Challenges in Automatic Speech Recognition

2010-2020: Speech Technology for the Next Decade - Visions from Academia and Industry

Ciprian Chelba, Michiel Bacchiani, Johan Schalkwyk

`{ciprianchelba,michiel,johans}@google.com`

Google

Case Study:Google Search by Voice

Carries 25% of USA Google mobile search queries!
What contributed to success:

- clearly set user expectation by existing text app
- excellent language model built from query stream
- clean speech:
    - users are motivated to articulate clearly
    - phones do high quality speech capture
    - speech tranferred error free to server over IP

Challenges:

- Making and measuring progress: manually transcribing data is at about same word error rate as system (15%)

Case Study: Google Labs GAudi Demo

This was the study for the YouTube feature that is now launched for all and integrated with translation.
Main challenge:

- lack of coverage due to ASR limitations:
  - noise-robustness
  - speaker/accent/channel variability
  - language model mismatches
  - web is multi-lingual

ASR for Retrieval and Ranking

On large document collections search is truly about Precision@N.
There is seldom a good reason to replace a result in the top-N with one that has hits in the (noisy) ASR transcript.
Future directions:

- improve retrieval for "hard queries" which return very few documents based strictly on keyword hits in the text metadata
- speech-rich sub-domains such as lectures/talks in English recorded in a controlled setup where current ASR capabilities are adequate after manual tuning to the sub-domain.

Core Technology

Current state:

- automatic speech recognition is incredibly complex
- problem is fundamentally unsolved
- data availability and computing have changed significantly since the mid-nineties

Challenges and Directions:

- re-visit (simplify!) modeling choices made on corpora of modest size; 2-3 orders of magnitude more data is available
- multi-linguality built-in from start
- noise-robustness and speaker/channel variability