Recognizing ATC English and Airbus ATC challenge

Jan “Yenda” Trmal, Lubos Smidl, Jan Svec
CLSP, Johns Hopkins University
NTIS, University of West Bohemia

Outline

Introduction to Air Traffic Control (for ASR purposes)

Airbus ATC Challenge

Some lesser-known techniques in Kaldi

Diagnostic techniques in Kaldi

Our team’s results

Conclusion

What ATC means?

- ATC = Air Trafic Control

Wikipedia: ATC is a service provided by ground-based air traffic controllers who direct aircraft on the ground and through controlled airspace, and can provide advisory services to aircraft in non-controlled airspace.

- To prevent collisions, organize and expedite the flow of air traffic
- To provide navigation assistance or information services and other support to pilots

ATC may issue **instructions** that pilots are required to obey, or **advisories** (known as flight information in some countries) that pilots may, at their discretion, disregard.

Areas of ATC

Generally, ATC can be divided into several broad areas.

- Ground control
  - all taxiways, inactive runways, holding areas and other areas of the airport
- Tower control
  - active runway surfaces (take-offs, landings, go-arround)
- Approach control (arrival/departure control)
  - handling traffic within a certain radius of the airport
- Area (lower/upper) control
  - provides air traffic control in the large area of controlled airspace (sometimes covering the whole of the country)

Areas of ATC (cont’d)

- Flight data and clearance delivery
  - To ensure the plane has correct information about weather and aircraft conditions
  - + some other types of restrictions and information specific for the given flight
  - ATIS -- recorded continuous loop on a specific frequency
    - weather changes, outages, airport ground delays/ground stops, runway closures

Communication in ATC

Data/text communication (CPDLC: controller-pilot data link communications)

Voice communication over radiotelephony

- Addressing: call-signs
  - Written: 3-letter combination plus flight number, assigned by ICAO
    - E.g. KLM785, AAL872
  - Radiotelephony (audio) callsigns
    - Speedbird 832 instead of BAW832

Callsigns in radiotelephony

Note that the airline telephony designator is different from the ICAO airline designator:

Airline commercial name	ICAO designator	Telephony designator
Lufthansa	DLH	Lufthansa
Thai Airways International	THA	Thai
Delta Airlines	DAL	Delta
Air France	AFR	Airfrans
British Airways	BAW	Speedbird
Southern Airways Express	FDY	Friendly

- full call sign shall be used in every message, be it from a controller or a pilot
- In reality shortened after first contact or can be even changed by the controller
- Prefixes and suffixes: **mayday** Speedbird seven niner, Speedbird seven niner **heavy**

- CONTROLLER: *Speedbird seven niner zero, contact metro tower one one niner decimal two*
- PILOT: *seven niner zero good day*
- CONTROLLER: *good day*

Table and the communication example from the Airbus ATC Challenge Starter pack, Technical note n° X23D18011900, Author: Estelle DELPECH (AIRBUS)

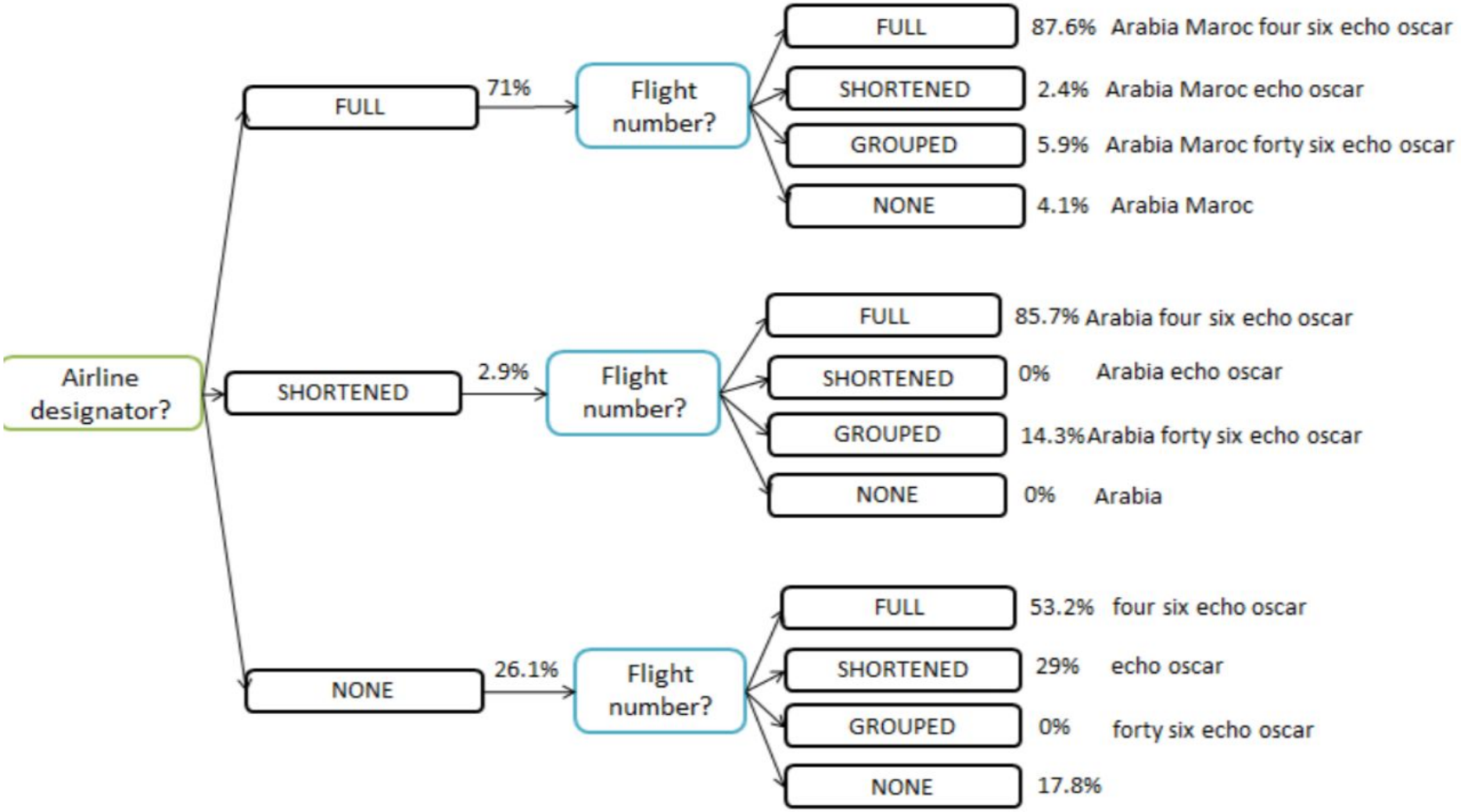
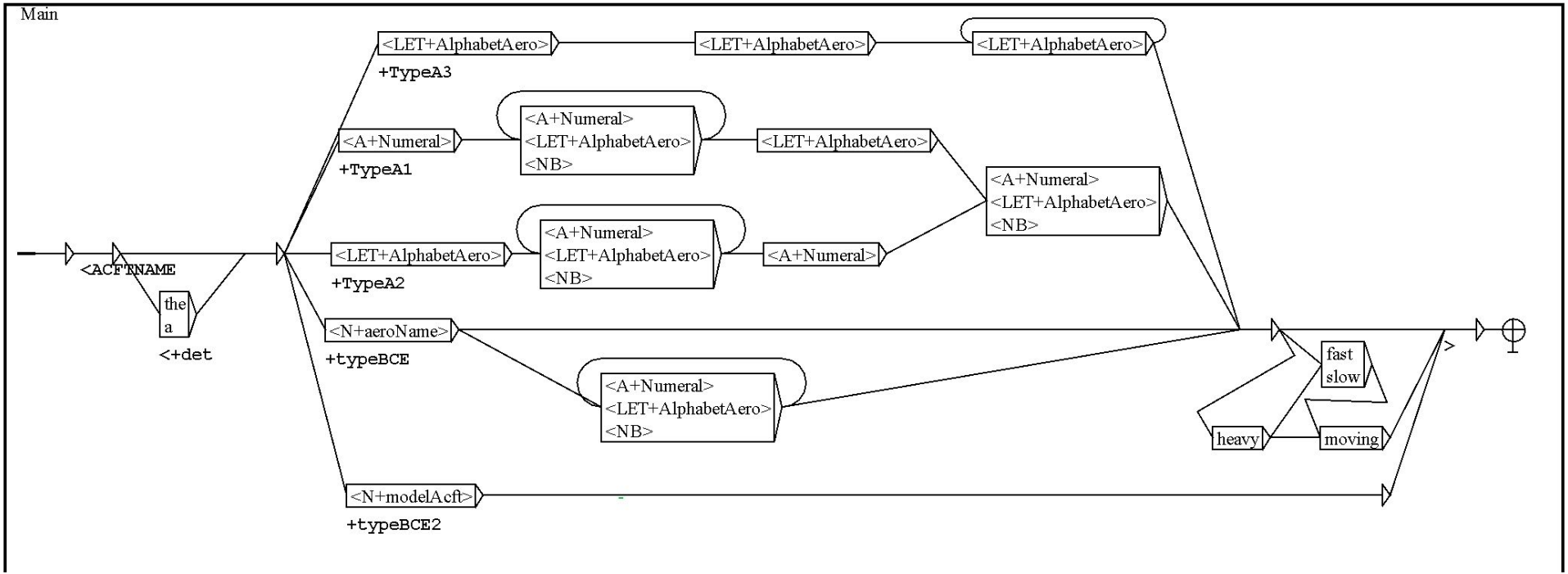


Figure 5 - Call signs in pilots' messages (Le Goff et al., 2017)

Image from the Airbus ATC Challenge Starter pack, Technical note n° X23D18011900, Author: Estelle DELPECH (AIRBUS)

Callsign grammar example



Doesn’t include dictionaries and how to insert it into a sentence

From ATC challenge started pack

ATC English

Domain specific English

Primarily to enhance communication between the air traffic controller and the pilot in command

- Extremely simplified and concise, defined by phraseology
  - To save radio bandwidth
  - To avoid ambiguities (3 = tree, 9 = niner)
- Use of slang/jargon is discouraged/prohibited
- Very carefully specified meaning of some words
  - roger (your message has been received)
  - wilco (I will comply with your request)

Conditional clearance to cross the intermediate runway:

Conditional phrases, such as "**behind** landing aircraft" or "**after** departing aircraft", shall not be used for movements affecting the active runway(s), except when the aircraft or vehicles concerned are seen by the appropriate controller *and* pilot. The aircraft or vehicle causing the condition in the clearance issued shall be the first aircraft/vehicle to pass in front of the other aircraft concerned.

NB: Beware - the ICAO phrase **'behind** instruction to 'get close to' the preceding incidents.

READ-BACK

Read-back is vital for ensuring mutual understanding between the pilot and the controller of the intended plan for that aircraft.

- Following correct read-back the flight crew must ensure that they carry out the correct action. Statistics show that one of the most common causes of a level bust in Europe is correct read-back followed by **incorrect** action.

On first contact with subsequent frequencies include **call-sign (and wake turbulence category if 'heavy')** and:

- **Level** , including passing and cleared level if not maintaining the cleared level
- **Cleared level** (if different from **current level**)
- **Speed** (if assigned by ATC), and
- **Other ATC** clearances assigned.

ng down the clearance
ew members listen to all
heck!

Almost grammar?

Syntax

CLEARED (or PROCEED) TO (significant point, name of facility or fix) [MAINTAIN (or CLIMB or DESCEND TO) (level)] HOLD [(direction)] [(specified) RADIAL, COURSE, INBOUND TRACK (three digits) DEGREES] [RIGHT (or LEFT) HAND PATTERN] [OUTBOUND TIME (number) MINUTES] EXPECT APPROACH CLEARANCE (or FURTHER CLEARANCE) AT (time) (additional instructions, if necessary).

Examples

LGL123 proceed to SKI, maintain FL100, hold two zero zero radial, course inbound track zero two zero degrees left hand pattern, outbound time three minutes expect further clearance at time five one.

https://contentzone.eurocontrol.int/phraseology/Phrase.aspx

Example communication

RTF En-Route Examples

Big Jet 345, fly heading 260 (degrees), climb to FL 100, no speed restrictions

Fly heading 260 (degrees), climb to FL 100, no speed restrictions, Big Jet 345

Big Jet 345, fly direct BONNY, climb to FL 360

Direct BONNY, climb to FL 360, Big Jet 345

Big Jet 345, contact Northern Control, 132.6

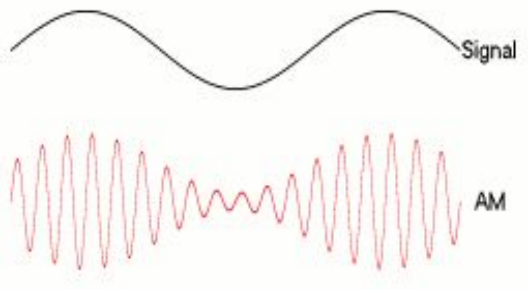
Contact Northern Control, 132.6, Big Jet 345

Northern Control, Big Jet 345, passing FL240 climbing to FL 360, direct BONNY

Big Jet 345, Northern Control, fly direct CLYDE

Direct CLYDE, Big Jet 345

Audio in ATC



- Communication typically over VHF (30 MHz to 300 MHz)
  - 108–118 MHz: Air navigation beacons VOR and Instrument Landing System localizer.
  - 118–137 MHz: Airband for air traffic control, AM, 121.5 MHz is emergency frequency
  - Sometimes additional country-dependent frequencies
  - Source: https://en.wikipedia.org/wiki/Very_high_frequency
- For speech (118kHz-137kHz), 25kHz or newly 8.33kHz is used per one channel, yielding 760 or 2280 channels (to resolve congestion)
- amplitude modulation (AM)
- 3-4kHz bandwidth for speech
- Still analog, different aircrafts, control towers and other users transmit with different bandwidths and audio characteristics

From: https://en.wikipedia.org/wiki/Airband

Examples of communication

ATIS example

Airbus ATC ASR Challenge

Airbus in collaboration with IRIT, ENAC and SAFETY DATA-CFH

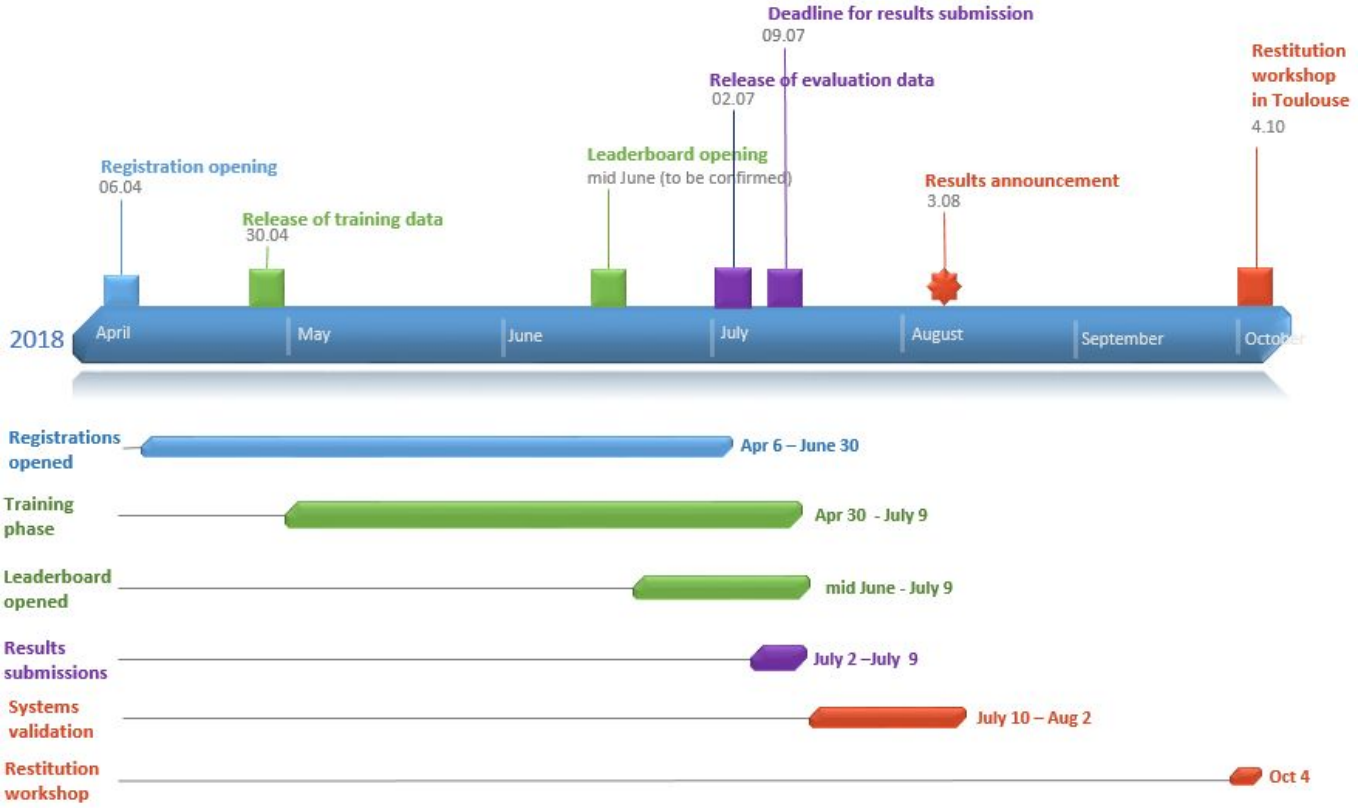
Purpose

- identify top players in the field for future collaboration;
- encourage research community to tackle the specific issues of ATC (accent / noise / code-switching robustness, highly specific vocabulary).

Tasks

- Standard audio transcription (evaluated with WER)
- Call sign extraction (evaluation with Precision/Recall/F1)

Challenge Schedule



Data

- 40h of transcribed, non-native, noisy, English, reconstituted ATC data
- +5h for validation/leader board (no transcriptions made available)
- +5h for final evaluation (no transcriptions made available)
- **you can use your own additional data**

Data segmented per utterance

No speaker info

Random filenames

No dev data (we split train to 35 train and 5 dev)

Scoring

Scoring metrics WER and F1,

BUT for F1:

Before comparison, hypothesis and reference text are set to lower case, characters other than letters _ and @ symbols are eliminated. Words are extracted by splitting on spaces. Then, call signs shall match exactly to be considered as correct. Inclusion or overlap between reference and hypothesis are considered as errors.

From: Air Traffic Control Speech Recognition Scientific Challenge Rules, https://s3-airbusaigym-web-1.s3.amazonaws.com/chal_speechtotext_rules_d8tQ42CmYE.pdf

hypothesis	reference
Nostrum seven six four	**Air Nostrum** seven six four
Easy two **zero** two bravo	Easy two **o** two bravo
sierra victor foxtrot eight one **huh** six	sierra victor foxtrot eight one six

Some of the problems (huh, @, _) resolved during the course of the challenge

Keyword search on the lattice not usable?

Scoring, continued

Similarly, for WER

‘@’ maps for 1 or more foreign words

I.e. we will need to handle OOVs well?

Quality zero five four sierra huh @ → where @ equals to *“alerte huh relief vérifiez votre altitude”* (terrain obstacle check your altitude)

In the example below, a pilot utters a sentence in English except for the numbers which are enunciated in French:

cleared for takeoff huh three two right Quality @ hotel → where @ equals to “*cent vingt quatre*” (hundred twenty four)

From: Delpeh et al.:A Real-life, French-accented Corpus of Air Traffic Control Communications

	Code in reference	Match in hypothesis
Foreign word in reference	@	@ or _
Not intelligible word in reference	_	@ or _ or any word

From: Air Traffic Control Speech Recognition Scientific Challenge Rules, https://s3-airbusaigym-web-1.s3.amazonaws.com/chal_speechtotext_rules_d8tQ42CmYE.pdf

Participants evaluation ranking (leaderboard scores)

The primary metrics are F1 and WER

For the leaderboard ranking, the harmonic mean was computed

$$Harm\ mean = \frac{2 * Norm\ pACC * Norm\ F1}{Norm\ pACC + Norm\ F1}$$

Where Norm pAcc and Norm F1 are normalized WER and F1 (so that the performer having the best performance has Norm pAcc = 1.0 or Norm F1 = 1.0

I.e. we won't see the performance of the other teams, just the relative ranking

Transcriptions

126546		one two nine decimal three one zero four four five
136654		toulouse Zulu airbus huh one seven four
146546		contact bordeaux one three three seven seven five goodbye
1564		air algerie one zero seven seven contact bordeaux one three three seven seven five goodby
1664556	one three three seven seven five	contact bordeaux one three three seven seven five goodbye
176456	zero zero bye	contact Toulouse one eight eight zero zero zero bye
18654	easy one zero seven eight	easy one zero seven eight contact Paris one two three decimal four seven zero

- all in lower case
- figures written in letters
- no punctuation
- gap fillers transcribed as ‘huh’
- unknown words transcribed with character ‘_’ (underscore)
- foreign (i.e. not English) words transcribed with character ‘@’ (at sign) or ‘_’ (underscore)
- no contractions (won’t -> will not)
- Noise ‘#’

Lexicon preparation

Out of 2500 types in the training list, around 500 were typos.

We checked against CMUdict

- Fixed manually typos
- Generated french pronunciation for french words (cities) using espeak + manually created table IPA->ARPAbet
- Verified specific words do exist (ATC terminology, waypoints)
- Trained G2P for correct, words not present CMUdict (phonetisaurus)
- Added ‘huh’ pronunciations (from WSJ)
- Two possible <UNK>: unknown word ‘_’ and foreign word (or phrase) ‘@’

Language Modeling

Used srilm toolbox

- 3-gram perplexity: 8.0
- 4-gram perplexity: 5.0 (MaxEnt LM, used for rescoring)

RNNLM didn’t help

Add perplexity:

> Rule of thumb says that for perplexity X, the LM has to, on average, choose between X alternatives. More precisely, it’s an weighted branching factor.
>
> For normal human language you’d see perplexity in 100 to 400 (depends on many factors, esp. vocabulary size).

Additional data available ?

Youtube channels (approx 100 hrs recordings)

LiveATC

- Fan-driven community page containing recordings of communication from various airports
- downloaded around 150k hours of recordings
- FR, CZE, SW, US, CAN accents

UWB corpus (proprietary corpus of approx 200hrs of CZE accented ATC)

Various sites with additional aux info: phraseology, spelling, aviation-safety, manuals, planecrashinfo, quora, skytalk, tailstrike

Handling the <UNK>

- Typically, detecting UNKs is fairly hard task
- Normally, you'd see something like this in a lexicon
  - <UNK> <unk>
  - I.e. word '<UNK>' maps to a single unit '<unk>'
  - This way, the training procedure is able to use the sentence for training, but the model of '<UNK>' won't be very good
- For decoding, it is a better idea to replace the pronunciation of '<UNK>' by a phoneme graph
  - Either all probabilities constant
  - Or you can train a LM on alignment of the training data

```
49 if [ $stage -le 4 ] ; then
50   utils/lang/make_unk_lm.sh data/local/dict exp/make_unk
51
52   utils/prepare_lang.sh \
53     --unk-fst exp/make_unk/unk_fst.txt --phone-symbol-table data/lang/phones.txt \
54     data/local/dict "<UNK>" data/local/lang_test data/lang_test
```

Handling the <UNK>

First idea: map both ‘_’ and ‘@’ to <UNK>

Second idea from listening to audio: map ‘@’ to <FOREIGN> with pronunciations of French greetings.

```
637 <FOREIGN> b oh n jh uw r ah
638 <FOREIGN> b ah n sh uh r
639 <FOREIGN> b oh n jh uw r
640 <FOREIGN> b aa n
641 <FOREIGN> b oh n
642 <FOREIGN> jh uh r n ea
643 <FOREIGN> ow r ah v w aa
644 <FOREIGN> ow r ah v w aa r
645 <FOREIGN> oh r eh v uh aa r
```

```
17 if [ $stage -le 4 ] ; then
18   utils/prepare_lang.sh \
19     --unk-fst exp/make_unk/unk_fst.txt --phone-symbol-table data/lang/phones.txt \
20     data/local/dict_foreign/ "<UNK>" data/local/lang_foreign_test data/lang_foreign_test
21
22   utils/format_lm.sh \
23     data/lang_foreign_test data/srilm_foreign/best_3gram.gz data/local/dict_foreign/lexicon.txt data/lang_foreign_test
24
25   utils/build_const_arpa_lm.sh \
26     data/srilm_foreign/best_4gram.gz data/lang_foreign_test data/lang_foreign_test_fg
```

Adding <FOREIGN>

- Hypothesis easy to test -- generate new lexicon and decoding graph, decode again
  - Make sure you use the '--phone-symbol-table' parameter for make_lang.sh
- Can we train? Remember not all <FOREIGN> can be salutations
  - Yes, we can
  - Utterances that fail the alignment will get removed automatically

```
25 LOG (gmm-align-compiled[5.4]:main():gmm-align-compiled.cc:127) 9kfucS6dFjWZkUyw
26 LOG (gmm-align-compiled[5.4]:main():gmm-align-compiled.cc:127) 9kgI3C8GTbsNv65m
27 WARNING (gmm-align-compiled[5.4]:AlignUtteranceWrapper():decoder-wrappers.cc:466) Retrying utterance 9kgI3C8GTbsNv65m with beam 40
28 WARNING (gmm-align-compiled[5.4]:AlignUtteranceWrapper():decoder-wrappers.cc:475) Did not successfully decode file 9kgI3C8GTbsNv65m, len = 5453
```

- Too many utterances dropped? Add line
  <FOREIGN> <foreign>
  Into the lexicon (and into phone list)

Pronunciation probabilities

- Most lexicons do not specify which pronunciation variant is more probable.
- For some words, the silence is more probable than after other (this probability is not modeled by LM)
- We can use our alignments to estimate these probabilities
- In practice, the conditional silence probability seems to be more important

PRONUNCIATION AND SILENCE PROBABILITY MODELING FOR ASR

Guoguo Chen[1], *Hainan Xu*[1], *Minhua Wu*[1], *Daniel Povey*[1,2], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

guoguo@jhu.edu, hxu31@jhu.edu, mwu56@jhu.edu, dpovey@gmail.edu, khudanpur@jhu.edu

Pronunciation probabilities

```
229 steps/get_prons.sh --cmd "$train_cmd" data/train_nodup data/lang_nosp exp/tri3b
230
231 utils/dict_dir_add_pronprobs.sh --max-normalize true \
232   data/local/dict_nosp exp/tri3b/pron_counts_nowb.txt exp/tri3b/sil_counts_nowb.txt \
233   exp/tri3b/pron_bigram_counts_nowb.txt data/local/dict
234
235 utils/prepare_lang.sh data/local/dict "<unk>" data/local/lang data/lang
```

1. First, get the stats from the alignments (of the training data)
2. Create a new dict dir
3. Generate lang directory the usual way
4. Add G.fst and regenerate decoding graph (not shown)

Data cleanup

- The transcribed data will often contain transcription errors, the segments are not correct, audio can be so noisy, that it causes harm using it…
- Idea: recognize using biased LM and use only those parts that were recognized correctly
- Used fairly often in kaldi egs
- Typically done before DNN training to get nice/correct alignments
- Script local/run_cleanup_segmentation.sh

```
45  # This does the actual data cleanup.
46  steps/cleanup/clean_and_segment_data.sh --stage $cleanup_stage \
47    --nj $nj --cmd "$cmd" \
48    $data $langdir $srcdir $dir $cleaned_data
```

Acoustic model

- Chain model (LF-MMI), factorized TDNN
- 12-layer, dim=1280, bottleneck=256, dropout
- Unconstrained egs
- Data cleanup (10 % of the data thrown away)
- Data augmentation: volume and speed (final system had 5-way, but performed only marginally better than “standard” 3-way)
- i-vectors (fairly small gain), tested two-pass i-vector estimation, again very tiny gain
- UNK = 4-gram phoneme loop
- Online decoder

Results

Baseline 9.28

+	Cleanup	9.02
+	iVectors	8.98
+	Pronprobs	8.83
+	LM Rescoring	8.45
+	<FOREIGN>	7.69
+	Two-stage ivectors	~0.03 (not included)
+	7-way augmentation	~0.00

Debugging a trained system

Assumes scoring using steps/score_kaldi.sh (or steps/scoring/score_kaldi_cer.sh)

```
jtrmal@t01 /export/b18/jtrmal/airbus/2/exp/chain_cleaned/tdnn4c_noivec_sp_online_p/decode_dev_fg/scoring_kaldi/wer_details $ ls -al
total 1564
drwxr-xr-x 2 jtrmal fax     110 Sep  7 03:28 .
drwxr-xr-x 7 jtrmal fax     155 Sep  7 03:28 ..
-rw-r--r-- 1 jtrmal fax       3 Sep  7 03:28 lmwt
-rw-r--r-- 1 jtrmal fax  160657 Sep  7 03:28 ops
-rw-r--r-- 1 jtrmal fax     545 Sep  7 03:28 per_spk
-rw-r--r-- 1 jtrmal fax 1417697 Sep  7 03:28 per_utt
-rw-r--r-- 1 jtrmal fax      49 Sep  7 03:28 wer_bootci
-rw-r--r-- 1 jtrmal fax       4 Sep  7 03:28 wip
```

- ops -- information about operation (in Levenshtein algorithm), i.e. about Insertions, Deletions, Substitutions
- pet_utt -- per utterance information, mostly decoded vs reference alignment
- per_spk -- scores per speaker (in our case not very interesting)

ops -- information about Levenshtein operations

```
 815 deletion              huh              ***      472
 816 deletion              and              ***       60
 817 deletion               to              ***       42
1144 insertion             ***              huh      148
1145 insertion             ***               to       38
1146 insertion             ***              one       35
1404 substitution         nine            niner       91
1405 substitution        niner             nine       62
1406 substitution        clear          cleared       55
1407 substitution      cleared            clear       46
1408 substitution      descent          descend       34
1409 substitution      descend          descent       20
1410 substitution           to              two       19
1411 substitution        three              two       15
1412 substitution          two            three       13
1413 substitution    dewpoint            point       12
1414 substitution          fox          foxtrot       12
1415 substitution          and               on        9
1416 substitution       descend       descending       9
```

per_utt -- information from alignments

```
5 2binqpq2ujm4nwsr ref  easy  i  call  you  back  huh  ***  ***  topswiss
6 2binqpq2ujm4nwsr hyp  easy  i  call  you  back  huh  top  swiss    is
7 2binqpq2ujm4nwsr op     C   C    C    C     C    C    I    I       S
8 2binqpq2ujm4nwsr #csid 6 1 2 0
```

```
jtrmal@t01 /export/b18/jtrmal/airbus/2/exp/chain_cleaned/tdnn4c_noivec_sp_online_p/decode_dev_fg/scoring_kaldi/wer_details $ grep csid per_utt | column -t |
sort -k6,6g | tail -n 3
jegh84mnnmhlnugd  #csid  0  1  0  11
vknuqq3hzjqldhto  #csid  6  5  0  12
nvnlfbk1vw0ldevd  #csid  0  0  0  18
```

```
jtrmal@t01 /export/b18/jtrmal/airbus/2/exp/chain_cleaned/tdnn4c_noivec_sp_online_p/decode_dev_fg/scoring_kaldi/wer_details $ grep vknuqq3hzjqldhto per_utt
vknuqq3hzjqldhto ref  huh  yes  we  will  see  how  we  get  on  in  the  next  couple  of  miles  huh  should  not  be  too  much  an  issue
vknuqq3hzjqldhto hyp  huh  yes  *** ***  ***  ***  we  do   not  have  the  latest  couple  and  miles  ***  ***  ***  ***  ***  ***  ***  ***
vknuqq3hzjqldhto op     C    C    D   D    D    D    C   S    S    S    C    S       C    S    C    D    D    D    D    D    D    D    D
vknuqq3hzjqldhto #csid 6 5 0 12
```

Debugging lexicon

```
jtrmal@t01 /export/b18/jtrmal/airbus/2  $ steps/cleanup/debug_lexicon.sh --nj 32 --cmd "retry.pl queue.pl" --alidir exp/tri5_ali_cleaned/ data/train_cleaned d
ata/lang_test exp/tri5_ali_cleaned/ data/local/dict/lexicon.txt exp/debug_lexicon
```

```
;; <count-of-this-pron> <rank-of-this-pron> <frequency-of-this-pron> CORRECT|INCORRECT <word> <pron>
657.0 1 0.24 INCORRECT <UNK> "<sil>", ref = "<unk>"
103.5 1 0.25 INCORRECT descent "d ih s eh n d", ref = "d ih s eh n t"
28.0 1 0.42 INCORRECT want "w ah n", ref = "w aa n t | w ah n t | w ao n t | w oh n t"
12.5 1 0.46 INCORRECT taxiing "t ae k s ih ng", ref = "t ae k s ih ih ng | t ae k s iy ih ng"
11.5 1 0.36 INCORRECT Nouvelair "n ow v eh l ea r", ref = "n uw v eh l ea r"
11.5 1 0.29 INCORRECT Clermont "k l eh r m aa n", ref = "k l eh r m aa n t"
11.0 1 0.17 INCORRECT triple "th r ih p ah l", ref = "t r ih p ah l | t r ih p l"
```

Call-sign detection

- 2-layer bidirectional LSTM
- Training data
  - Recognized ASR hypothesis with ground truth callsign (alignment!)
- LSTM tagging
  - Output classes: no CS, beginning of CS, middle of CS, end of CS
- Expert knowledge (word classes) by additional embedding layer
  - Company name, numbers, spelling alphabet
- Ensembling to average over different initializations of LSTM training

Leaderboard results

team	rank	score	entries
sharp-pup	1	1.00	20
intent-bear	**2**	**0.98**	**30**
known-squid	3	0.98	12
nearby-gecko	4	0.98	24
decent-skink	5	0.97	3
	...		
	27		

- WER: 0.0934
- F1: 0.8340

Final results

	RANK	SCORE	WER_rank	WER	F1_rank	F1
sharp-pup	1	1.0	1	0,0763	1	0.8241
sharp-perch	2	0.97	2	0.0842	3	0.7939
nearby-gecko	3	0.96	5	0.0941	2	0.8017
intent-bear	**4**	**0.95**	**4**	**0.0876**	**5**	**0.7704**
known-squid	5	0.94	6	0.0955	4	0,7762

Conclusion

- ATC english is a specific style of english
- Very noisy, accented
- Single online decode (+rescoring) ranked #4 in the challenge.
- Introduced some of the lesser known techniques and diagnostic tools for Kaldi

Thanks for your attention