# Airbus Air Traffic Control Speech Recognition Challenge

*Julien Despres, Jodie Gauvain, Viet-Bac Le, Igor Swiecicki, Lori Lamel†, Jean-Luc Gauvain†*

Vocapia Research & CNRS-LIMSI†

October 4, 2018 Toulouse

# CNRS-LIMSI

- Speech processing research at CNRS-LIMSI since the late 1970s
- Participation in many national and international research projects and technology evaluations
- Innovative research (early uptake): unsupervised learning, neural networks for acoustic and language modeling
- License of speech processing technology to Vocapia Research in 2002

# CNRS-LIMSI: Our new building

# Vocapia Research

# Vocapia Research

- R&D company and software publisher
- Founded in 2000
- Privileged partnership with CNRS-LIMSI
- Leading-edge multilingual speech processing technologies
- VoxSigma® software suite: Audio segmentation, Speaker diarization, Language identification, Multilingual speech-to-text transcription, Spoken term detection Speech-text synchronization

# Speech recognizer

Our standard approach:

- PLP-like front-end (use since 1997)
- Audio partitionner (use since 1998)
- HMM-NN phone models (use MLP since 2008)
- N-gram LM
- Optional NN LM (use since 2002)
- Dynamic network decoder
- Consensus decoding

# Speech recognizer

Our standard approach:

- PLP-like front-end (use since 1997)
- Audio partitionner (use since 1998)
- HMM-NN phone models (use MLP since 2008)
- N-gram LM
- Optional NN LM (use since 2002)
- Dynamic network decoder
- Consensus decoding

**But small system (small training corpus)**

# Training and development data

- Used only the challenge ATC data
- Tried to use other ATC data (LDC, Clarin ATCC, LiveATC) without success
- Corrected evident errors in manual transcripts of training data
- Shared typos with other participants via forum
- Defined a heldout dev set from train (700 utterances, 1h of data)
- Completed annotation of call signs in our heldout dev

# Language model

- 4-gram language model
- 2.2K words including a word <foreign> for non-English speech
- Only 140K 4-grams
- Perplexity of heldout dev with 4g LM: 8.4

# Language model

- 4-gram language model
- 2.2K words including a word <foreign> for non-English speech
- Only 140K 4-grams
- Perplexity of heldout dev with 4g LM: 8.4

  LM much smaller than usual, estimated on much less data
  - Typically > 1B words of texts and >100K audio transcripts
  - 100k-200k word vocabulary, 50M n-grams

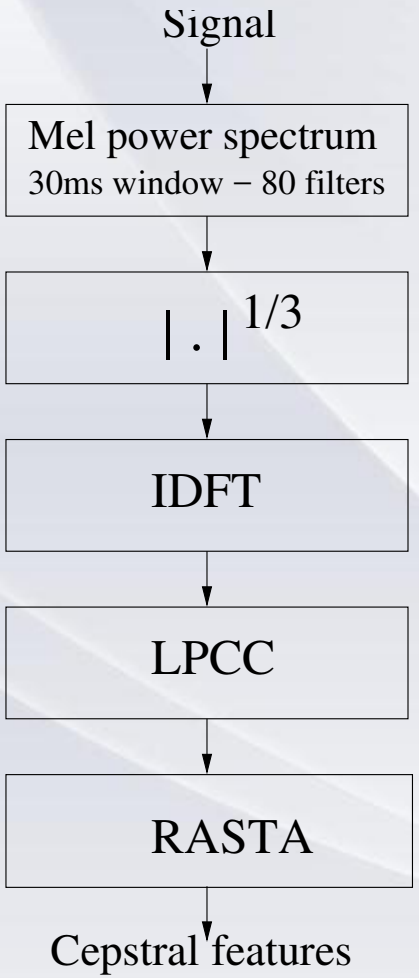
- NN model with word embeddings (first use in ASR by LIMSI 2002)
- Tried MLP (2 hidden layers) and RNN
- Both gave just a small gain on our dev set

# Pronunciation lexicon

- Pronunciations based on LIMSI American English pronunciation dictionary (1996)
- 48 phones, including 3 non-speech symbols
- Pronunciation probabilities
- 2200 words
- Specific prononciations for certain words: letters (Q Quebec), nine/niner, three/tree
- Average 1.3 pronunciations/word

# Front-end

- PLP-like front-end (since 1997)
- 40 cepstrum coefficients
- Rasta filtering



# HMM-DNN acoustic model

- 2100 phones in context (monophones, diphones, triphones)
- About 2K HMM tied-states
- Divisive decision tree clustering using a set of 184 questions about phone position in word, classes of current and surrounding phones
- Maximize training data likelihood using single Gaussian models
- DNN with only 6M parameters

# Decoder

- Our standard decoder dynamic network decoder can directly use a 4-gram LM and supports real-time decoding
- No warmup time
- Generate a word graph for each utterance, which is then used in consensus decoding (to optimise WER instead of SER) or to rescore with more complex LMs

# Speech recognition results

| Test set | heldout dev | leaderboard-dev | eval |
|---|---|---|---|
| WER (del, ins) | 6.2% (2.5, 1.3) | 8.1% | 7.6% |

- WER on heldout dev lower than other data sets (part of train)
- WER on eval lower than leaderboard dev (better verification?)
- Would be nice to know del, ins on leaderboard and eval data

# Call sign detection

- Call signs should adhere to a pre-defined format

  **AIRLINE CODE  NUMBERS (3-5)  LETTERS (0-2)**

- Explored two pattern matching methods:
  - Consensus network search with confidence scores
  - Regular expressions using 1-best hypothesis
- Detect only one call sign per segment
- Try to detect full call signs first, and if unsuccessful, look for partial matches

# Where is the call sign?

Example 1:

- Pilot: five thousand one zero zero seven november four tango

# Where is the call sign?

Example 1:

- Tower: november four tango continue descent altitude five thousand feet and QNH one zero zero seven
- Pilot: five thousand one zero zero seven november four tango

# Where is the call sign?

Example 1:

- Tower: november four tango continue descent altitude five thousand feet and QNH one zero zero seven
- Pilot: five thousand one zero zero seven november four tango

Example 2:

- Pilot: one eighteen one one eight one three zero nine juliet bye

# Where is the call sign?

Example 1:

- Tower: november four tango continue descent altitude five thousand feet and QNH one zero zero seven
- Pilot: five thousand one zero zero seven november four tango

Example 2:

- Tower: Easy three zero nine juliet speed one eight zero knots contact tower one one eight decimal one goodbye
- Pilot: one eighteen one one eight one three zero nine juliet bye

# Where is the call sign?

Example 1:

- Tower: november four tango continue descent altitude five thousand feet and QNH one zero zero seven
- Pilot: five thousand one zero zero seven november four tango

Example 2:

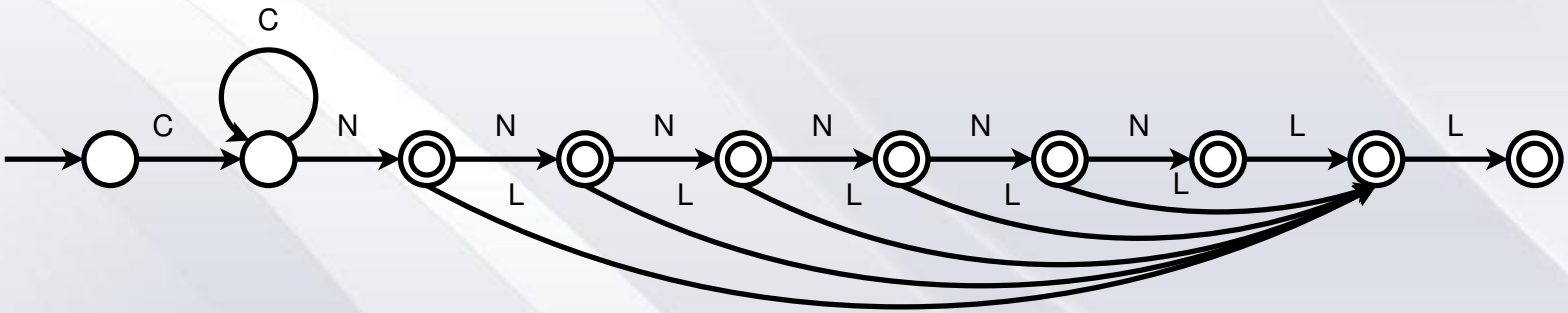
- Tower: Easy three zero nine juliet speed one eight zero knots contact tower one one eight decimal one goodbye
- Pilot: one eighteen one one eight one three zero nine juliet bye

Need to identify and filter possible false matches: runway, taxiway, radio frequency, flight level, QFE, QNH, ...

# Call sign detection

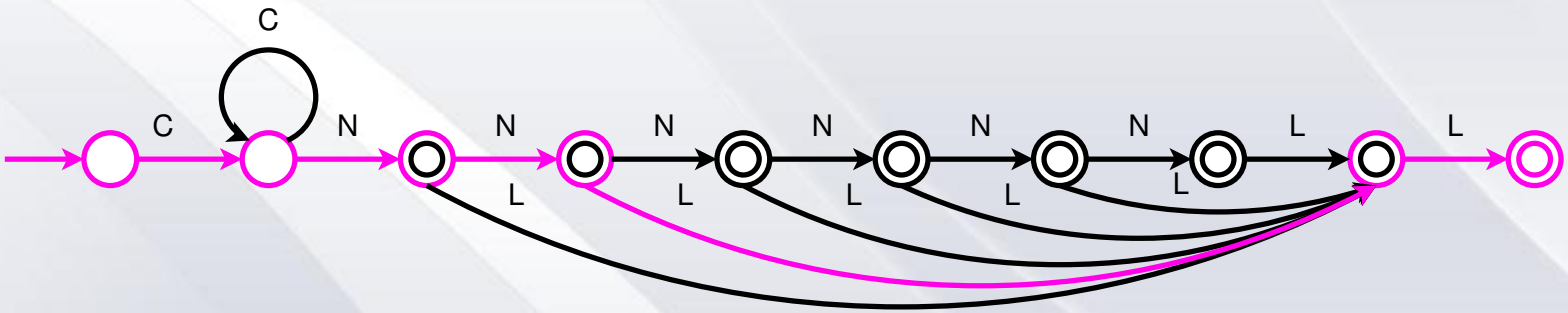
- Classes
  - C – company names (Aeroflot, Airbus, Ryanair, ...)
  - L – aeronautical letters (alpha, bravo, charlie, ...)
  - N – numbers (one, two, three, ...)
- Classes are slightly different for two methods
  - Consensus network search (CNS):
    compound words included in C, N contains only digits
  - Regular-expression based (RE):
    no compounds, just individual words in C, N contains some numbers
    10, 100, ...

# Call sign detection - RE Full call signs




Graph covers all full call signs

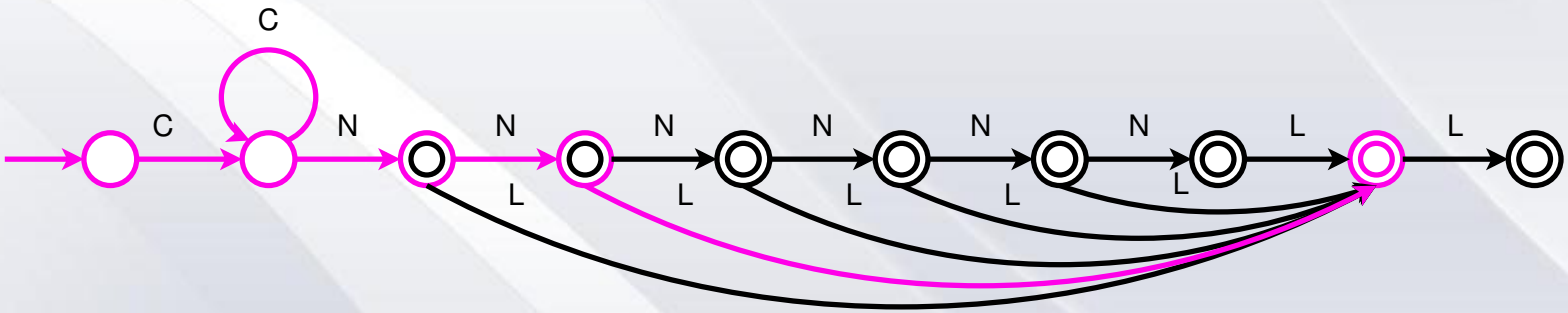
# Call sign detection - RE Full call signs




Ryanair niner five whiskey uniform

C N N L L

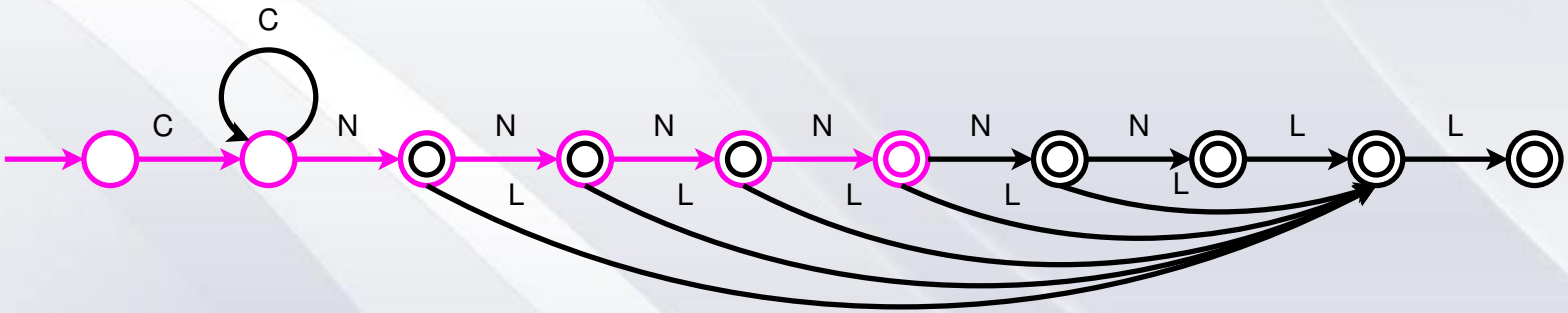
# Call sign detection - RE Full call signs




Enac School six seven november

C C N N L

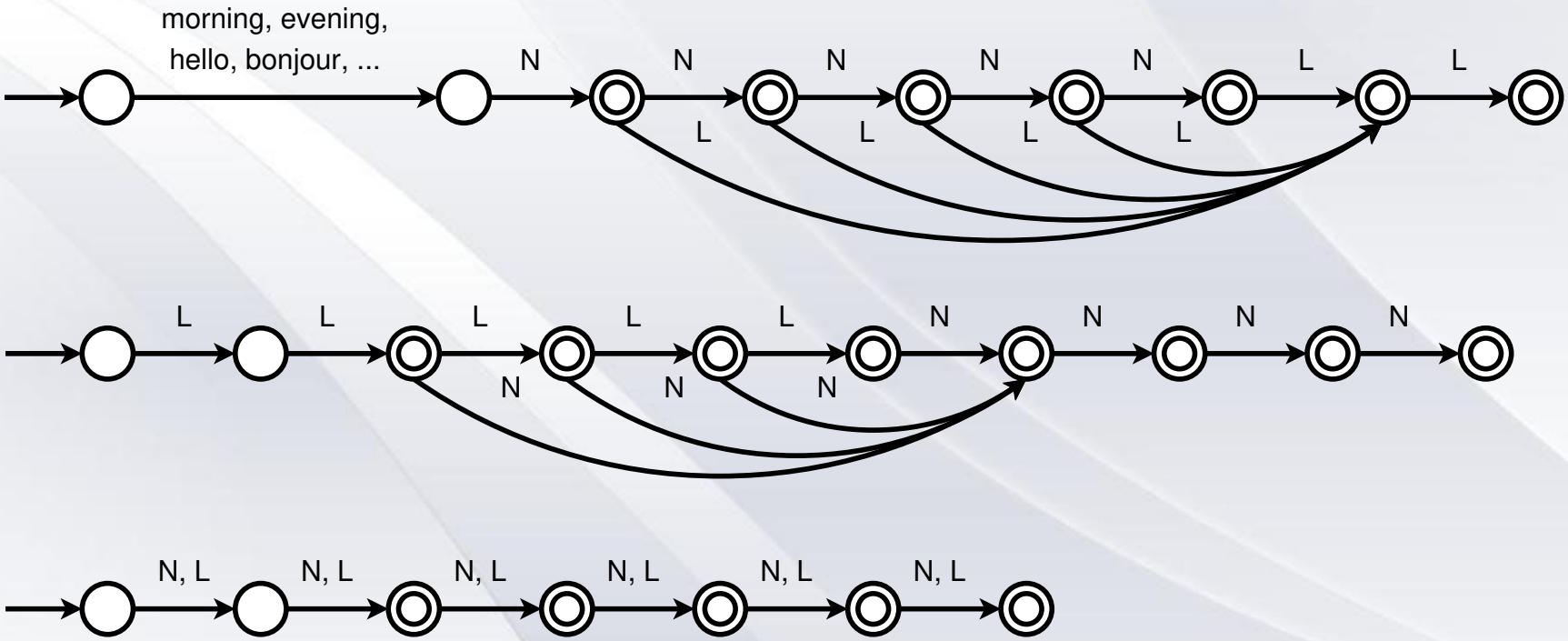
# Call sign detection - RE Full call signs



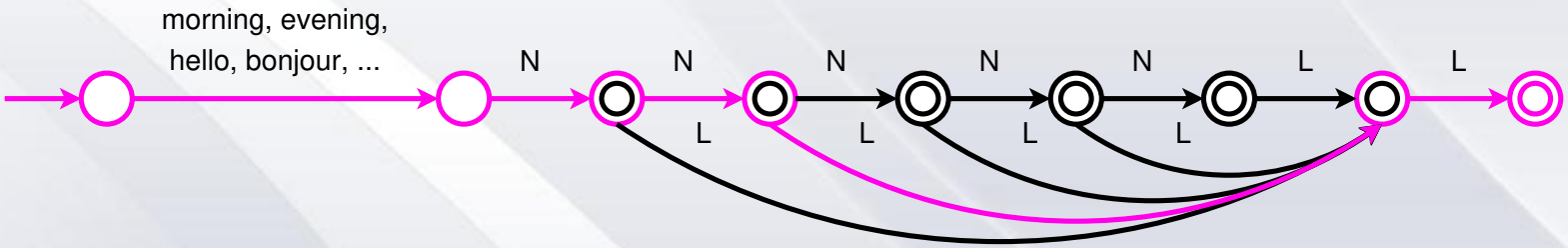

Germania one eight eight one

C N N N N

# Call sign detection - RE partial match



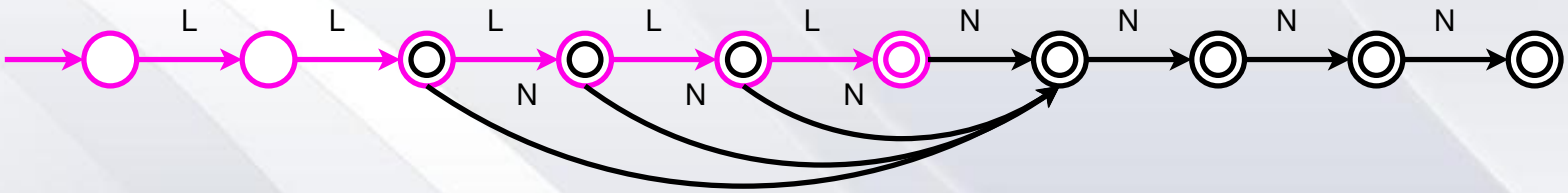
# Call sign detection - RE partial match




morning six seven yankee november

x N N L L

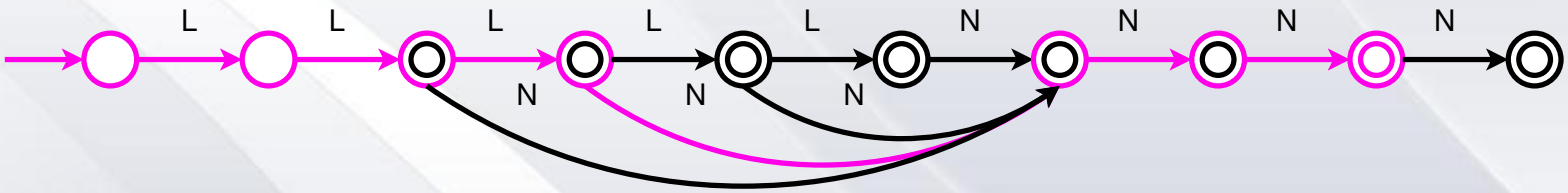
# Call sign detection - RE partial match




delta alpha juliet oscar yankee

L L L L L

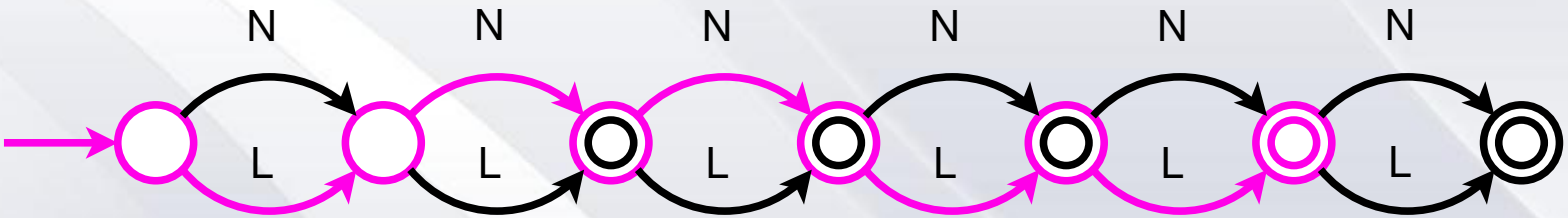
# Call sign detection - RE partial match




sierra victor foxtrot eight one six

L L L N N N

# Call sign detection - RE partial match




november six one alpha juliet

L N N L L

# Call sign detection via consensus network search

Pattern matching with a predefined call sign list

- Extracted a list of 1500 call signs from training data
- Each call sign is considered as a pattern (with a prior probability)
- Perform search on the confusion networks
- Advantages:
  - Confidence scores from ASR are taken into account
  - Low detection of false alarm
- Drawback: cannot detect "unseen" call signs
- Also experimented with a more generic version

# Call sign detection - Results

| Call sign detection | heldout dev | | lead-dev | eval |
|---|---|---|---|---|
| | full only | full+partial | full only | full+partial |
| CNS (full) | 0.8962 | 0.8559 | 0.8501 | 0.8132 |
| RE(full+partial) | 0.8105 | 0.9012 | 0.7858 | 0.8176 |
| CNS+RE | 0.8207 | 0.9038 | 0.7899 | 0.8241 |

| WER | 6.2% (2.5, 1.3) | 8.1% | 7.6% |
|---|---|---|---|

- Only full call signs annotated in training data
- Partial matches increase false alarms

# Call sign detection - Results

| Call sign detection | heldout dev | | lead-dev | eval |
|---|---|---|---|---|
| | full only | full+partial | full only | full+partial |
| CNS (full) | 0.8962 | 0.8559 | 0.8501 | 0.8132 |
| RE (full+partial) | 0.8105 | 0.9012 | 0.7858 | 0.8176 |
| CNS+RE | 0.8207 | 0.9038 | 0.7899 | 0.8241 |

| WER | 6.2% (2.5, 1.3) | 8.1% | 7.6% |
|---|---|---|---|

- Similar trends seen on leaderboard dev

# Call sign detection - Results

| Call sign detection | heldout dev | | lead-dev | eval |
|---|---|---|---|---|
| | full only | full+partial | full only | full+partial |
| CNS (full) | 0.8962 | 0.8559 | 0.8501 | 0.8132 |
| RE (full+partial) | 0.8105 | 0.9012 | 0.7858 | 0.8176 |
| CNS+RE | 0.8207 | 0.9038 | 0.7899 | 0.8241 |

| WER | 6.2% (2.5, 1.3) | 8.1% | 7.6% |
|---|---|---|---|

- RE detects partial call signs in heldout dev (recall we manually corrected these)
- Best result using CNS combined with RE for partial match

# Call sign detection - Results

| Call sign detection | heldout | | lead-dev | eval |
|---|---|---|---|---|
| | full only | full+partial | full only | full+partial |
| CNS(full only) | 0.8962 | 0.8559 | 0.8501 | 0.8132 |
| RE (full+partial) | 0.8105 | 0.9012 | 0.7858 | 0.8176 |
| CNS+RE | 0.8207 | 0.9038 | 0.7899 | 0.8241 |

| WER | 6.2% (2.5, 1.3) | 8.1% | 7.6% |
|---|---|---|---|

- Eval data does not follow heldout or leaderboard dev
- Could be that eval chosen to have mainly full call signs or partial not completely annotated
- Would be nice to have FA and miss on leaderboard dev and eval data

# Some other ideas we explored

- Unsupervised AM training using the dev (no difference on internal heldout dev or dev on leader board)
- Generating texts from provided ATC transcripts
- Using other ATC data (LDC, Clarin ATCC, live ATC)
- Live ATC was often French, so not well suited to task at hand
- Bilingual English/French system
- Enhanced audio partitioner

# Bilingual system

- Given the nature of the ATC data (accented English + some foreign speech, mostly French) we built a bilingual English/French system
- Method:
  1. Decode with English system $\longrightarrow$ English word graph
  2. Decode with French system $\longrightarrow$ French word graph
  3. Merge word graphs and generate a bilingual confusion network
- Note that the language can switch at each word
- Since non-English words are not scored in this challenge, also tried using a single model for all foreign words <foreign>
- For this challenge, the bilingual model gave no significant gain over the simple <foreign> method
- The bilingual model can be used to process other data (e.g., live ATC)

# Conclusions and some observations

- Challenge was a good start, but more issues to address
- Need to handle wide range of native and non-native accents
- Handling multiple languages, bilingual systems
- Localization of transcription system, e.g., destinations, waypoints
- Dynamic information use: what planes are currently in, entering and leaving airspace
- Wider task: coverage of all airports in France (or Europe) will certainly require some of the things we tried that did not help for this challenge

Thank you to Airbus and IRIT for organizing this challenge

Thank you to Airbus and IRIT for organizing this challenge

and

thank you for your attention

Questions?

# Audio partitioning

- Partitioner designed to process a continuous audio stream
- "Audio Stream Mixture Model" (introduced in 1998) to segment the audio and cluster the segments
- On the ATC data, this step does not impact the results even with an enhanced partitioner with specific speech/non-speech models trained using the LDC ATC corpus
- Expect this step will be important in a system targeting more general data
- Enhanced partitioner performs better than the original partitioner on Live ATC data

# Call sign detection

Two stage call sign detection

- Generalized pattern matching approach
- Define 3 lists: COMPANY, NUMBER and LETTER
- Perform a first search using these 3 lists
- Define call sign templates with prior probabilities. Eg:

| | | |
|---|---|---|
| Callsign01 | $p1$ | COMPANY NUMBER LETTER LETTER |
| Callsign02 | $p2$ | COMPANY NUMBER NUMBER |
| Callsign03 | $p3$ | COMPANY NUMBER NUMBER LETTER |

- Perform a 2nd search with callsign template
- Advantages:
  - Confidence scores from ASR can be taken into accout
  - More flexible: can easily add new call sign templates and new companies
- Drawback: Detection is less robust than using predefined lists