Knowledge-Based Word Lattice Rescoring in a Dynamic Context

Todd Shore, Friedrich Faubel, Hartmut Helmke, Dietrich Klakow
Section I

Motivation
Motivation

- **Problem**: difficult to incorporate higher-level knowledge sources into automatic speech recognition (ASR)
Motivation

- **Problem:** difficult to incorporate higher-level knowledge sources into automatic speech recognition (ASR)

- **However:** there are domains in which the situational context of utterances is available, e.g. air traffic control (ATC) or command and control tasks
- **Problem:** difficult to incorporate higher-level knowledge sources into automatic speech recognition (ASR)

- **However:** there are domains in which the situational context of utterances is available, e.g. air traffic control (ATC) or command and control tasks

**Approach taken in this work:** incorporate contextual knowledge into ASR by rescoring word lattice output of an ATC task
Section II

The Air Traffic Control Task
The Air Traffic Control Task

- Air traffic controllers at their workstations
The Air Traffic Control Task

- Air traffic controllers at their workstations

- **Primary objective:**
  - maintain aircraft separation
  - safely guide approaching aircraft to their runway threshold
  - integrate departing and passing aircraft
The Air Traffic Control Task

- Air traffic controllers at their workstations

- Have access to:
  - radar screens revealing aircraft positions and speeds
  - flight plans
  - weather reports, indicators for speed of wind, etc.
The Air Traffic Control Task

- Air traffic controllers at their workstations

**Implementation:**
- issue verbal commands to aircraft pilots
- use of a standardized subset of English which is formally specified by the International Civil Aviation Organization (ICAO)
ICAO Phraseology for ATC commands comprises:

- single aircraft callsign (identifying the aircraft)
- goal actions to execute
- goal values to be achieved
The Air Traffic Control Task

- **ICAO Phraseology** for ATC commands comprises
  - single aircraft callsign (identifying the aircraft)
  - goal actions to execute
  - goal values to be achieved

<table>
<thead>
<tr>
<th>Type</th>
<th>Values</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESCENT</td>
<td>ALT</td>
<td><em>descend altitude</em> ALT feet</td>
</tr>
<tr>
<td>DESCENT</td>
<td>FL</td>
<td><em>descend flight level</em> FL</td>
</tr>
<tr>
<td>REDUCE</td>
<td>SPD</td>
<td><em>reduce speed</em> SPD knots</td>
</tr>
<tr>
<td>TURN</td>
<td>DIR, HDG</td>
<td><em>turn</em> DIR heading HDG</td>
</tr>
</tbody>
</table>
**The Air Traffic Control Task**

- **ICAO Phraseology** for ATC commands comprises
  - single aircraft callsign (identifying the aircraft)
  - goal actions to execute
  - goal values to be achieved

- **Example:** „*Delta four three niner turn right heading two two zero*“
The Air Traffic Control Task

- **ICAO Phraseology** for ATC commands comprises
  - single aircraft callsign (identifying the aircraft)
  - goal actions to execute
  - goal values to be achieved

- **Example:** "Delta four three niner turn right heading two two zero"

- **Relevant Information:**
  - DL 439
  - TURN, DIR=right, HDG=220
The Air Traffic Control Task

- **Use of a recognition grammar**
  - is sufficient for recognizing ICAO phraseology
  - simplifies extraction of semantic content
Use of a recognition grammar
- is sufficient for recognizing ICAO phraseology
- simplifies extraction of semantic content

Example: „Delta four three niner turn right heading two two zero“
The Air Traffic Control Task

- **Use of a recognition grammar**
  - is sufficient for recognizing ICAO phraseology
  - simplifies extraction of semantic content

- **Example:** “*Delta four three niner turn right heading two two zero*”

- **XML-tags for easy extraction of semantic frames:**
  
  ```xml
  <callsign> delta four three niner </callsign> 
  <cmd_turn> turn </cmd_turn> 
  <direction> right </direction> 
  <heading> two two zero </heading>
  ```
Knowledge-based word lattice rescoring in a dynamic context

- **Use of a recognition grammar**
  - is sufficient for recognizing ICAO phraseology
  - simplifies extraction of semantic content

- **Example:** „Delta four three niner turn right heading two two zero“

- **XML-tags for easy extraction of semantic frames:**
  
  ```xml
  <callsign> delta four three niner </callsign>
  <cmd_turn> turn </cmd_turn> <direction> right </direction> heading
  <heading> two two zero </heading>
  ```
The Air Traffic Control Task

- **Use of a recognition grammar**
  - is sufficient for recognizing ICAO phraseology
  - simplifies extraction of semantic content

- **Example:** „*Delta four three niner turn right heading two two zero***“

- **XML-tags for easy extraction of semantic frames:**


callsign> delta four three niner </callsign> cmd_turn> turn </cmd_turn> direction> right </direction> heading
<heading> two two zero </heading>
Use of a recognition grammar
- is sufficient for recognizing ICAO phraseology
- simplifies extraction of semantic content

Example: „Delta four three niner turn right heading two two zero“

XML-tags for easy extraction of semantic frames:

```xml
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>
```
The Air Traffic Control Task

- **Use of a recognition grammar**
  - is sufficient for recognizing ICAO phraseology
  - simplifies extraction of semantic content

- **Example:** "Delta four three niner turn right heading two two zero"

- **XML-tags for easy extraction of semantic frames:**

```
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>
```
The Air Traffic Control Task

Main idea of this work: rescore recognized utterances through use of context knowledge about the situation in the airspace
The Air Traffic Control Task

- **Main idea of this work:** rescore recognized utterances through use of context knowledge about the situation in the airspace

- But where does this information come from?
The Air Traffic Control Task

- Modernized ATC workspace as envisioned by the German Aerospace Center (DLR)
The Air Traffic Control Task

- including the arrival manager **4D-CARMA**, which assists controllers in managing aircraft arrivals
The Air Traffic Control Task

... including the arrival manager 4D-CARMA, which assists controllers in managing aircraft arrivals.

This system allows us to extract:
- the callsigns of aircraft in the airspace
- the aircraft positions relative to the radar
- Their speeds, altitudes, climb/descend rates, reduce rates, etc.
Section III
Knowledge-Based Lattice Rescoring
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- graphical representation of a lattice with states and transitions labeled with phonemes and words.
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly.

Diagram:

```
q_{21} \quad \langle\text{callsign}\rangle \quad q_{22} \quad \text{speedbird} \quad q_{23} \quad \text{eight} \quad q_{167} \quad \langle/\text{callsign}\rangle
```
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search.

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly.

![Diagram showing the flow of lattice with nodes and transitions labeled with words like `callsign`, `speedbin`, `eight`, etc. and a semantic template box labeled `callsign:`.]

Knowledge-based word lattice rescoring in a dynamic context
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly

![Diagram of a lattice with semantic frames](image)
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly

![Diagram](image-url)

- **callsign**: BA 8

- **speedbird**: eight

- **semantic template**

- **callsign**: BA 8
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- This allows us to directly extract semantic frames by simply adding a pointer to a semantic template structure, which is filled on the fly

```
q_{21} <\text{callsign}> \rightarrow q_{22} \quad \text{speedbird} \rightarrow q_{23} \quad \text{eight} \rightarrow q_{167} </callsign>
```

Semantic template:

```
callsign: BA 8 ←
```
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

**semantic template**
callsign: BA 8 ← disclosure representation
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea:** penalize invalid callsigns and unlikely command values based on discourse representation system

```
semantic template
callsign: BA 8 ←
```

check if exists
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea:** penalize invalid callsigns and unlikely command values based on discourse representation system

<table>
<thead>
<tr>
<th>semantic template</th>
</tr>
</thead>
<tbody>
<tr>
<td>callsign: BA 8 ←</td>
</tr>
</tbody>
</table>

- check if exists
- if no
- apply penalty
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea:** penalize invalid callsigns and unlikely command values based on discourse representation system

Semantic template:

- **Callsign:** BA 8

Edge information:

- Start / end time
- Acoustic / LM score
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

### Semantic Template
- **callsign**: BA 8

### Edge Information
- start / end time
- acoustic / LM score
- knowledge score
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

<table>
<thead>
<tr>
<th>semantic template</th>
</tr>
</thead>
<tbody>
<tr>
<td>callsign: BA 8 ←</td>
</tr>
</tbody>
</table>

discourse representation
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search.

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system.

**semantic template**
callsign: BA 8
cmd: REDUCE
val: SPD=220 ←

discourse representation
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea:** penalize invalid callsigns and unlikely command values based on discourse representation system

<table>
<thead>
<tr>
<th>semantic template</th>
<th>discourse representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>callsign: BA 8</td>
<td>check speed</td>
</tr>
<tr>
<td>cmd: REDUCE</td>
<td>spd=220 ←</td>
</tr>
</tbody>
</table>
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea:** penalize invalid callsigns and unlikely command values based on discourse representation system

<table>
<thead>
<tr>
<th>semantic template</th>
</tr>
</thead>
<tbody>
<tr>
<td>callsign: BA 8</td>
</tr>
<tr>
<td>cmd: REDUCE</td>
</tr>
<tr>
<td>val: SPD=220 ←</td>
</tr>
</tbody>
</table>

apply penalty

check speed
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

### Semantic Template

- **callsign**: BA 8
- **cmd**: REDUCE
- **val**: SPD=220

### Cost Function

\[
\partial_{kn}(q_i, q_j) = \sum_{k=1}^{K} \rho_k c_k (\text{callsign}, \text{cmd}, \text{val})
\]
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search.

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system.

**Semantic Template**
- callsign: BA 8
- cmd: REDUCE
- val: SPD=220

$$\partial_{kn}(q_i, q_j) = \sum_{k=1}^{K} \rho_k c_k \text{(callsign, cmd, val)}$$

cost: $$\varepsilon : \text{/speed}$$
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

```
semantic template

callsign: BA 8
cmd: REDUCE
tval: SPD=220
```

```
\[
\partial_{kn}(q_i, q_j) = \sum_{k=1}^{K} \rho_k \ c_k (\text{callsign, cmd, val})
\]

\[
\text{constraint penalty function } c_k (\cdot) \rightarrow [0,1]
\]
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Principal Idea**: penalize invalid callsigns and unlikely command values based on discourse representation system

**Semantic Template**
- callsign: BA 8
- cmd: REDUCE
- val: SPD=220

**Total Cost**
\[
\hat{c}(\cdot) = \omega_{ac}\hat{c}_{ac}(\cdot) + \omega_{lm}\hat{c}_{lm}(\cdot) + \omega_{kn}\hat{c}_{kn}(\cdot)
\]
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties

1. Propagate scores along the nodes
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties

![Diagram](image)
**Knowledge-Based Lattice Rescoring**

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties

```
\begin{align*}
\text{acScore} + = & \omega_{ac} \delta_{ac}(\cdot) \\
\text{lmScore} + = & \omega_{lm} \delta_{lm}(\cdot) + [?lp] \\
\text{knScore} + = & \omega_{kn} \delta_{kn}(\cdot)
\end{align*}
```
Knowledge-Based Lattice Rescoring

- **WFST Decoder:** can be used to generate phone-to-word transducer lattice during Viterbi search.

- **Rescoring:** in analogy to rescoring with different language model scales and word insertion penalties.

\[
\begin{align*}
\text{q}_{21} & \xrightarrow{\text{callsign}} \text{q}_{22} & \text{speedbin} & \xrightarrow{\text{eight}} \text{q}_{23} & \xrightarrow{\text{callsign}} \text{q}_{167}
\end{align*}
\]

- **Token:**
  - acScore
  - lmScore
  - knScore

\[
\begin{align*}
\text{acScore} & = \omega_{ac}\partial_{ac}(\cdot) \\
\text{lmScore} & = \omega_{lm}\partial_{lm}(\cdot) + [?/p] \\
\text{knScore} & = \omega_{kn}\partial_{kn}(\cdot)
\end{align*}
\]
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties

1. Propagate scores along the nodes
Knowledge-Based Lattice Rescoring

- **WFST Decoder**: can be used to generate phone-to-word transducer lattice during Viterbi search

- **Rescoring**: in analogy to rescoring with different language model scales and word insertion penalties

1. Propagate scores along the nodes
2. Trace back from best final state
Section IV

Rescoring Experiments
Rescoring Experiments

- **Corpus:**
  - recorded using the 4D-CARMA software from DLR
  - includes aircraft state vectors (5 second intervals)
  - total of 1,107 ATC commands
  - 9.5 words per sentence
  - approx. 100 minutes of speech
Rescoring Experiments

- GUI used by the participants
Rescoring Experiments

- **Recognition Grammar:**
  - branching factor of 7.68
  - 244,220 unique callsigns
  - sentence perplexity of 3.9E+09
Rescoring Experiments

- **Recognition Grammar:**
  - branching factor of 7.68
  - 244,220 unique callsigns
  - sentence perplexity of $3.9 \times 10^9$

- **Used Constraints:**
  - callsign
  - speed
  - altitude
Rescoring Experiments

- **Recognition Grammar:**
  - branching factor of 7.68
  - 244,220 unique callsigns
  - sentence perplexity of 3.9E+09

- **Used Constraints:**
  - callsign
  - speed
  - altitude

<table>
<thead>
<tr>
<th>Rescoring</th>
<th>WER</th>
<th>SER</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (baseline)</td>
<td>2.81</td>
<td>22.58</td>
<td>0.849</td>
</tr>
<tr>
<td>Callsign</td>
<td>0.55</td>
<td>4.61</td>
<td>0.966</td>
</tr>
<tr>
<td>Callsign, Spd, Alt</td>
<td>0.52</td>
<td>4.52</td>
<td>0.967</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.31</td>
<td>2.07</td>
<td>0.979</td>
</tr>
</tbody>
</table>
Section V

Conclusions
We have shown how dynamic context knowledge can profitably be used for rescoring ASR hypotheses.
We have shown how dynamic context knowledge can profitably be used for rescoring ASR hypotheses.

This is of particular interest in scenarios where explicit context information is available, such as:
- ATC: radar-derived aircraft state vectors
- video games
- virtual reality
Thank you very much for your attention!
The Air Traffic Control Task

- XML-tags for easy extraction of semantic frames:

```xml
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>
```
The Air Traffic Control Task

- XML-tags for easy extraction of semantic frames:

  `<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>`

- Can easily be extended to class-based N-grams

  `[callsign] [cmd_turn] [direction] heading [heading] .`
The Air Traffic Control Task

- XML-tags for easy extraction of semantic frames:

```
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>
```

- Can easily be extended to class-based N-grams

```
[callsign] [cmd_turn] [direction] heading [heading] .
```
XML-tags for easy extraction of semantic frames:

```xml
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading <heading> two two zero </heading>
```

Can easily be extended to class-based N-grams

```
[callsign] [cmd_turn] [direction] heading [heading] .
```
The Air Traffic Control Task

- XML-tags for easy extraction of semantic frames:

```xml
<callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading 
<heading> two two zero </heading>
```

- Can easily be extended to class-based N-grams

```plaintext
[callsign] [cmd_turn] [direction] heading [heading] .
```
The Air Traffic Control Task

- XML-tags for easy extraction of semantic frames:

  ```xml
  <callsign> delta four three niner </callsign> <cmd_turn> turn </cmd_turn> <direction> right </direction> heading
  <heading> two two zero </heading>
  ```

- Can easily be extended to class-based N-grams

  ```text
  [callsign] [cmd_turn] [direction] heading [heading] .
  ```