Automated Speech Recognition in Controller Communications applied to Workload Measurement

Third SESAR Innovation Days
27th November 2013
Stockholm, Sweden

José Manuel Cordero – CRIDA
jmrcordero@e-crida.aena.es

José Miguel de Pablo - CRIDA
Manuel Dorado – AENA
Natalia Rodríguez-CRIDA
General Overview

• Final objective of the system:
  - **Workload measurement**
  - *in an automated way*
  - *in operation environment*

• Dual approach:
  - Automated voice controller events detection
  - Underlying technology (ATC semantic speech recognition)

“Unlocks” the way for ASR applications in operation
Setting the scenario: what is this all about?

- ASR in ATM has proved to be very challenging

- Various reasons:
  - Immaturity of natural speech recognition technology
  - Separation from standard ICAO phraseology
  - Multilingual
  - Need of a highly reliable system (less than that may even increase workload)
  - Difficult to access to real ATC communications
  - High user expectations (and growing!)

- Applications mainly in Simulation environment, until recently
Setting the scenario: A long story short

- **AENA: Initial research around 2006**
  - Pseudo pilots scheme in real-time simulation environment

- **Extremely difficult in initial stages to achieve effective speech recognition**
  - COTS didn’t provide acceptable detection rates (under 30%)
  - For simulation purposes, the integration with the ATC Platform allowed to mitigate the problem (however, speech recognition itself was poor)

- **Decission to make a “non-contextual information” approach**
• What does it mean “non contextual information”?
  – No integration with ATC Platform, so no information of Flights to help on detection (*standalone ASR*)

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Independency from ATC Platform (easy adaptation)</td>
<td>-Increases difficulty of detection (wider constellation)</td>
</tr>
<tr>
<td>-Usable (as a service) in many other applications</td>
<td>-Requires more training/modelling to get similar results</td>
</tr>
<tr>
<td>-Better ASR</td>
<td></td>
</tr>
</tbody>
</table>

– However, ATM logic is inside the detection model (even the scenario can be included).
Setting the scenario: On the other side, some strengths

- **Wide set of real ATC communications available**

- **Close collaboration with operational staff (trainers)**
  - Validation/calibration
  - Event model refinement
  - Language interpretation

- **Reoriented objective**: *Workload estimation by voice recognition (in real operation recordings)*
  - Calculation through detected controller events
  - Voice is an essential source of information
The underlying technology: ATC Event Detection Functional Scheme

- **Preprocessing**
- **Segmentation/Labelling** (silence removal, …)
- **Speech recognition** (HMM)
  - Language Model
  - Acoustic Model
  - Extensive training
- **CS detection/Event detection**->**Algorithms**, keywords+logic
- **Postprocessing/Refining**
- **XML (Output)**
The enabler: System Training

- The ASR Module needs to be trained with transcriptions (from real ATC communications)
- Transcriptions are very time-consuming and done manually -> Transcription-aid strategy
- Current prototype contains more than 100 net (no-silence transcribed) hours (both en-route and TMA), with 100% reliability (human check), corresponding to aprox. 500 raw hours (with silences)

Evolution strategy: limit 100% accuracy manual transcriptions, use those with automated confidence index >95% -> Improvement in WDR
Automation Architecture

- Sector configuration in CWP to be extracted from the ATC system
- VoIP recording (NICE System)
- Double Workload calculation based on controller events (Wickens/MWM and NORVASE)
The output

• 1 XML file per sector, per hour, combining channels: set of events

```
- <event>
  <name>Av</name>
  <callsign>ANE8390</callsign>
  <UCE>Ruta 15</UCE>
  <date>17/09/2011</date>
  <time>08:11:33</time>
  <duracion>4</duracion>
  <coord>No</coord>
  <message>air nostrum ocho tres nueve cero descienda a nivel de vuelo uno cinco cero</message>
</event>
- <event>
  <name>Av</name>
  <callsign>ANE8390</callsign>
  <UCE>Ruta 15</UCE>
  <date>17/09/2011</date>
  <time>08:11:46</time>
  <duracion>3</duracion>
  <coord>No</coord>
  <message>air nostrum ocho tres nueve cero esta establecido cuarenta y cinco millas</message>
</event>
```

Level Change Event

Communication Automated Transcription
• Metrics for an ASR applied to Workload estimation:

  - **WDR:** Word Detection Rate. Is more usual to find WER (Word Error Rate = substitutions + deletions + insertions/total real words), WDR = 100 - WER

  - **EDR:** Event Detection Rate (An event is considered correct when type of event and CS are OK)

  - **EDR_{no callsign}:** Event Detection Rate without callsign (Only considers event categorisation)

  - **FPR:** False Positives Rate
Some numbers (results)-Feb 2013

- Results obtained from a set of 60 raw hours not included in the training of the system (control group) (6591 events)

<table>
<thead>
<tr>
<th></th>
<th>WDR</th>
<th>EDR</th>
<th>FPR</th>
<th>EDR_callsign</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-route</td>
<td>67.3%</td>
<td>74.6%</td>
<td>5.8%</td>
<td>95.9%</td>
</tr>
<tr>
<td>Approach</td>
<td>69.8%</td>
<td>72.5%</td>
<td>5.3%</td>
<td>91.2%</td>
</tr>
<tr>
<td>Overall</td>
<td>68.9%</td>
<td>73.5%</td>
<td>5.6%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>

- Rates better than any other known product applied to ATC communications, continuously evolving

- Later on, the workload calculation can be performed using diverse methodologies
A look to workload measurement

• After the events are obtained, they are cross-checked with those detected from pure FP and radar data.

• A set of events for a period of time is determined, and send to two different workload calculation modules:
  – MWM (MultiWorkload Model, based on Wickens cognitive workload model
  – NORVASE (Sector Validation Normative), based on Spanish normative
A look to workload measurement

• NORVASE is particularly relevant as automation of the workload measuring process allows a bigger number of samples for all sectors, thus increasing the accuracy of the measure (versus manual takes, very limited and selective).

  • More workload samples
  • More sectors measured
  • Fully automated
  • Cost efficient
Which events necessarily need voice?

- Focus is put in three of them, where voice analysis is key for effective event detection:
  
  i) **Direct/heading determination**
  - From simple radar data analysis is difficult to determine
  - Voice is the most reliable source
  - Mistake in this determination has a big impact in workload

  ii) **Effective sector exit**
  - Radar data allows geographical exit, but not frequency transfer
  - Key as the moment when the event happens is relevant for workload
  - Only obtainable through voice analysis

  iii) **Inter-sector coordination**
  - Unavailable from any other source
## Events Detected (rates Feb 2013)

<table>
<thead>
<tr>
<th>Event code</th>
<th>Event Description</th>
<th>$EDR_{\text{no callsign}}$</th>
<th>Com. duration (s)</th>
<th>Ocurr</th>
<th>Ocurr</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTEv</td>
<td>Sector Entry Communication</td>
<td>96.2%</td>
<td>3.1, 1.8</td>
<td>33.26%</td>
<td>17.88%</td>
</tr>
<tr>
<td>Csv</td>
<td>Sector change Communication to Pilot</td>
<td>98.5%</td>
<td>3.9, 1.8</td>
<td>32.42%</td>
<td>19.87%</td>
</tr>
<tr>
<td>Dv</td>
<td>Direct Communication</td>
<td>92.1%</td>
<td>2.7, 0.9</td>
<td>2.01%</td>
<td>0.33%</td>
</tr>
<tr>
<td>Xv</td>
<td>Heading Type 1 Communication</td>
<td>90.1%</td>
<td>3.7, 1</td>
<td>0.13%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Sv</td>
<td>Heading Type 2 Communication</td>
<td>91.5%</td>
<td>4.6, 2.9</td>
<td>0.13%</td>
<td>28.48%</td>
</tr>
<tr>
<td>Vv</td>
<td>Speed change Communication</td>
<td>94%</td>
<td>3.3, 0.9</td>
<td>1.34%</td>
<td>7.28%</td>
</tr>
<tr>
<td>Av</td>
<td>Level change Communication</td>
<td>96.6%</td>
<td>1.8, 2</td>
<td>17.38%</td>
<td>9.38%</td>
</tr>
<tr>
<td>Cov</td>
<td>Inter-sector controller-controller coordination</td>
<td>79.7%</td>
<td>7.8, 7.6</td>
<td>8.56%</td>
<td>2.32%</td>
</tr>
<tr>
<td>Ac3.4.11v</td>
<td>Clearance or instruction Communication</td>
<td>93%</td>
<td>3.3, 1.3</td>
<td>0.87%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Ac7v</td>
<td>ILS Authorization Communication</td>
<td>91.3%</td>
<td>3.6, 1.4</td>
<td>0.53%</td>
<td>3.64%</td>
</tr>
<tr>
<td>Ac13.1v</td>
<td>STAR assignment Communication</td>
<td>90%</td>
<td>3.9, 2.5</td>
<td>N/A</td>
<td>0.53%</td>
</tr>
<tr>
<td>Ac9v</td>
<td>Essential information Communication</td>
<td>80%</td>
<td>5.8, 5.3</td>
<td>2.41%</td>
<td>8.83%</td>
</tr>
<tr>
<td>H1v</td>
<td>Holding stack Communication</td>
<td>87.2%</td>
<td>2.3, 0.8</td>
<td>0.40%</td>
<td>1.10%</td>
</tr>
<tr>
<td>CRv</td>
<td>Clearance/authorization Correction communication</td>
<td>88.8%</td>
<td>2.4, 1.1</td>
<td>0.67%</td>
<td>0.33%</td>
</tr>
</tbody>
</table>

**Model optimised for en-route detection**

En-route: 43.25% events voice detection has a key role for workload

TMA: 51% events voice detection has a key role for workload
What’s next?

• As stated, underlying technology unlocks and enables the way for new applications

• SESAR Exercise EXE-04-07.01-VP-003, “Resolving Complexity by dynamic management of airspace”

• V2 exercise, OFA05.03.04

• Voice will be analysed on-line for complexity indicators calculation, using the same ASR technology described.
Conclusions

• Automated measurement of controller workload, based on ASR, in operation environment

• Dual approach: Application and enabler

• Set of events provided, for later WL calculation in two modules

• Nice EDR (around 75%), very good EDR if not considering callsigns (over 90%)
Conclusions (II)

- Voice information key element in >40% events en-route, >50% events in TMA

- Need to evolve some detection algorithms (especially callsigns)

- Plan to include Airports (2014)

- Final integration with ATC system for virtual 100% EDR (2014-2015)

- Other applications now feasible (even virtual pseudo-pilots)
Thank you!

Any questions?

Centro de Referencia I+D+i ATM