From AUDREY to Siri.
Is speech recognition a solved problem?

Roberto Pieraccini
Director, ICSI
the International Computer Science Institute at Berkeley

roberto@ICSI.berkeley.edu
http://www.icsi.berkeley.edu/
Follow us on Twitter @ICSIatBerkeley
What is ICSI?

- The International Computer Science Institute
- Started in 1988, located in downtown Berkeley
- An independent research organization affiliated with (but not part of) the University of California at Berkeley
- 80-100 people, including staff, principal investigators, postdoctoral fellows, researchers, international visitors, and students
- Pursuing advanced research in many areas of computer science
  - Networking, Security, Speech, Vision, Artificial Intelligence, Algorithms, Computational Biology, Computer architectures
- Funded through federal grants, industry contracts, and collaborations with foreign countries
1952 AUDREY

• First known and documented speech recognizer
• Built in 1952 by Davis, Biddulph, and Balashek at Bell Laboratories
• Fully analogic
• Recognized strings of digits with pauses in the between
• 97-99% accuracy if “adapted” to speaker
...why was it not exploited?

Given these early successes, why were they not exploited? They were not economically attractive. [...] AUDREY occupied a six-foot high relay rack, was expensive, consumed substantial power and exhibited the myriad maintenance problems associated with complex vacuum-tube circuitry. More important, its reliable operation was limited to accurate recognition of digits spoken by designated talkers. It could therefore be used for voice dialing by, say, toll operators, or especially affluent telephone customers, but this accomplishment was poorly competitive with manual dialing of numbers. In most cases, digit recognition is faster and cheaper by push-button dialing, rather than by speaking the successive digits.

What happened after AUDREY?

- Early 1960s – exploration, hybrid systems, phonetic segmentation
- Late 1960s – brute force approach, templates: IT WORKS! Hard to scale...
- Early 1970s – first big ARPA project, Speech Understanding Research (SUR). The AI hype... not a great success, except template based brute force (HARPY)
- Late 1970s – first appearance of Hidden Markov Models (HMMs): IBM (Jelinek), Baker (Dragon)
- Early 1980s – More templates, HMMs are still a secret cult
- Late 1980s – New DARPA projects, HMMs become popular (Rabiner @Bell Labs)
- Early 1990s – More DARPA projects, better HMMs. AT&T’s first large scale deployment (VRCP), the birth of VUI art (Wildfire)
- Mid 1990s – Better HMMs. The industry starts (Nuance, SpeechWorks)
- Late 1990s – Better HMMs. IVRs
- Early 2000s – Better HMMs. IVRs
- Early 2010s – Better HMMs. Mobile voice

IMPROVEMENTS MOSTLY DUE TO MOORE’s LAW


2011 Siri

• Practically infinite vocabulary
• Contextual language understanding
  – ANSWERS ... NOT LINKS
• Voice access to calendar and contacts, help make reservations, gives answer on lots of things, including the meaning of life
• Integrated within iPhone, freely available to everyone (who buys an iPhone)
...why is Siri successful?

• Perception of intelligence
• Fun to use it, witty, catchy personality
• iPhone design and Apple marketing
• Works relatively well for a certain number of tasks
• Improves with time

So ... is speech recognition a solved problems?
...is speech recognition a solved problem?

• NO...and language understanding is even less solved.
  – Fails where humans don’t
  – Little basic science
  – More data, more improvements ... but the rate of improvement is diminishing
  – Looks like we are hitting the intrinsic limitations of the underlying models
  – Each new task requires almost the same new level of effort
READ SPEECH

1000 words

20,000 words

WALL STREET JOURNAL

RANGE OF HUMAN TRANSCRIPTION ERROR

SPONTANEOUS INTERACTIVE SPEECH

ATIS

SWITCHBOARD

SWITCHBOARD II

SWITCHBOARD CELLULAR

RANGE OF HUMAN TRANSCRIPTION ERROR

CONVERSATIONAL SPEECH

NEWS ENGLISH UNLIMITED

NEWS ENGLISH 10X

NEWS ENGLISH 1X

RANGE OF HUMAN TRANSCRIPTION ERROR

BROADCAST SPEECH

RANGE OF HUMAN TRANSCRIPTION ERROR

MEETING SPEECH

RANGE OF HUMAN TRANSCRIPTION ERROR
The evolution of speech recognition

1992
- Feature extraction: frame-based measures
  - Mel frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction (PLP)
  - Delta cepstra (and delta delta, etc)
- Acoustic modeling: Hidden Markov Models (HMMs)
  - representing context-dependent phoneme-like units
- Language modeling: Statistical language models
  - representing context-dependent words

2012
- Feature extraction: frame-based measures
  - Mel frequency cepstral coefficients (MFCC)
  - Perceptual linear prediction (PLP)
  - Delta cepstra (and delta delta, etc)
- Acoustic modeling: Hidden Markov Models (HMMs)
  - representing context-dependent phoneme-like units
- Language modeling: Statistical language models
  - representing context-dependent words

MORE DATA, FASTER CPUs
Normalization, Adaptation, Combination of different systems, …
The evolution of language understanding

1992
- Data-driven statistical models of semantic attributes
  - Concepts
  - Semantic classification
- Handcrafted grammar based semantic parsing
  - Context-free grammar tagging
  - Robust parsing

2012
- Data-driven statistical models of semantic attributes
  - Concepts
  - Semantic classification
- Handcrafted grammar based semantic parsing
  - Context-free grammar tagging
  - Robust parsing

MORE DATA, FASTER CPUs, Standards (SRGS), Tools, …
Where do we go from here?

• Data is not a problem today, models are
  
• Better features and models
  – Models of hearing/production -> better features
  – Models of these features -> better acoustic models
  – Models of understanding -> better language models, dialog models, pragmatics, etc.

• Understanding the errors
  – Examine statistical assumptions
  – Experiments to determine relative importance
  – Look for the cause, rather than for the cure
Wegmann/Gillick: Test model assumptions

Recognize some speech data using an HMM

- output fits HMM distribution
- satisfies independence assumptions

<table>
<thead>
<tr>
<th>Test</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data</td>
<td>13.0</td>
</tr>
</tbody>
</table>
Wegmann/Gillick: Test model assumptions

Simulate pseudo speech data from the HMM

- output fits HMM distribution
- satisfies independence assumptions

<table>
<thead>
<tr>
<th>Test</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data</td>
<td>13.0</td>
</tr>
<tr>
<td>Simulated data</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Wegmann/Gillick: Test model assumptions

Resample real speech frames, respecting the model

- output fits HMM distribution
- satisfies independence assumptions

<table>
<thead>
<tr>
<th>Test</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data</td>
<td>13.0</td>
</tr>
<tr>
<td>Simulated data</td>
<td>0.2</td>
</tr>
<tr>
<td>Resampled data</td>
<td>0.4</td>
</tr>
</tbody>
</table>
OUCH: Outing Unfortunate Characteristics of HMMs

• An ICSI project sponsored by AFRL (Air Force Research Lab) and IARPA (Intelligent Advanced Research Projects Activity)


• In-depth study of acoustic modeling and effects of assumptions in current statistical models  
  – Resampling, mismatch, advanced frond-ends

• Broad survey of field  
  – Literature, expert survey
Gartner’s 2009 Hype Cycle
Gartner’s 2011 Hype Cycle
Conclusions

• Speech recognition has a long history (60 years) of research, failures, and successes

• It feels like we are at a tipping point for the technology

• But the most general speech recognition problem is far from solved

• We do not want to see user expectations outgrow the actual capabilities.

• Continuing on the slope of enlightenment ... or back to the trough of disillusionment?
Stanley Kubrick’s 1968 film 2001: A Space Odyssey famously featured HAL, a computer with the ability to hold lengthy conversations with his fellow space travelers. More than forty years later, we have advanced computer technology that Kubrick never imagined, but we do not have computers that talk and understand speech as HAL did. Is it a failure of our technology that we have not gotten much further than an automated voice that tells us to “say or press 1”? Or is there something fundamental in human language and speech that we do not yet understand deeply enough to be able to replicate in a computer? In The Voice in the Machine, Roberto Pieraccini examines six decades of work in science and technology to develop computers that can interact with humans using speech and the industry that has arisen around the quest for these technologies. He shows that although the computers today that understand speech may not have HAL’s capacity for conversation, they have capabilities that make them usable in many applications and are on a fast track of improvement and innovation.

Pieraccini describes the evolution of speech recognition and speech understanding processes from waveform methods to artificial intelligence approaches to statistical learning and modeling of human speech based on a rigorous mathematical model—specifically, hidden Markov models (HMM). He details the development of dialog systems, the ability to produce speech, and the process of bringing talking machines to the market. Finally, he asks a question that only the future can answer: will we end up with HAL-like computers or something completely unexpected?