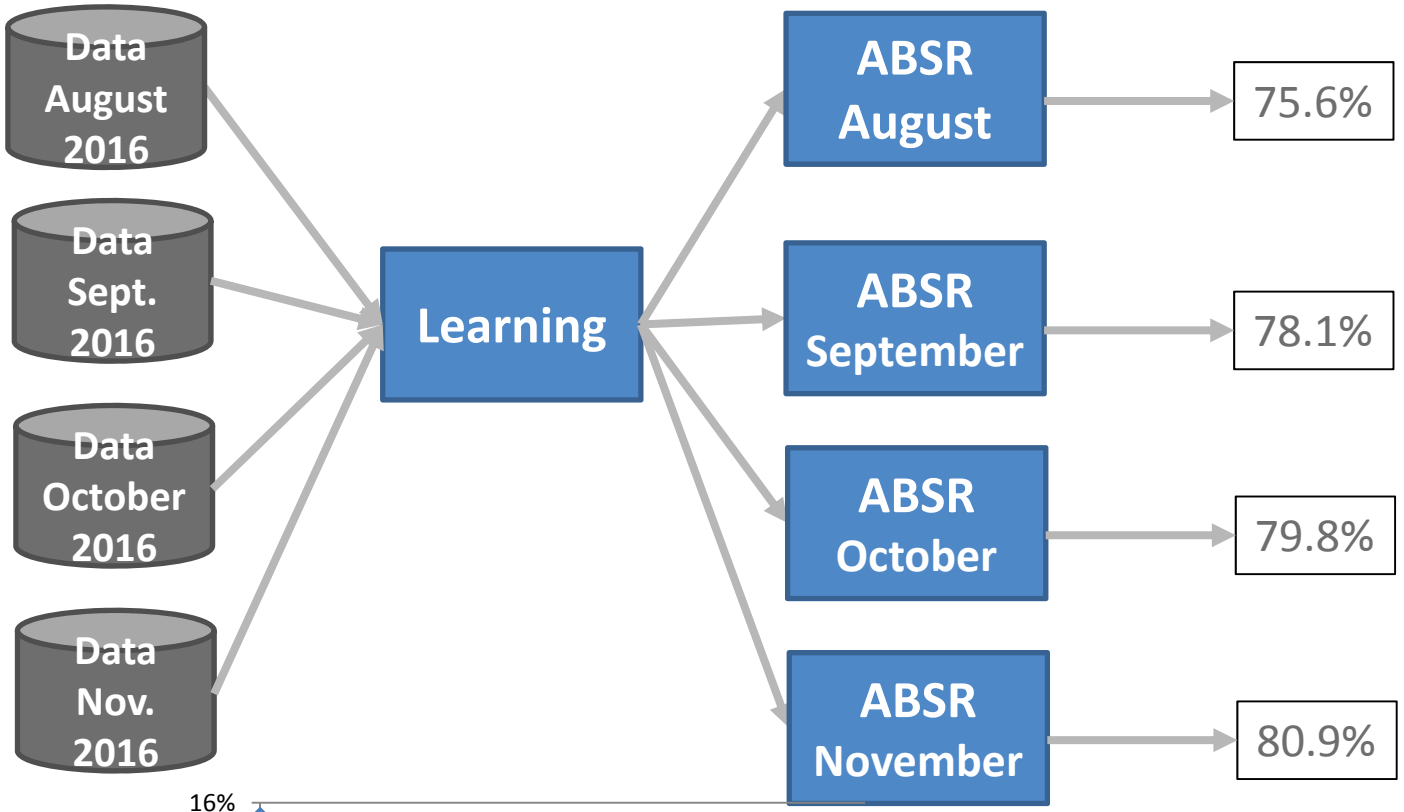
Main question:
What do you prefer?
Complete rejection or partly correct output?

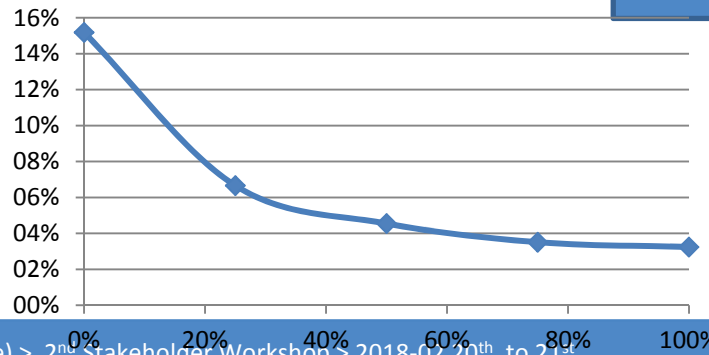
90%/2% versus **95%/4%** recognizer?

Contents

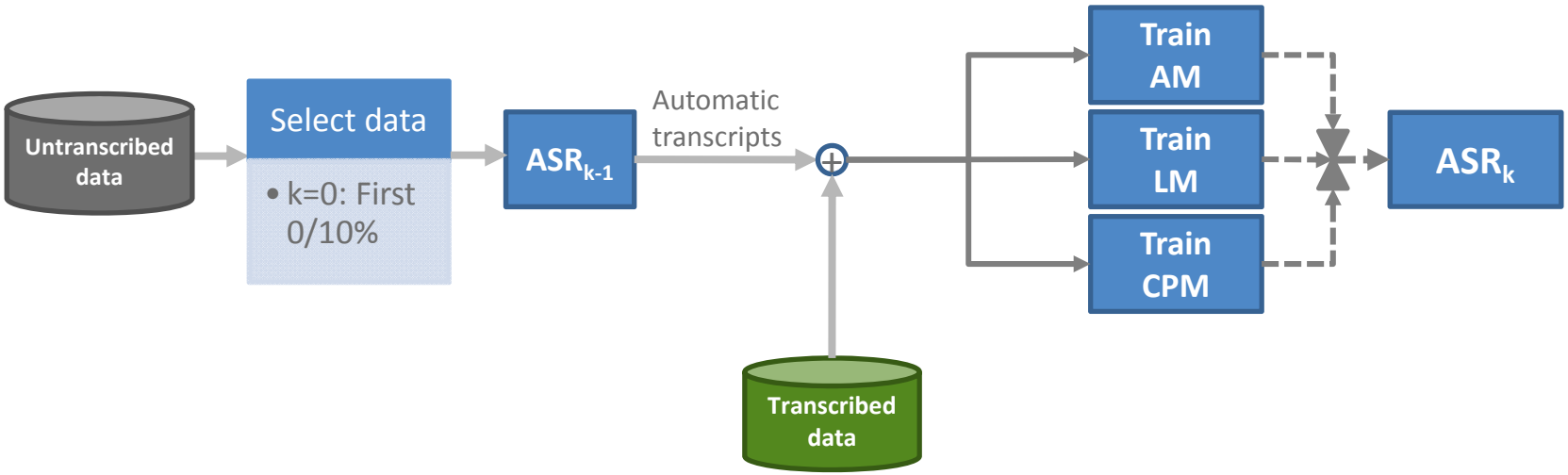
- T1: workshop to evaluate the ABSR against the requirements.
- O2: controller uses ABSR system for radar label maintenance
- O1: Baseline and trained system:
  36 different situations with always 2 different outputs
  Which one is better?
- T2: Simulate the effect of monthly adding more and more training data

T2: Proof-of-Concept for Continuous Learning



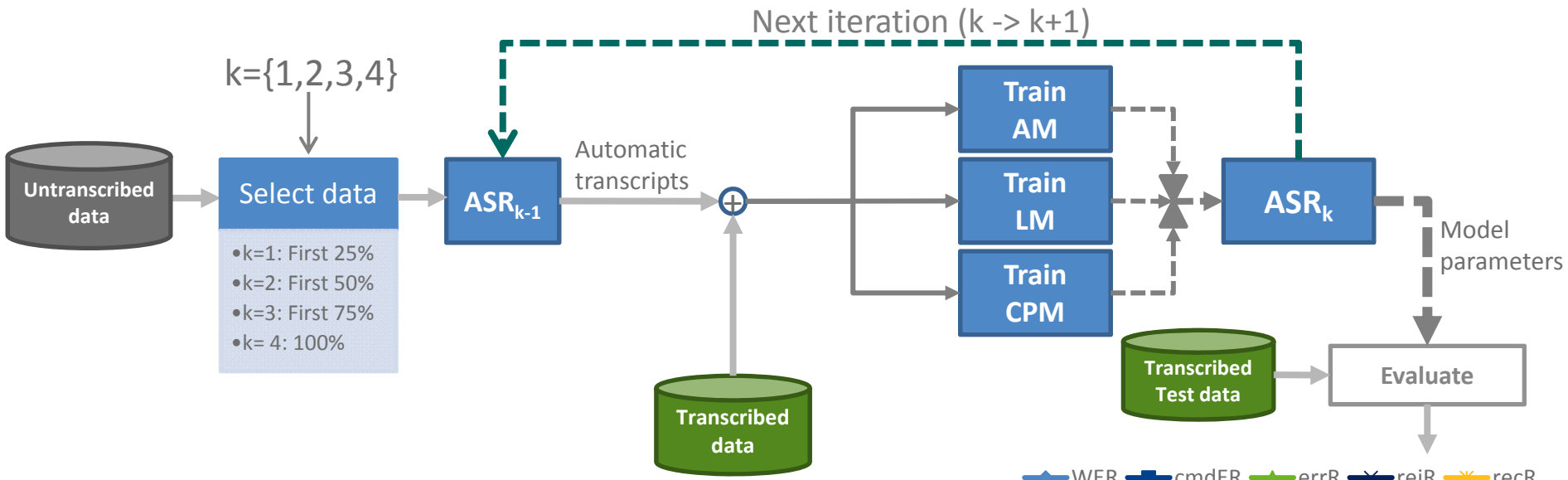


Explanations



- CPM-10 trained from 10% of untranscribed controller commands is used to predict commands for 25% of untranscribed controller commands
- AM/LM are trained now on 70% of transcribed data and 25% of untranscribed controllers commands
- Training of AM/LM uses output of CPM-10 , i.e. recognitions, which are predicted are used as good examples resp. bad examples if not predicted (it is not known whether recognitions resp. predictions are correct or not

T2: Flow diagram

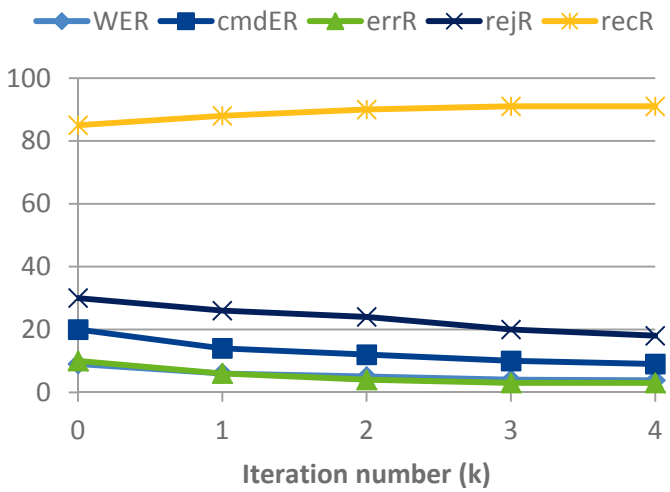


Data flow

Store across iterations

Model parameters flow

Base system, **ASR$_0$** trained with out-of-domain data and adapted with transcribed in-domain data



Computer Times for 100% of training data

Automatic transcript generation

- Between 8 and 40 hours on a single CPU (1 job)
- Utterance transcription from 0.6 to 5 times real time
  1s to 8s for 5s audio file (1-best, no context to 5-best, context

Speaker dependent DNN AM training

- 3 hours on single GPU + 8 hours on a CPU cluster (10 jobs)

Language model training

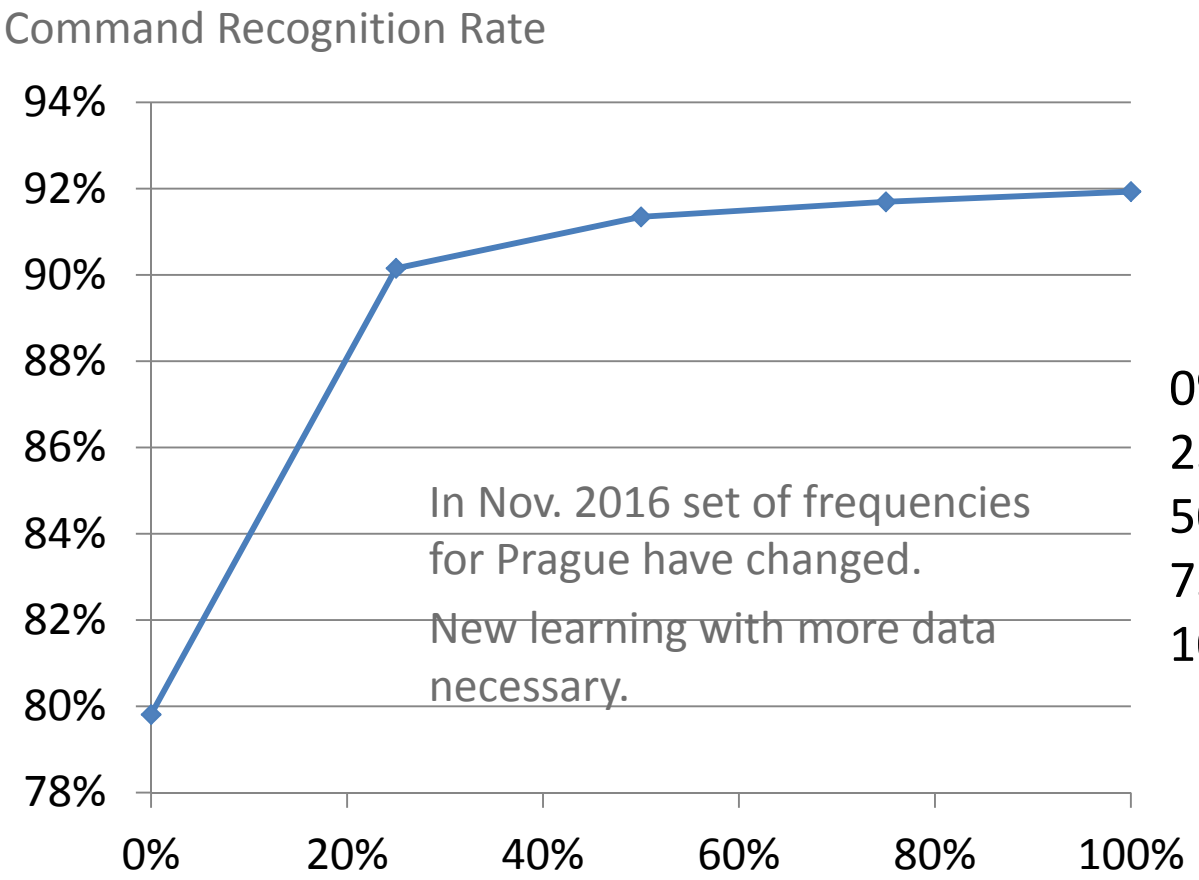
- 3 hours on a CPU cluster (10 jobs) for first iteration and
  0.5 hours for the others (1 job sufficient)

Command prediction model training

- 3 hours on a Windows Laptop (no parallel processing)

One iteration step lasts approx. 3 days, so 2 weeks for the iteration with 5 steps.

Learning Curve for Prague

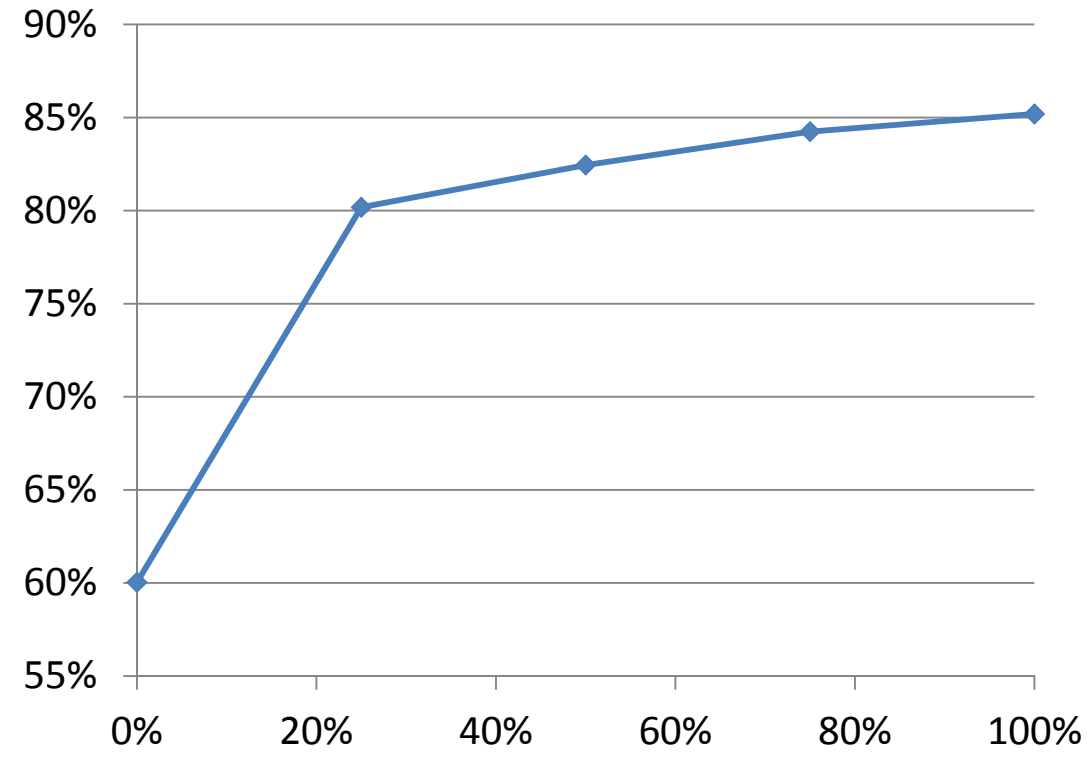


	Recognition Rate
0%/10%	79.8%
25%	90.2%
50%	91.3%
75%	91.7%
100%	91.9%

Doubling from 50% to 100% still results in an improvement of 0.6%

Perhaps 800% result in 92.6%.

Learning Curve for Vienna



	Recognition Rate
0%/10%	60.0%
25%	80.2%
50%	82.4%
75%	84.2%
100%	85.2%

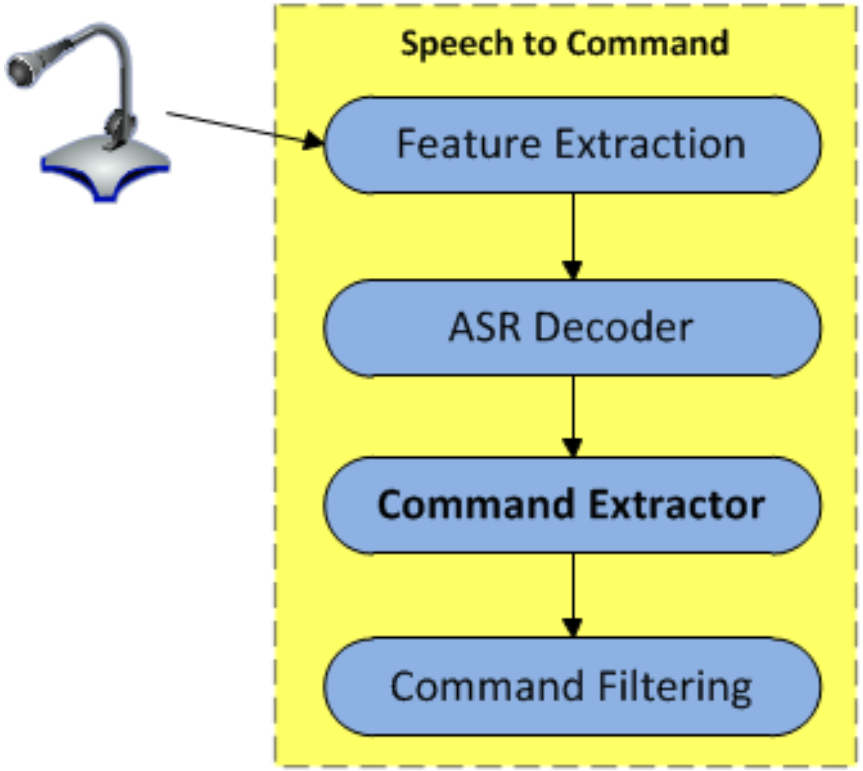
Doubling from 50% to 100% still results in an improvement of 2.8%

Perhaps 800% result in 90.2%.

Results of ”Normal“ ASR System

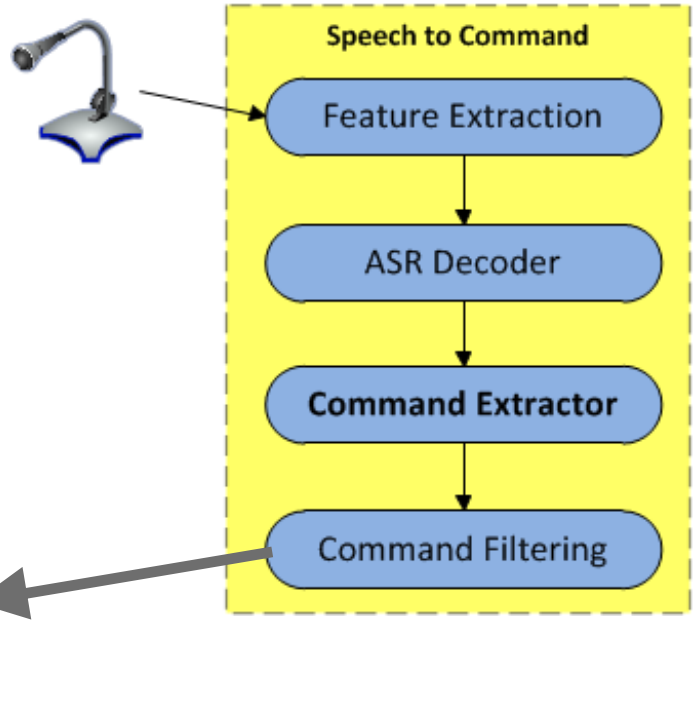
Recognition Rate: 87.5%
Error Rate: 6.7%

Rejection Rate > 6%



Results of ”Normal“ ASR System

Recognition Rate: 87.5%
Error Rate: 6.7%



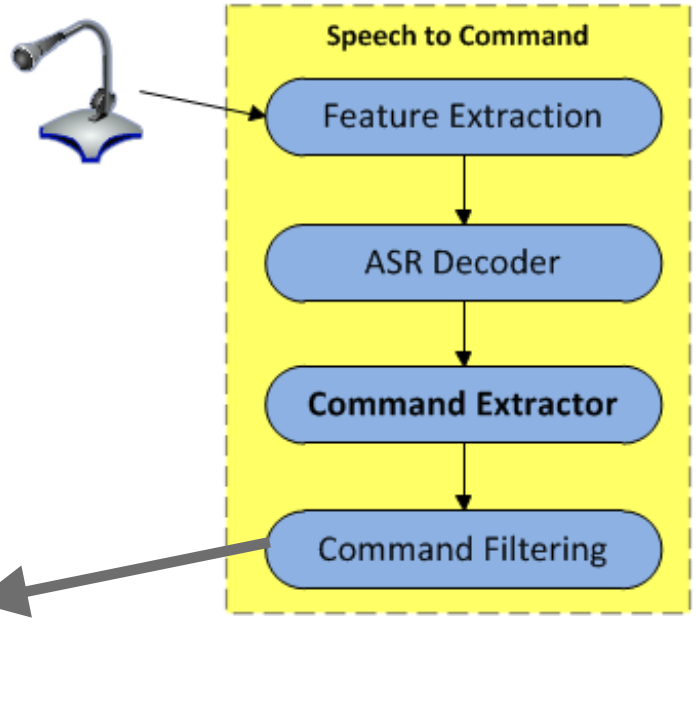
In the best case
- all 6.5% errors are filtered out
- None of 87.5% are filtered out

In reality.
- Some false rejections (we get below 87.5%)
- Some remaining errors (error rate > 0%)

Results of „Normal“ ASR System

Recognition Rate: 87.5%
Error Rate: 6.7%

Recognition Rate: 85.7%
Error Rate: 0.5%



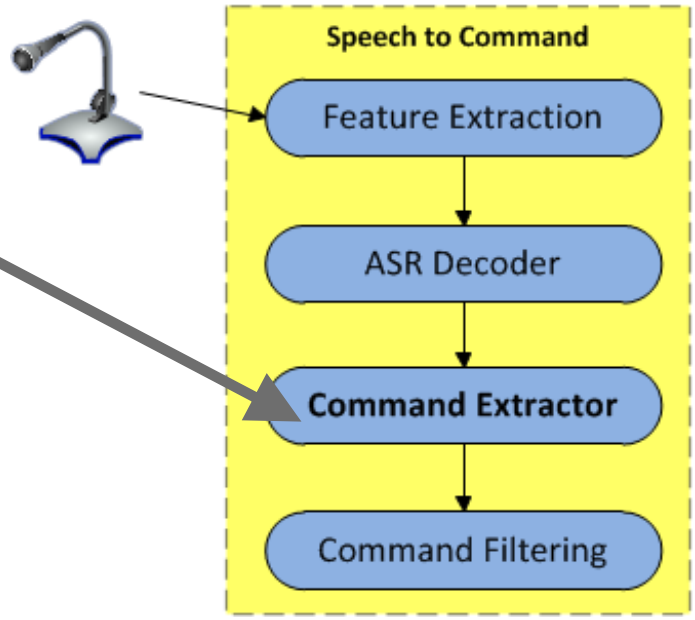
Filtering is good,
But we do “only” improve error rate and not recognition rate.

Results of ABSR System

Normal Recognition Rate: 87.5%
Normal Error Rate: 6.7%

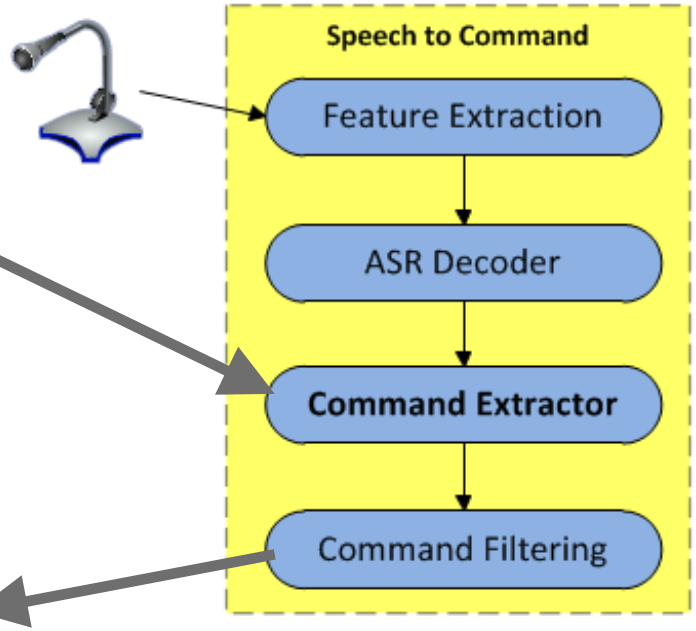
Hypotheses Generator

Recognition Rate: 93.8%
Error Rate: 2.0%



Results of ABSR System with checker for Prague

Normal Recognition Rate: 87.5%
Normal Error Rate: 6.7%



Hypotheses Generator

Recognition Rate: 93.8%
Error Rate: 2.0%

Plausibility Checker

Recognition Rate: 91.8%
Error Rate: 0.57%

Any value between 0.57% and 2.0% is possible

Trade-off between error rate
and recognition rate.

Results of ABSR System with Checker for Vienna

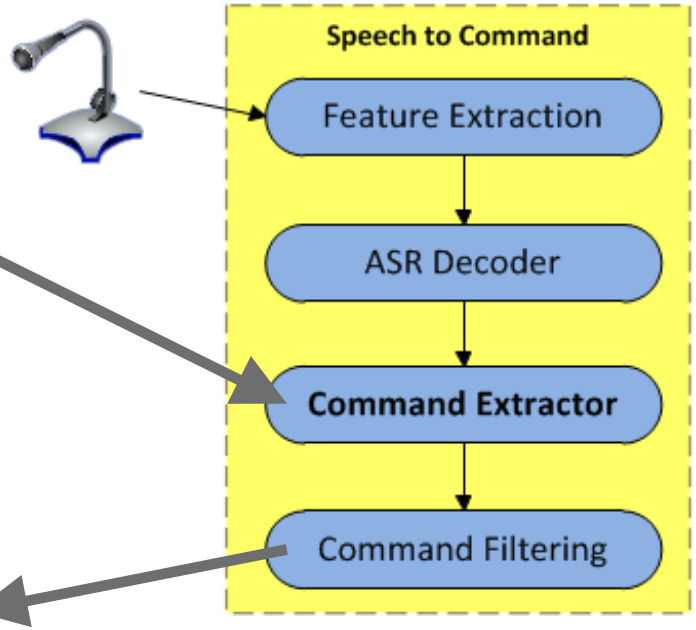
Normal Recognition Rate: 71.3% (87.5%)
Normal Error Rate: 15.7% (6.7%)

In brackets Prague

Hypotheses Generator

Recognition Rate: 84.4% (93.8%)
Error Rate: 6.8% (2.0%)

Plausibility Checker



Recognition Rate: 83.3% (91.8%)
Error Rate: 3.7% (0.57%)

Any value between 3.7% and 6.8% is possible

Trade-off between error rate
and recognition rate.

Different Noise Levels for Prague and Vienna

Prague

Vienna

Prague

Vienna

Conclusions

- 90%/2% or 95%/4% : Which performance is best?
- Automatic Transcription reduces manual costly transcription effort
- And increases command recognition rate from 80% to 92% (Prague) resp. 60% to 85% (Vienna)

- Context integration in recognition process already increases recognition rate from 87.5% to 93.8 (Prague) resp. 71.3% to 84.4% (Vienna)

- Using also Context in checker in Post-Recognition reduces command recognition error rate further from 2.0 to 0.57% (Prague) resp. 6.8% to 3.7% (Vienna)

MALORCA

Machine Learning of Speech Recognition Models for Controller Assistance

DLR

UNIVERSITÄT DES SAARLANDES

idiap RESEARCH INSTITUTE

austro CONTROL

Air Navigation Services of the Czech Republic

Covering the sky...

Thank you very much for your attention!

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 698824.

SESAR JOINT UNDERTAKING

Founding Members

EUROPEAN UNION EUROCONTROL