Speech technologies for interactive mobile applications – a primer

Jason D. Williams

AVIxD / IxD workshop – August 2011
Roadmap

• A quick tour of the toolbox
• Technology and usage issues (that influence design)
• A few pointers on getting started
What is a spoken dialog system?

- Converts audio to words
- Converts words to meaning
- Language understanding: Maintains dialog state; decides what to say or do
- Dialog manager
- Converts meaning to words
- Text to speech
- Speech recognition
- "Starbucks in Boston"
- "In Austin?"
- "starbucks in austin"
Speech recognition

Audio to phonemes

End-pointing

Feature Extraction

Pattern Classification (Decoding, Search)

Acoustic Model

Language Model

Word Lexicon

Confidence Scoring

“uh benefits in florida”

“uh benefits information”

“tomato: t o m a ^ o
t o m A ^ o”

[ (?(i live in) [
alaska~0.001
california~0.2
arkansas~0.01
...])]

Allowed or likely sequences of words
Language understanding

- Example input/output:

<table>
<thead>
<tr>
<th>TRANSCRIPTION (input)</th>
<th>INTERPRETATION (output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_m_l_a claim</td>
<td>&lt;call_request &quot;fmla&quot;&gt; &lt;action &quot;claim&quot;&gt;</td>
</tr>
<tr>
<td>payroll</td>
<td>&lt;ambig_key &quot;payroll&quot;&gt;</td>
</tr>
<tr>
<td>customer service</td>
<td>[none]</td>
</tr>
<tr>
<td>i'd [fragment] ~</td>
<td>[none]</td>
</tr>
<tr>
<td>f_m_l_a agent</td>
<td>&lt;call_request &quot;fmla&quot;&gt; &lt;action &quot;operator&quot;&gt;</td>
</tr>
<tr>
<td>employment</td>
<td>&lt;call_request &quot;employment&quot;&gt;</td>
</tr>
<tr>
<td>[side_speech]</td>
<td>[none]</td>
</tr>
<tr>
<td>more examples</td>
<td>&lt;command &quot;moreoptions&quot;&gt;</td>
</tr>
<tr>
<td>medical benefits</td>
<td>&lt;ambig_key &quot;healthplan&quot;&gt;</td>
</tr>
</tbody>
</table>

- Can be done with rules, or by a pattern classifier
Search by name

Place: Austin, Boston

Business name: Starbucks

GPS: Boston

Update current state

Decide what to say or do
How dialog systems are designed today

Typical commercial spoken dialog system contains ~100 pages of flowchart
if (action.type == 'disambig'):
    out.text = 'Sorry, was that %s or %s?' % (action.place1, action.place2)

elif (action.type == 'confirm'):
    out.text = '%s, is that right?' % (action.item)
    out.wavs = [GetWav(action.item), is_that_right.wav]

elif (action.type == 'display'):
    out.text = '%s in %s' % (action.business, action.bizname)
    out.display = DisplayMap(action.lat, action.lon)

...
'Barack Obama was elected...'

Text-to-speech

Unit database

Pre-processing

Unit selection

Post-processing

+ + ... =
Putting it all together: example

- Uni-modal telephone-based spoken dialog system
  - SLM speech recognition
  - Language understanding (~250 categories)
  - Speaker verification (~300K users)
An example multi-modal system
Two central challenges for deploying speech technology

AVIxD / IxD workshop – August 2011
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR/SLU errors are common

---

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
<tr>
<td>% in-grammar/in-domain</td>
<td>92.3%</td>
<td>91.0%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>92.1%</td>
<td>77.6%</td>
<td>77.7%</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
ASR/SLU errors are common

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Yes/no</th>
<th>City &amp; state</th>
<th>How may I help you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-grammar/in-domain accuracy</td>
<td>99.8%</td>
<td>85.1%</td>
<td>89.5%</td>
</tr>
<tr>
<td>% in-grammar/in-domain</td>
<td>92.3%</td>
<td>91.0%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>92.1%</td>
<td>77.6%</td>
<td>77.7%</td>
</tr>
<tr>
<td>Accepted utts (False accepts)</td>
<td>89.6%</td>
<td>60.3%</td>
<td>73.3%</td>
</tr>
<tr>
<td></td>
<td>(1.8%)</td>
<td>(4.9%)</td>
<td>(8.3%)</td>
</tr>
</tbody>
</table>

Source: Two different deployed commercial applications running two different speech recognizers
The "theory of mind" problem

Users must think simultaneously about what language the system can understand, and what the system can do – they must form a "theory of mind" about the dialog system.
Responses to "How may I help you?"

• Silences and hesitations while users think
  - Leads to end-pointing problems
  - Leads to users confusing themselves
• "Robot" language (hence examples, "speak naturally")
  - Example 1
  - Example 2
• Recognition errors confused with competences
  - "i need to sign up for a get off benefit" [no parse]
  - "i would like to enroll in a get one" [no parse]
  - "i would like to get help with my dental insurance" <HELP>
  - "dental insurance" <INSURANCE>

Source: Live calls, human resources dialog system, AT&T
Getting started with speech technology

AVIxD / IxD workshop – August 2011
How do I add ASR to my mobile application?

- ASR on mobile devices is usually a bad idea
  - ASR will kill your battery
  - Mobile processors are probably too small
  - Can't benefit from cross-user acoustic data
- Probably better to run ASR in the cloud
  - Stream audio to server; server sends back ASR result
- Example: AT&T Speech Mash-up
Language models: constrain/weigh word sequences

Usually, each "dialog state" has its own language model (LM)

Rule-based LMs

```
GET_STATE [ (?i live in) [
  alaska~0.001
  california~0.2
  arkansas~0.01
  ...
]]
```

Statistical LMs

```
i want 0.0035
want to 0.0023
want benefits 0.0034
want need 0.000002
```
What to include in the language model?

What the user says

**Too narrow:** some caller speech can't be recognized

**Just right:** only things caller actually says can be recognized

**Too broad:** things caller will never say could be recognized, increasing misrecognitions
Setting parameters: beam width

![Graph showing accuracy vs. speed with beam widths 50, 100, 150, and 200](image)

- **beam=50**
- **beam=100**
- **beam=150**
- **beam=200**
Another parameter: language model scale factor

**LOW**

Too low: good acoustic matches but words make no sense; low accuracy

**HIGH**

Too high: word sequences make sense but too much acoustic information is ignored; low accuracy

Just right: a little acoustic information is ignored in favor of words that make sense; best accuracy
Example  "mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
</table>


Example  "mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storr sade a mon in ck a california</td>
</tr>
</tbody>
</table>
Example  "mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storre a mon in ck a california</td>
</tr>
<tr>
<td>0.01</td>
<td>maxtor sadd a monnin ke california</td>
</tr>
</tbody>
</table>
Example  "mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storr sade a mon in ck a california</td>
</tr>
<tr>
<td>0.01</td>
<td>maxtor sadd a monnin ke california</td>
</tr>
<tr>
<td>1</td>
<td>maxtor sad a monica california</td>
</tr>
<tr>
<td>2</td>
<td>maxtor sata monica california</td>
</tr>
</tbody>
</table>
Example

"mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storrsade a mon in ck a california</td>
</tr>
<tr>
<td>0.01</td>
<td>maxtor sadd a monnin ke california</td>
</tr>
<tr>
<td>1</td>
<td>maxtor sad a monica california</td>
</tr>
<tr>
<td>2</td>
<td>maxtor sata monica california</td>
</tr>
<tr>
<td>3-19</td>
<td>mac store santa monica california</td>
</tr>
</tbody>
</table>
## Example

"mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storr sade a mon in ck a california</td>
</tr>
<tr>
<td>0.01</td>
<td>maxtor sadd a monnin ke california</td>
</tr>
<tr>
<td>1</td>
<td>maxtor sad a monica california</td>
</tr>
<tr>
<td>2</td>
<td>maxtor sata monica california</td>
</tr>
<tr>
<td>3-19</td>
<td>mac store santa monica california</td>
</tr>
<tr>
<td>20</td>
<td>maxtor pharmacies</td>
</tr>
</tbody>
</table>
### Example

"mac store santa monica california"

<table>
<thead>
<tr>
<th>LM Scale</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>mm aq ck storr sade a mon in ck a california</td>
</tr>
<tr>
<td>0.01</td>
<td>maxtor sadd a monnin ke california</td>
</tr>
<tr>
<td>1</td>
<td>maxtor sad a monica california</td>
</tr>
<tr>
<td>2</td>
<td>maxtor sata monica california</td>
</tr>
<tr>
<td>3-19</td>
<td>mac store santa monica california</td>
</tr>
<tr>
<td>20</td>
<td>maxtor pharmacies</td>
</tr>
<tr>
<td>25+</td>
<td>restaurants</td>
</tr>
</tbody>
</table>
Summary

Brief tour of the toolbox
Challenges for building ASR systems
A few pointers for getting started
But what about design?! 

That's next!
Speech technologies for interactive mobile applications – a primer

Thanks!

Jason D. Williams

AVIxD / IxD workshop – August 2011

© 2011 AT&T Intellectual Property. All rights reserved.