Adapting Speech Technology to Fit Your Application

Jeff Adams
CEO
Cobalt Speech & Language
The Fallacy of the Average

• The average American woman is:
  – 37 years old
  – Got married at age 27 to a man 2 years older
  – Will have 1.9 children
  – Lives in California
  – Is of German ancestry

• And yet, there are few or no women meeting this description!
There is no average Speech Application

• Let’s imagine you determine to develop a new mobile speech application.

• The very things that make your application interesting and compelling will distance it from the average.

• As a result, very few applications will work well with an “out of the box” speech solution.
  – Customization: Things you can do up-front
  – Adaptation: Things you can do on the fly
“One Size Fits All” = “One Size Fits None”
Easy vs. Hard

• Speech is not a solved problem.
  – All applications will have some errors.
  – Key: design the system so some aspects are easy.

<table>
<thead>
<tr>
<th>easier</th>
<th>harder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small vocabulary</td>
<td>Unlimited vocabulary</td>
</tr>
<tr>
<td>Quiet environment</td>
<td>Noisy environment</td>
</tr>
<tr>
<td>Near microphone</td>
<td>Distant microphone</td>
</tr>
<tr>
<td>Deliberate, careful speech</td>
<td>Spontaneous speech</td>
</tr>
<tr>
<td>Fixed speaker</td>
<td>Multiple speakers</td>
</tr>
</tbody>
</table>
A Few Interesting Examples

Siliconic Home

AGVOICE™

fluent WORLDS

VocalZoom

Canary
• SmartCreature ("Smarty")
• Targeted to kids, it can...
  – Read stories
  – Play music
  – Call parents’ phones
  – Answer questions
  – Etc.
What is special about Siliconic Home? (as a speech app)

• Children speak differently from adults.
• We can’t always assume connection to a network.
• Need to recognize a variety of names of songs, stories, etc.
• The goal is not just speech recognition, but extracting useful, actionable information
• User may be close or “across the room”

(Note: Siliconic Home is here. Go talk to them!)
AgVoice service concept

Capture
- Field Observations
- Work orders
- Activity stream
- Phenotype data
- Photos

Share
- Automated Reports
- Compliance
- Documentation
- Food source tracking
- Workflow Records

Know
- Searchable history
- Query of Ag sources
- Predictive analytics
Before:

- Too many things to carry
- Both hands needed
- Papers get lost
- Manual data entry error prone

Hands-free, eyes free
Digital data uploaded
Timestamp and geo-location
Real-time error correction
What is special about AgVoice? (as a speech app)

• Users will use specialized terminology & phrasing.
• We can’t always assume connection to a network.
• May be windy / noisy.
• The goal is not just speech recognition, but extracting useful, actionable information

(Note: AgVoice is here, & presenting tomorrow morning. Go talk to them!)
The VocalZoom Challenge

How do we combine the best of both signals?

Audio Signal
Full spectrum, but noisy

Optical Signal
Noise-free, but limited to low frequencies

ASR

“Turn off the radio”
The VocalZoom opportunity
The VocalZoom Solution

• Cobalt & VocalZoom have partnered to develop a solution that does speech recognition on both signals simultaneously.
• Innovative approach: “Early DNN Fusion”
  – Cuts errors in noisy environments by about 2/3(!)

(Note: VocalZoom is here. Go talk to them!)
¡Vamos a viajar a lugares divertidos para aprender inglés!

Seleccionar una Aventura
Selecciona la respuesta correcta.

- Hello Lisa.
- Where are you?
- Goodbye.
- Ok.
What is special about FluentWorlds? (as a speech app)

• By definition, speakers are speaking a language they don’t know well.

• There is a difference between “What did they say?” and “Did they say X, and did they say it correctly?”
Disease Screening by Voice
Canary aims to detect disease

• Hypothesis: Some diseases are detectable by voice
  – Alzheimer’s
  – Parkinson’s
  – Depression
  – Brain Tumors
  – Some Cancers

• A simple screening app can encourage people to see a doctor.
What is special about Canary? (as a speech app)

• The main problem is speech classification (not recognition).
• Challenge to get data to train and test models.
• Extra pressure to “get it right”.

(Note: Canary is here. Talk to them!)
The Speech Recognition Chain

Audio

Acoustic Model (AM)

Phonemes

Lexicon

Words

Language Model (LM)

Phrases

Natural Language Understanding (NLU)

Recognizer.start()

Action
## Common Areas of Customization & Adaptation

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Phonemes</th>
<th>Words</th>
<th>Phrases</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acoustic Model</strong></td>
<td>Noise, Reverberance, Accents</td>
<td>Novel words, Names</td>
<td>Specialized text, Personalization</td>
<td>Custom apps (no real “generic” NLU)</td>
<td></td>
</tr>
<tr>
<td><strong>Lexicon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Language Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NLU</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Typical Problems
- Noise
- Reverberance
- Accents
- Novel words
- Names
- Specialized text, Personalization
- Custom apps (no real “generic” NLU)

### Customization
- Train new AM from hours of labeled speech
- Add new words (& prons) to the lexicon
- Train LM from lots of text (or describe by a grammar)
- Design specialized NLU system

### Adaptation
- Learn speech, noise patterns “on the fly”
- Learn new words & prons automatically
- Learn language patterns “on the fly”
- Learn on the fly what the user means
### Customization in our Examples

<table>
<thead>
<tr>
<th></th>
<th>AUDIO</th>
<th>PHONEMES</th>
<th>WORDS</th>
<th>PHRASES</th>
<th>ACTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acoustic Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lexicon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Language Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NLU</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Siliconic Home</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train for kids’ speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Far-field AMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Add words from stories,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>songs, etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train to reflect how kids’ say things</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop NLU for Smarty apps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AgVoice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train for outdoor speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Add specialized Ag terminology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train to reflect typical reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information extraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VocalZoom</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop novel AM (&amp; engine)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FluentWorlds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train for non-native speakers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus on words to be learned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus on phrases to be learned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Canary Speech</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elderly speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognize anomalies, i.e. repetitions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

• Your application is not average. It’s special.
• Consult your ASR provider about possibilities for customization & adaptation.
• Find a consultant that has experience in your area of special-ness.

Feel free to contact me: jeff@cobaltspeech.com