

Easy Adaptation of Speech Recognition to Different Air Traffic Control Environments using the DeepSpeech Engine

Matthias Kleinert, Narasimman Venkatarathinam, Hartmut Helmke, Oliver Ohneiser (DLR)
Maximilian Strake, Tim Fingscheidt (TU BS)

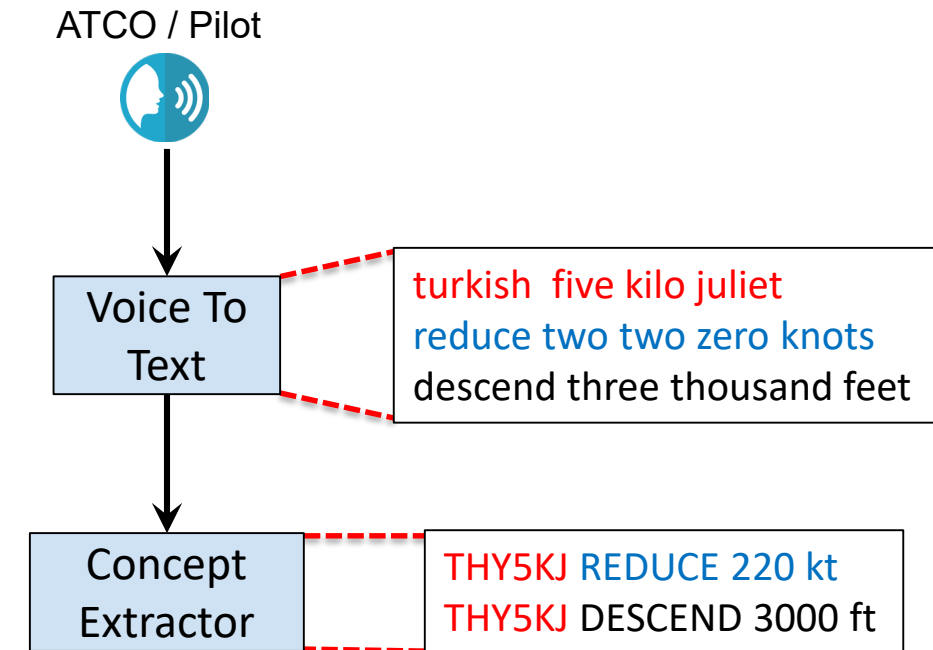
German Aerospace Center (DLR), Braunschweig, Germany
Technical University Braunschweig (TU BS), Braunschweig,
Germany

A large, curved image of the Earth from space occupies the right half of the slide. It shows a blue horizon, white clouds, and green landmasses, including parts of Europe and Africa.

Knowledge for Tomorrow

Speech in Air Traffic Control

- Speech is mankind's most important means of communication
- Speech also plays a major role in ATC
- Communication between ATCO and pilot primarily consist of:
 - Instructions
 - Reports
 - Readbacks
- However → Majority of ATC communication is not recognized and understood automatically as input for other tools/systems



What is the Problem?

- **Why not use an existing speech recognition system?**
 - Systems from Amazon, Google etc. not really suitable: Created for a different purpose and proprietary
 - PJ.16-04 CWP HMI : COTS-System with adapted ATC grammar achieved 31% to 82% recognition rate
 - **Custom Research solutions:**
 - AcListant® and MALORCA: Introduced / Improved ABSR with recognition rates between 83% and 95%
 - Speech technologies are not easily transferred to ATC-industry
 - **Is an easier solution possible?**
 - In recent years, end-to-end ASR models have received interest which simplify the training ASR systems
- So does a implementation exist which is mature enough and also easy to use?

ABSR = Assistant Based Speech Recognition
ASR = Automatic Speech Recognition
COTS = Commercial-of-the-Shelf

Content

1. DeepSpeech
2. Evaluation of recognizing speech
3. Is understanding more than just words?
4. Evaluation of understanding speech
5. Conclusion



DeepSpeech

- Open source, speech-to-text engine from Mozilla Corporation
- Inspired by research paper published from Baidu
- Input → Features extracted from the audio signal
- Neural Network with overall six layers
- Output → Characters over time – Results in a sequence words
- Optionally a language model can be added on top for improvement
- Provides descriptions and tools for training, recognition and deployment

→ So lets use it to recognize ATC speech

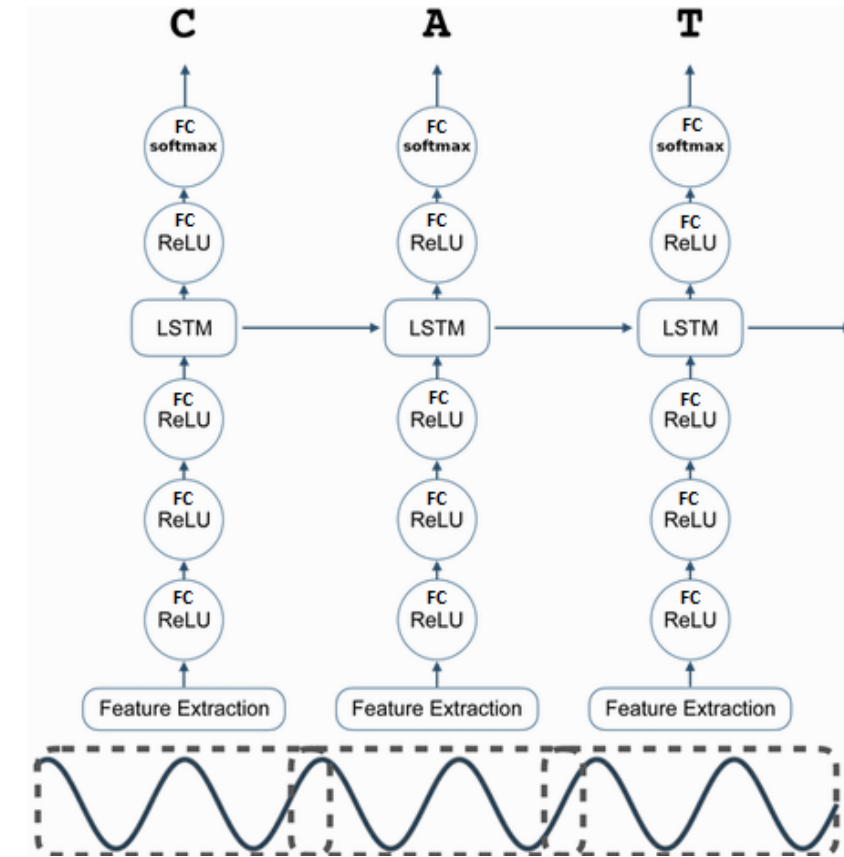


Fig. 2: DeepSpeech RNN model
<https://deepspeech.readthedocs.io/en/v0.9.3/DeepSpeech.html>

DeepSpeech – How to get the right model?

- Pre-trained model for English language:
 - 6500h of data – Fisher, Librispeech, Common Voice etc.
 - Model has never seen any ATC data before
- Train a model from scratch
 - Ignore existing model and train a completely new model
 - Only use ATC related data for model training
- Fine tuning:
 - Keep existing pre-trained model
 - Resume training with ATC related data

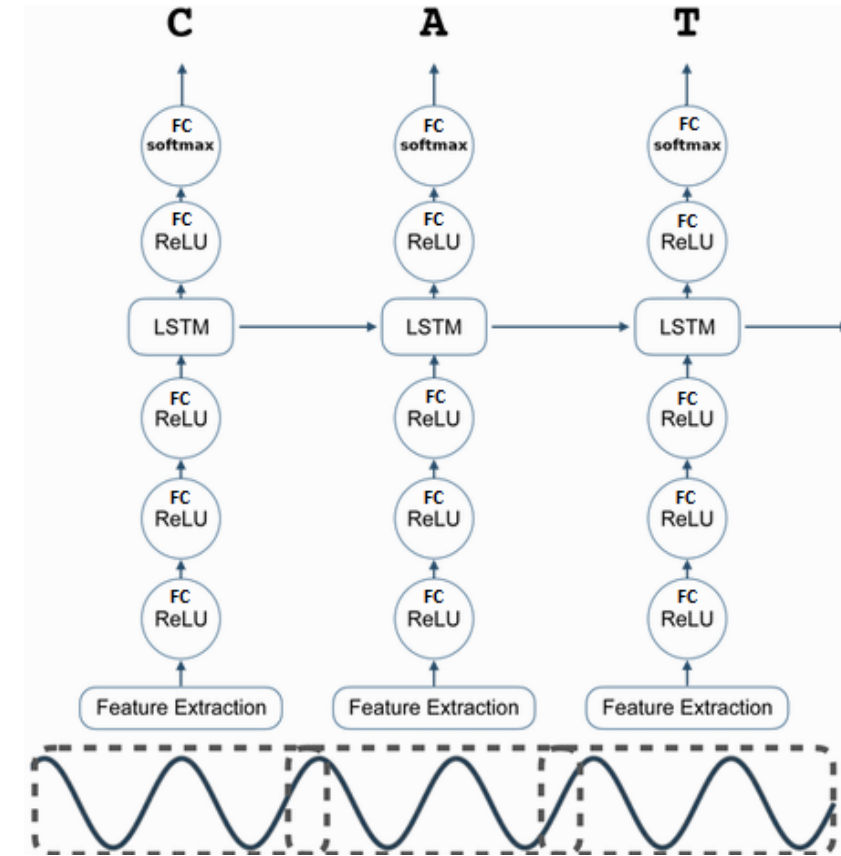


Fig. 2: DeepSpeech RNN model
<https://deepspeech.readthedocs.io/en/v0.9.3/DeepSpeech.html>

DeepSpeech – How to get the right model?

- Transfer Learning
 - Suggested to adapt pre-trained models for a different character set
 - Remove the weights of certain layers of the pre-trained model, reinitialize them and re-start the training process
- Layer Freezing
 - Approach for adapting an existing model to new languages [1]
 - Freeze weights of certain layers of a pre-trained model throughout the training process

→ Training of language models in addition to the DeepSpeech models

[1] Eberhard & Zesch, „Effects of Layer Freezing when Transferring DeepSpeech to New Languages“, 2021

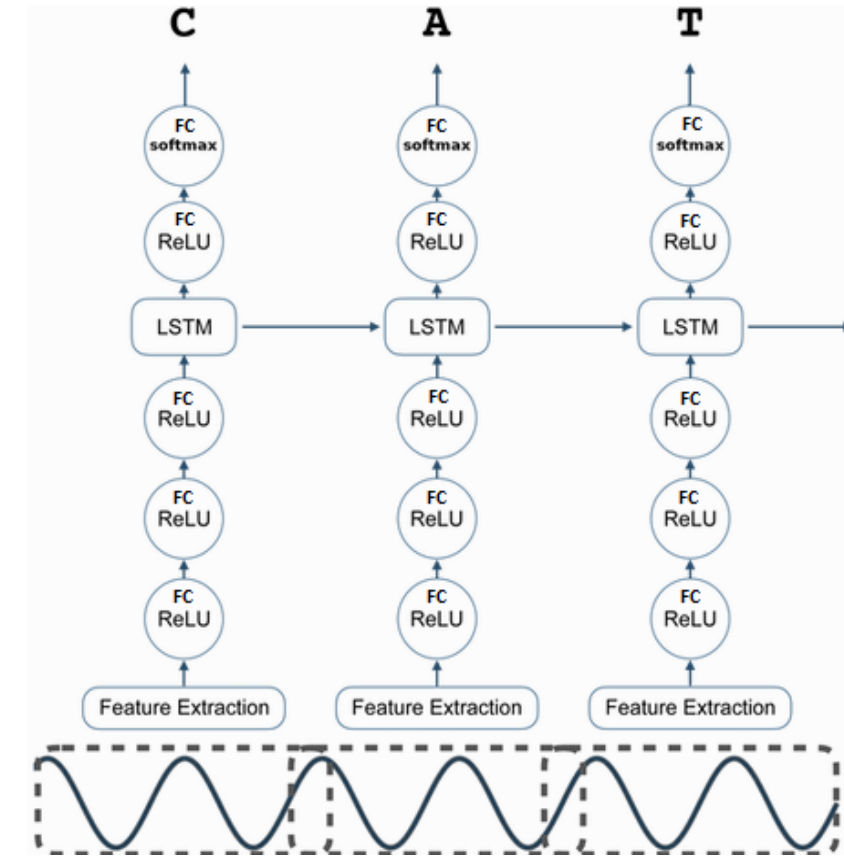


Fig. 2: DeepSpeech RNN model
<https://deepspeech.readthedocs.io/en/v0.9.3/DeepSpeech.html>

Content

1. DeepSpeech
- 2. Evaluation of recognizing speech**
3. Is understanding more than just words?
4. Evaluation of understanding speech
5. Conclusion



Evaluation of recognizing speech: Data

- Data description:

Data Set	ATCOSIM	DLR
Airspace	Söllingen, Zürich and Geneva	Vienna and Prague
Controller Position	En-Route	Approach
Recording Situation	100% simulation	62% operational, 38% simulation
Speaker Profile	60% male, 40% female, Germany and Switzerland	86% male, 14% female, Austria and Czech
Duration	10.7 h	22.9 h

- Data split:

Data Set	# Utterances	Duration
Training	17,877	21.0 h
Validation	7,662	9.0 h
Test	2,913	3.6 h

Evaluation of recognizing speech: Word Error Rate

- Evaluated 12 different models combined with different Language Models (LM)
- Word Error Rate (WER) based on Substitutions (S), Insertions (I), Deletions (D) and Number of Words in ground truth (N) → How accurately does an ASR recognize speech

$$WER = \frac{S + I + D}{N}$$

Model	No LM WER %	3-gram LM WER %	4-gram LM WER %
Pre-train	100.0	85.2	86.6
Scratch	20.0	8.6	8.2
Fine-Tune	15.5	6.3	6.0
Trans-Learn-1	22.6	7.7	7.4
Trans-Learn-2	24.3	8.3	7.8
Trans-Learn-3	31.4	9.1	8.6
Trans-Learn-4	32.9	10.0	9.4
Trans-Learn-5	33.1	10.4	9.8
Layer-Freeze-1	15.8	6.3	6.0
Layer-Freeze-2	16.7	6.8	6.5
Layer-Freeze-3	17.1	7.7	7.4
Layer-Freeze-4	18.5	7.9	7.6

11th SESAR Innovation Days

Content

1. DeepSpeech
2. Evaluation of recognizing speech
3. **Is understanding more than just words?**
4. Evaluation of understanding speech
5. Conclusion



Is understanding more than just words?

- Natural Language offers a lot of different possibilities to express a certain intention
 - Also ATCOs are no exception, even though ATC phraseology is standardized
 - Example:
 - lufthansa three echo romeo make it heading two two zero degrees left
 - echo romeo turn left two two zero degrees
 - lufthansa three echo romeo turn left now onto a heading of two two zero
 - three echo romeo turn left now heading two two zero degrees
 - lufthansa three echo romeo heading left of two twenty
 - lufthansa three echo romeo turn further left heading two two zero degrees
- All meaning DLH3ER should turn LEFT onto a HEADING of 220 degrees

PJ.16-04 Ontology

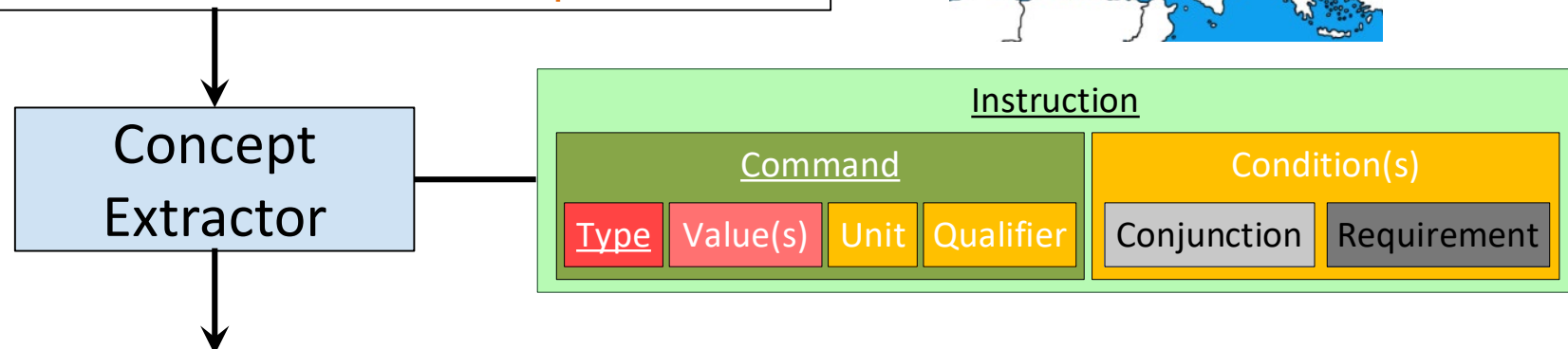


Input:
Transcription

lufthansa two alfa four
turn left heading three two zero
descend flight level one two zero when passing gunpa
be careful airbus three two zero at two o'clock position

Concept Extractor

Implementation to automatically apply Ontology
on a given word sequence



Expected Output:
Annotation

DLH2A4 HEADING 320 LEFT
DLH2A4 DESCEND 120 FL WHEN PASSING GUNPA
DLH2A4 INFORMATION TRAFFIC

Content

1. DeepSpeech
2. Evaluation of recognizing speech
3. Is understanding more than just words?
- 4. Evaluation of understanding speech**
5. Conclusion



Evaluation of understanding speech

- All experiments based on best word recognition model with WER of 6%
- Evaluation only possible for data set with Prague and Vienna – for others no ground truth available
- Data:

# Commands	All	Training + Validation	Test	Cmds / Utt (Average)
Ops Prague	6,120	4,765	1,355	2.0
Lab Prague	6,904	6,313	591	1.6
Ops Vienna	4,713	4,419	294	1.6
Lab Vienna	6,014	5,384	630	1.7

- Used metrics:

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

Evaluation of understanding speech: Results

- Gold = Based on manual created or verified word recognitions (WER almost 0%)
- Auto = Based on word recognitions with DeepSpeech fine tuned model (WER see table)

	CaE Gold	CaE Auto	WER
Ops Prague	0.6%	5.7%	9.1%
Lab Prague	0.0%	3.9%	1.1%
Ops Vienna	0.6%	6.4%	9.0%
Lab Vienna	1.6%	3.0%	2.6%

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

	RcR Gold	RcR Auto	WER
Ops Prague	98.2%	85.0%	9.1%
Lab Prague	99.7%	97.0%	1.1%
Ops Vienna	96.3%	87.4%	9.0%
Lab Vienna	97.5%	95.6%	2.6%

Conclusion

- Various models were trained with DeepSpeech providing very good results for ATC speech (WER 6%)
- The tools and scripts of DeepSpeech are mature and easy to use
- Although a language model is not necessary by definition, the results show that it can significantly improve the recognition (improvement by a factor of 2.5 for the best model with a 4-gram LM)
- Results of the semantic interpretation showed that the WER varies for different data sets between 1.1 and 9.1%, but improvement is expected with
 - More data
 - Specific model for the controlled area
- Future Work
 - Train a model with more data to get a general ATC model which can be tuned for different environments
 - Look for ways to incorporate context knowledge into recognition process (e.g. Surveillance data)



11th SESAR Innovation Days

