Read Back Error Detection using Automatic Speech Recognition

Shuo Chen
Hunter Kopald
Ronald S. Chong
Dr. Yuan-Jun Wei
Zachary Levonian

June 27, 2017
Outline

3. Introduction

4. Background

Concept and Design

5

6

7

Speech Recognition Evaluation

8

9

10

11. 'Case Study'

12. Takeaways
After a controller issues a clearance or instruction to a pilot, he or she must then “ensure that items read back are correct, ensure the read back of hold short instructions…, and ensure pilots use call signs and/or registration numbers in any read back acknowledging an air traffic clearance or ATC instruction”

– FAA Order JO 7110.65W
Background

Automatic read back error detection is a popular idea with both current and former air traffic controllers and managers.

The numbers on read back/hear back errors:

In the 1990’s, Cardosi et al. reported:
- ≤ 1% read back error rates in the tower domain
- 40% of read back errors were uncorrected
- On average, one read back error per hour per frequency across all domains

More recent study by Prinzo et al. reported:
- 6% read back error rates in the TRACON domain
- 92% hear back error rate

Eurocontrol found read back hear back errors were the most common controller-pilot communication problem.

Kopald and Goring found hear back errors accounted for 10.7% of runway incursions in a 6-year runway incursion dataset – 129 runway incursions in 6 years.

The risk of a runway incursion caused by a hear back error is very small, but the potential consequences of a single incursion may still warrant capabilities to mitigate this risk.
Concept and Design

A simple and intuitive design:
1. Passively monitor controller and pilot communications over the radio
2. Match transmission pairs, by ACID recognition and/or dialogue model
3. Extract commands for comparison
4. Indicate when a read back contains a mismatch or is absent altogether

Designed a graphical user interface to satisfy the need for the following basic information, identified by subject matter experts:

- Aircraft identifier
- Indication that a read back error has occurred
- What the discrepancy was
- An audible alert if the read back error is not corrected

Key decision for the automation: is the read back error worth alerting?
Alert Rules

Not all read back errors are created equal!

An analysis of 150 dialogues between controller and pilots at John F. Kennedy Airport (JFK) highlighted many interesting nuances...

[CALLSIGN], kennedy ground, 31L KE, taxi right A, hold short of J.

And A, short of J, for 31L KE, [CALLSIGN].


Okay, behind the [CALLSIGN-1] from the right, follow her down K for 04L, and 19 1 to monitor, [CALLSIGN-0]

[CALLSIGN], cross runway 31L at K, and monitor tower 19.1.

Roger, cleared to cross runway 13L at K, and then monitor tower 119.1, [CALLSIGN].

[CALLSIGN], cross runway 31L at K, and monitor tower 19.1.

[Uh] Roger, cross runway 13L at K, and then monitor tower 1191, [CALLSIGN].

[CALLSIGN], it’s runway 31L at K, cross runway 31L at K.

Sorry, runway 31L at K then, my apologies, [CALLSIGN], thank you.
Alert Rules (continued)

Summarizing observations and conclusions from the dialogue analysis:

- **Order of elements** in read back does not need to match the original clearance ordering
- **Transfer of control/frequency**: controllers do not require facility name and frequency in the read back, and an acknowledgement of the transfer may suffice
- **Callsigns**: pilots almost always used their full callsign but controllers do not require a full or partial callsign when the identity of the speaker is unambiguous
- **Taxi commands**: pilots regularly dropped the left/right turn specification
- **Follow**: pilots almost always acknowledged the “follow” instruction with something semantically equivalent (e.g., “after”, “behind”).
- **Crossing instructions**: controllers almost always used the form “cross <runway> at <intersection>” and pilots always provide some form of read back. However the controller does not always require a complete read back.
- **Hold Short**: pilots almost always correctly read back both the hold short instruction and position and controllers may always require a correct and complete read back to this command

Not all discrepancies in a read back warrant an alert. A good design will consider the risk associated with the read back error.
Three types of acoustic models were evaluated:

- A base English acoustic model released by CMUSphinx
- Adaptations of the English acoustic model, enhanced using controller and pilot recordings and transcriptions
- Custom acoustic models created using only controller and pilot recordings

Two statistical language models were created:

- One trained using transcriptions of local and ground controller radio transmissions
- One trained using transcriptions of pilot radio transmissions heard at the pilot and ground controller positions

A semantic parse algorithm was used to extract command concepts and aircraft identifiers that were robust to word variations.

**Speech Recognition Evaluation**

**Digitized Audio Signal**

**Acoustic Model**

**Basic Speech Sounds**

\[ \text{r} \quad \text{ʌ} \quad \text{n} \quad \text{w} \quad \text{e} \quad \text{i} \quad \text{w} \quad \text{ʌ} \quad \text{n} \]

**Language Model**

**Speech Hypothesis**

\[ \text{three hold short of runway one niner right} \]

**Command Concept**

\[ \text{Hold Short 19R} \]
Speech Recognition Evaluation

Model Comparison

Word Error Rate (WER) was used as a rough performance measure to compare different speech recognition tuning methods and configurations.

Word Error Rate = omissions + additions + substitutions

- Custom acoustic model performed best
- Bigger advantage on pilot speech as size of training set increases (see right)

But concept recognition is a better predictor of application performance

<table>
<thead>
<tr>
<th>Command Concept</th>
<th>Nominal Command Phrase</th>
<th>List of Parameter Alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTL</td>
<td>“cleared to land”</td>
<td>Runway</td>
</tr>
<tr>
<td>LUAW</td>
<td>“line up and wait”</td>
<td>Runway</td>
</tr>
<tr>
<td>CFT</td>
<td>“cleared for takeoff”</td>
<td>Runway</td>
</tr>
<tr>
<td>Hold Short</td>
<td>“hold short”</td>
<td>Runway or Taxiway</td>
</tr>
<tr>
<td>Cross</td>
<td>“cross”</td>
<td>Runway or Taxiway</td>
</tr>
<tr>
<td>Turn</td>
<td>“turn left” or “turn right”</td>
<td>Runway or Taxiway</td>
</tr>
<tr>
<td>Taxi</td>
<td>“taxi”</td>
<td>Runway(s) or Taxiway(s)</td>
</tr>
<tr>
<td>Continue</td>
<td>“continue”</td>
<td>Runway(s) or Taxiway(s)</td>
</tr>
</tbody>
</table>

- Test data consisted of 10 hours of audio recordings from JFK containing clearance and read back exchanges
  - 3,689 pilot transmissions
  - 2,970 controller transmissions

- Training data consisted of 127 hours of transcribed, silence-reduced audio from the air traffic control tower domain
  - Several facilities (including JFK)
  - Both controller and pilot speech
# Speech Recognition Evaluation

## Concept Recognition Accuracy

- Better concept recognition of controller speech than of pilot speech.
- ACID recognition in this test is poor; previously, we have demonstrated better results with restrictive context-informed grammar.
- The addition of dialogue context across transmissions could improve clearance to read back matching, despite low ACID recognition.

<table>
<thead>
<tr>
<th>Controller Transmissions</th>
<th>ACID</th>
<th>LUAW</th>
<th>CFT</th>
<th>CTL</th>
<th>Hold Short</th>
<th>Cross</th>
<th>Turn Right</th>
<th>Turn Left</th>
<th>Taxi</th>
<th>Continue</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Recognized Correctly</td>
<td>79%</td>
<td>94%</td>
<td>98%</td>
<td>91%</td>
<td>96%</td>
<td>96%</td>
<td>98%</td>
<td>97%</td>
<td>93%</td>
<td>97%</td>
</tr>
<tr>
<td>% Partially Recognized</td>
<td>NA</td>
<td>5%</td>
<td>1%</td>
<td>1%</td>
<td>3%</td>
<td>3%</td>
<td>2%</td>
<td>1%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>% Not Recognized</td>
<td>21%</td>
<td>1%</td>
<td>2%</td>
<td>7%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>2%</td>
<td>5%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pilot Transmissions</th>
<th>ACID</th>
<th>LUAW</th>
<th>CFT</th>
<th>CTL</th>
<th>Hold Short</th>
<th>Cross</th>
<th>Turn Right</th>
<th>Turn Left</th>
<th>Taxi</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>% Recognized Correctly</td>
<td>63%</td>
<td>86%</td>
<td>80%</td>
<td>84%</td>
<td>88%</td>
<td>92%</td>
<td>87%</td>
<td>84%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>% Partially Recognized</td>
<td>NA</td>
<td>7%</td>
<td>8%</td>
<td>3%</td>
<td>8%</td>
<td>4%</td>
<td>4%</td>
<td>5%</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>% Not Recognized</td>
<td>37%</td>
<td>8%</td>
<td>12%</td>
<td>13%</td>
<td>4%</td>
<td>5%</td>
<td>9%</td>
<td>11%</td>
<td>7%</td>
<td></td>
</tr>
</tbody>
</table>

But application performance driven by concept recognition in sequence, not aggregate (see next slide)
Case Study: Hold Short

A look at potential system performance with current speech recognition performance on the Hold Short instruction data set

- Alert logic only considered controller clearance and pilot readback pairs
- Required readback of both command and runway

220 Hold Short Instructions

199 Complete Readbacks
12 Partial Readbacks
9 Non-Readbacks

181 correctly did not trigger an alert
19 triggered a correct alert
18 triggered a false alert
2 did not trigger a necessary alert
2 did not trigger a necessary alert
Key Takeaways

- Not all read back discrepancies warrant an alert
- Key to application design is appropriate consideration of what warrants an alert
- Using real ATC voice data enables better speech recognition accuracy
- Recognizing pilot audio is particularly challenging
- Assessing the application requires both speech recognition accuracy and a set of alert rules, and the two must be set in consideration of each other
This work was produced for the U.S. Government under Contract DTFAWA-10-C-00080 and is subject to Federal Aviation Administration Acquisition Management System Clause 3.5-13, Rights In Data-General, Alt. III and Alt. IV (Oct. 1996). The contents of this document reflect the views of the author and The MITRE Corporation and do not necessarily reflect the views of the Federal Aviation Administration (FAA) or the Department of Transportation (DOT). Neither the FAA nor the DOT makes any warranty or guarantee, expressed or implied, concerning the content or accuracy of these views. ©2017 The MITRE Corporation. ALL RIGHTS RESERVED.

<table>
<thead>
<tr>
<th>Fiscal Year:</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome Number:</td>
<td>8</td>
</tr>
<tr>
<td>PBWP Reference:</td>
<td>8-1.1</td>
</tr>
</tbody>
</table>