Automatic Callsign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications

Juan Zuluaga-Gomez, **Karel Veselý**, **Alexander Blatt**, **Igor Szöke**, Petr Motlíček, Dietrich Klakow, Allan Tart, Amrutha Prasad, Saeed Sarfjoo, Pavel Kolčárek, Martin Kocour, Jan "Honza" Černocký, Claudia Cevenini, Khalid Choukri, Mickael Rigault, Fabian Landis

Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne, Brno University of Technology, Saarland University, OpenSky Network, ReplayWell, Honeywell, Romagna Tech, Evaluations and Language Resources Distribution Agency (ELDA)

The 8th OpenSky Symposium

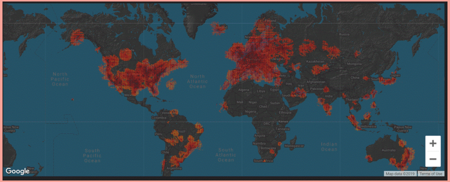
Highlights

- ATCO^2 EU project (Clean Sky 2 Joint Undertaking & European Union)
- Main challenges in speech recognition and callsign detection for ATC communications
- Standardization and data collection of ATC speech + radar data
- Deployment of first VHF receivers in different airports/countries
- ATCO^2 approach to use context-information such as radar (from OSN servers)
- Introduced state-of-the-art speech-to-text technologies for ATC communications
- First version of ATCO^2 text-to-callsign system

Outline

What is ATCO2[1]?

- EU project, with goal to collect, organize and pre-process air-traffic control audio data
- We collect voice communications between pilots and Air Traffic Controllers (ATCos)
- And generate meta-data useful for further processing



Funding Parties:

[1] https://www.atco2.org/.

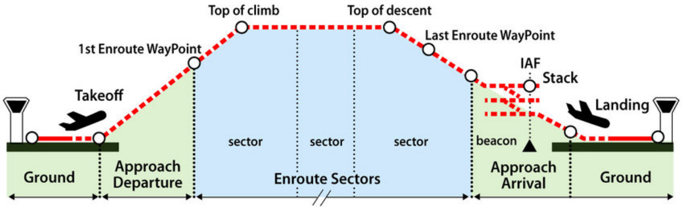
ATC speech is not an easy task!

Speech

- Speech variations (stress, fatigue)
- Inter-speaker variations
- Different accents/dialects
- Neither spontaneous, read nor pure command speech

ATC scenarios

- Clearances, arrivals, takeoffs
- Climb, cruise, descent, ground taxiing



Automatic Speech Recognition for ATC in previous projects

- Previous projects MALORCA[2] and AcListant[3], audio data limitations:
  - Focused on ATCo speech, not on pilot's speech
  - Limited amount of training data <100 hr
  - Data from few airports (Prague, Vienna), few speakers and English accents
  - Only clean data without noise
- ATC communications follow ICAO[4] regulations (constrained vocabulary/lexicon and grammar)
  - 1 message = one callsign (airplane name) and one command (in normal conditions)

[2] MAchine Learning Of speech Recognition models for Controller Assistance, http://www.malorca-project.de/wp

[3] Active Listening Assistant, www.AcListant.de

[4] International Civil Aviation Organization

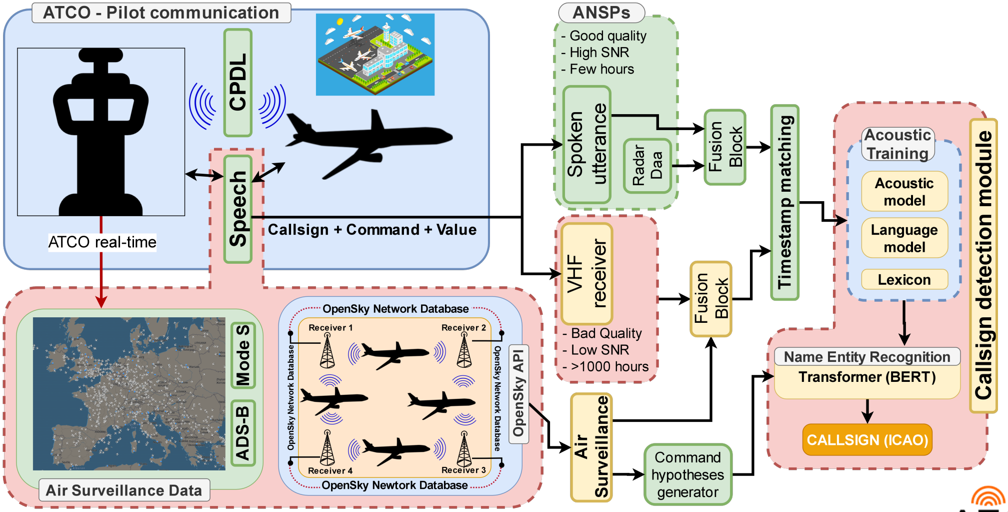
What we want to solve?

- Manually transcribed ATC data **EXPENSIVE** to produce (one man-week work = 1 hour speech[5,6])
- First phase: Collect public ATC data (LiveATC, some are quite noisy)
- Second phase: Design, build and deploy a system based on VHF receivers to record our own data (complying with legal regulations)
  - Data with different accents, protocols, speakers, quality. Also include pilot speech.
- Generate automatic transcripts by Automatic Speech Recognition (ASR)
  - Correct transcripts manually on a subset of data
  - Use surveillance data to assist the recognition
- Build ASR and **Callsign detection** system for ATC communications[7]
  - This non-trivial task is the current research goal of **ATCO2** project

[5] Ferreiros et al., "A speech interface for air traffic control terminals".

[6] Cordero, Dorado, and Pablo, "Automated speech recognition in ATC environment".

[7] Kleinert et al., "Semi-supervised adaptation of assistant based speech recognition models for different approach areas".



The 8th OpenSky Symposium

Done so far:

1. Gathered around 195 hours of pre-existing annotated ATCo speech (different speakers, accents and quality)
2. We are recording data with VHF receivers in several pilot locations
3. We have a Speech-to-text recognition system, the sub-tasks were:
   - Data pre-processing, unification of transcripts across ATC databases
   - Voice Activity Detection and Diarization
   - Building the lexicon and language model (LM)
   - Training the acoustic model (AM) with two state-of-the-art approaches
   - Extraction of lists of call-signs from surveillance data, for given location and timestamp
4. We designed a test set to evaluate our **speech-to-text** and **text-to-callsign** systems

What we can do with surveillance data?

- Main target: extract the callsign from a given utterance
- Initial solution: use speech-to-text and text-to-callsign system
- Previous results: good performance BUT only with **good quality data**
- New target: improve the performance on good/**BAD quality data**
- New solution: use surveillance data from **OSN servers** either in:
  - The output of the speech-to-text system
  - The input and/or output of the text-to-callsign system, or,
  - Simultaneously in both systems, a hybrid approach
- Expected results: much better performance in data

The 8th OpenSky Symposium

Air surveillance data and timestamp with spoken communications

1. Gather an ATC segment with a VHF receiver
2. Record the timestamp, add location (airport/receiver)
3. Send a query to OSN servers composed of:
   - Time range: input from timestamp
   - Location: area to search centered around the receiver/airport
4. OSN sends back the matching callsigns in ICAO format

Such as: LUF189AF (ICAO format)
LUFTHANSA ONE EIGHT NINE ALFA FOXTROT (verbalized format)

Nevertheles...

Non-standard abbreviations

- LUFTHANSA → HANSA
- SCANDINAVIAN → SCAN
- SCANWING → SCAN
- TRANSAVIA → AVIA
- RYANAIR → RYAN
- SPEEDBIRD → BIRD

Then what?

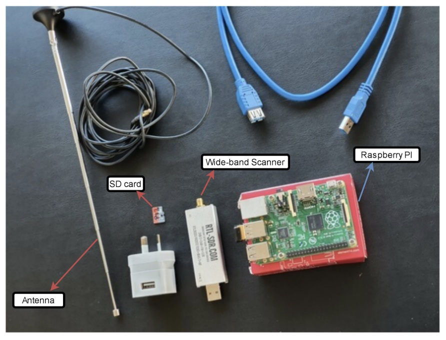
- Several "deviations" per callsign
- Initial contact: full callsign
- After: abbreviations are allowed

We work with many possible verbalizations of the ICAO callsign

How do we record our own data?

Very High Frequency Receivers

- Capture raw audio data from different airports/countries
- Recording software: RTL-SDR-Airband
- Output: complex I/Q format, converted by csdr to flac
- We evaluated SNR of 2 different hardware setups
  - RTL-SDR receiver, dipol antenna → RSP1A receiver, Watson WBA-20 antenna
  - More expensive setup, better SNR (less noise)
  - The WER was 8.3% smaller for more expensive setup (33.0% → 24.7%)



The 8th OpenSky Symposium

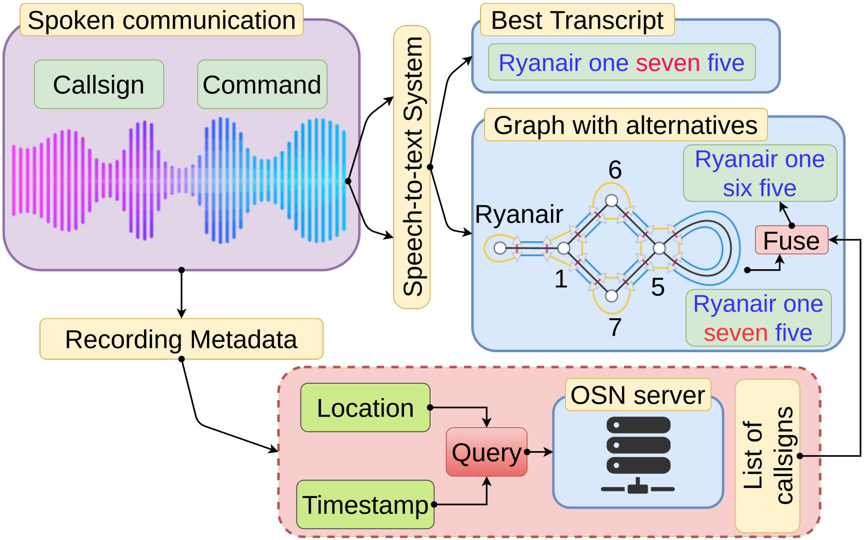
Speech-to-text system

(= ASR)

How we produce transcripts from speech?

Speech-to-text system

- One of the main components of the overall pipeline
- Speech-to-text system has as:
  - Input: audio signal with speech of ATCOs and pilots
  - Output: hypothesised transcript
- The output could be "best transcript" or a lattice (weighted graph of alternatives)
- Challenge: use context information (radar) to increase the performance


Spoken communication
Callsign
Command
Speech-to-text System
Best Transcript
Ryanair one seven five
Graph with alternatives
Ryanair
6
1
5
7
Ryanair one six five
Fuse
Ryanair one seven five
Recording Metadata
Location
Query
Timestamp
OSN server
List of callsigns

Proposed speech-to-text systems

Hybrid approach

- Pronunciation lexicon, acoustic and language model trained separately
- Based on HMM systems and Deep Neural Networks (DNN)
- More complex, but more freedom during training
- Requires less data to generalize and it is faster to train

End-to-End approach

- Pronunciation lexicon, acoustic model and implicit language model are trained jointly as one model
- Less complex, but less freedom
- Requires more data and more training time to generalize
- Word-piece output-symbols, difficult to generate lattices

Data, data and more data

Audio related

- Training on 7 databases, 195 hours (Table 1, manuscript)
- Doubled the data by adding noise corresponding to LiveATC audio channels (helps a lot!)
- The test sets:
  - Test set 1 Airbus database (see Airbus challenge[8])
  - Test set 2 from MALORCA (previous EU project)
  - Two test sets from LiveATC; low quality data → more challenging
  - Test set gathered from LKTB airport (Brno, Czechia) with better equipment

Delpech et al., "A Real-life, French-accented Corpus of Air Traffic Control Communications".

Text related pre-processing works

- Unify the transcripts from all databases
  - Use same ICAO alphabet and "number words"
  - Standardize the word-splitting such as: "take off" "take-off" or "takeoff" to single type
  - Ligature the multi-word airline designators: "air berlin" → "air_berlin"
- Create airline designator table for callsign verbalization: LUF → LUFTHANSA
- Preparing external data for language model training:
  - Verbalized callsigns from OSN flight-lists (2019/2020)
  - Adding all possible runaways numbers
  - Adding waypoints from Europe

Performance of current speech-to-text system

Table: Performance measured in Word Error Rates (WER).

Test set	WER%	
	Hybrid	End-to-End
AIRBUS	8.1	10.2
MALORCA	5.0	7.2
LiveATC set1	34.5	44.8
LiveATC set2	33.0	40.4
LKTB	24.7	32.6

Quick takeaways

Performance of speech-to-text system

- Test sets with better audio quality (AIRBUS, MALORCA & LKTB) → better performance
- Drop of performance for LiveATC test sets → due to worse audio quality
- Performance difference LKTB vs MALORCA (24.7 vs 5.0):
  - Lexical difference: callsigns, runways, local names. And accent.
- Solution: "boost" the system with "local words" (names and also verbalised callsigns from OSN)

What is boosting? And, how this can help?

A-priori boosting

1. Applied to "recognition grammar" prior to the time-taking recognition search
2. Give score-discounts to certain words/phrases (callsign)
3. Output: Increase chance that the speech-to-text system outputs the correct "phrase"

Ex-post boosting

1. Applied to "lattices" with alternative outputs generated by time-taking recognition search
2. Give score-discounts to certain words/phrases (callsign)
3. Output: Increase chance to get correct "phrase" as best hypothesis (faster, but less effective than "a-priori" boosting)

Boosting the speech-to-text system with retrieved callsigns from OSN

Table: A-priori boosting with list of callsigns from surveillance data retrieved from OSN (query: location/radius and timestamp), boosting done once per test-set, hybrid recognizer

Test set	WER%	
	non-boosted	boosted
LiveATC set1	34.5	33.6
LiveATC set2	33.0	30.8

A-priori boosting with phrases is slow (about 5min per run).
→ It cannot be used per-utterance.
→ **We are on a good way. More experiments as future work!**

Callsign identification system

Callsign identification system

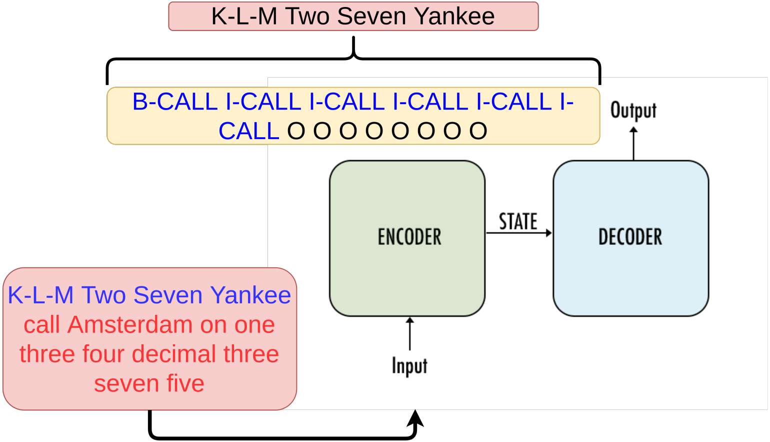
Detection Module - Transformer (BERT)

Sequence labeling task → Input seq. (transcripts) → Output seq. CALLSIGN
IOB format used; i) **I**: Inside; ii) **O**: Outside; iii) **B**: Beginning

Example

Input:
KLM Two Seven Yankee call Amsterdam on one three four decimal three seven five
Output:
B-CALL I-CALL I-CALL I-CALL I-CALL I-CALL O O O O O O O O

If you are wondering how the Callsign Detection Module works...



The 8th OpenSky Symposium

Callsign Detection Module - Generalities

1. Named Entity Recogition system, fine-tuned BERT model (Transformer)
2. Only, AIRBUS data used (only training set with callsign per each utterance)
3. System tested on AIRBUS (dev set) and LiveATC set 1
4. Currently testing on the "ground truth transcripts" (not the output of the speech-to-text system)

Table: Current performance of the NER module (with ground truth labels). Note: F1 roughly corresponds to "accuracy", the value 1.0 would be a perfect detection.

Test set	**F1**	**Precision**	**Recall**
AIRBUS	0.953	0.978	0.934
LiveATC set1	0.738	0.897	0.638

The 8th OpenSky Symposium

Recap

- Introduced ATCO^2 project and the main goals
- We are working on improving our callsign detection system
- Context information from OSN helps!
- Introduced each sub-system:
  - Data recording: VHF receivers
  - Speech-to-text: Hybrid and end-to-end systems
  - Text-to-callsign system: Based on BERT (Transformer neural network)

Thank you. Questions?

Twitter: @PabloGomez3

The 8th OpenSky Symposium

References

Cordero, José Manuel, Manuel Dorado, and José Miguel de Pablo. "Automated speech recognition in ATC environment". In: *Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems*. 2012, pp. 46–53.

Delpech, Estelle et al. "A Real-life, French-accented Corpus of Air Traffic Control Communications". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

Ferreiros, J et al. "A speech interface for air traffic control terminals". In: *Aerospace Science and Technology* 21.1 (2012), pp. 7–15.

Kleinert, Matthias et al. "Semi-supervised adaptation of assistant based speech recognition models for different approach areas". In: *37th Digital Avionics Systems Conference (DASC)*. IEEE. 2018, pp. 1–10.