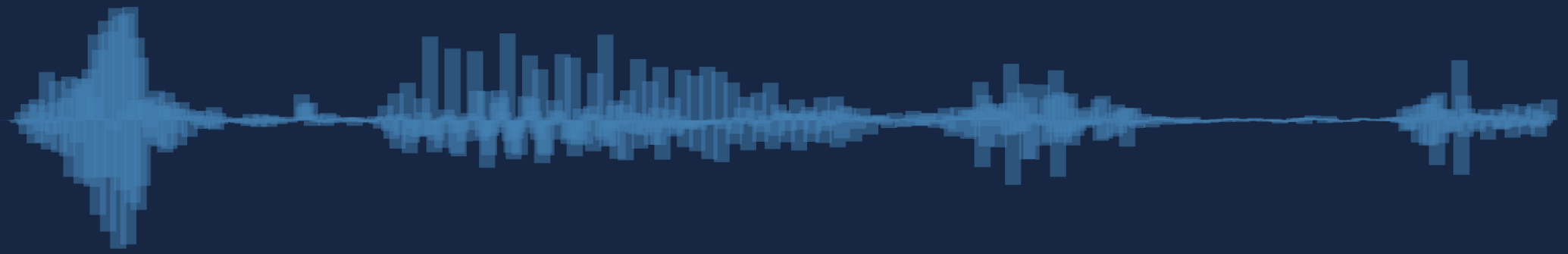
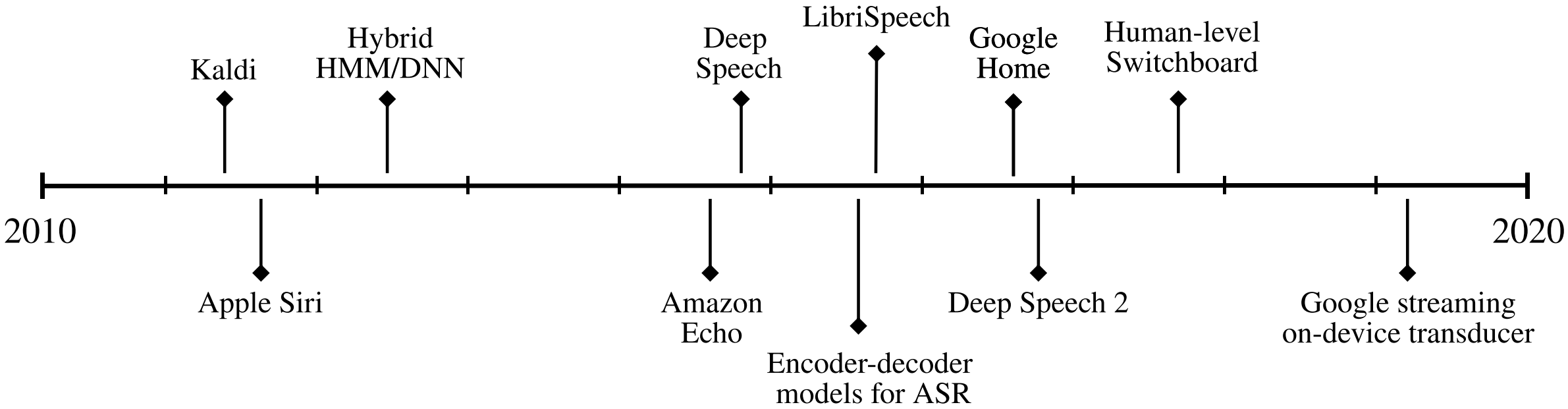
The History of Speech Recognition to the Year 2030

Awni Hannun, awni.hannun@gmail.com



Recap


Kaldi
Hybrid HMM/DNN
Deep Speech
LibriSpeech
Google Home
Human-level Switchboard
2010
2020
Apple Siri
Amazon Echo
Encoder-decoder models for ASR
Deep Speech 2
Google streaming on-device transducer

Predicting the Future

Richard Hamming in *The Art of Doing Science and Engineering* predicted:

- General-purpose rather than special-purpose hardware
- Digital over analog
- Importance of high-level programming languages
- Fiber-optic in place of copper

Predicting the Future

Richard Hamming in *The Art of Doing Science and Engineering* predicted:

- Neural networks “represent a solution to the programming problem”, and that “they will probably play a large part in the future of computers.”

Predicting the Future

Richard Hamming in *The Art of Doing Science and Engineering* predicted*:*

- By “the year 2020 it would be fairly universal practice for the expert in the field of application to do the actual program preparation rather than have experts in computers (and ignorant of the field of application) do the program preparation.”

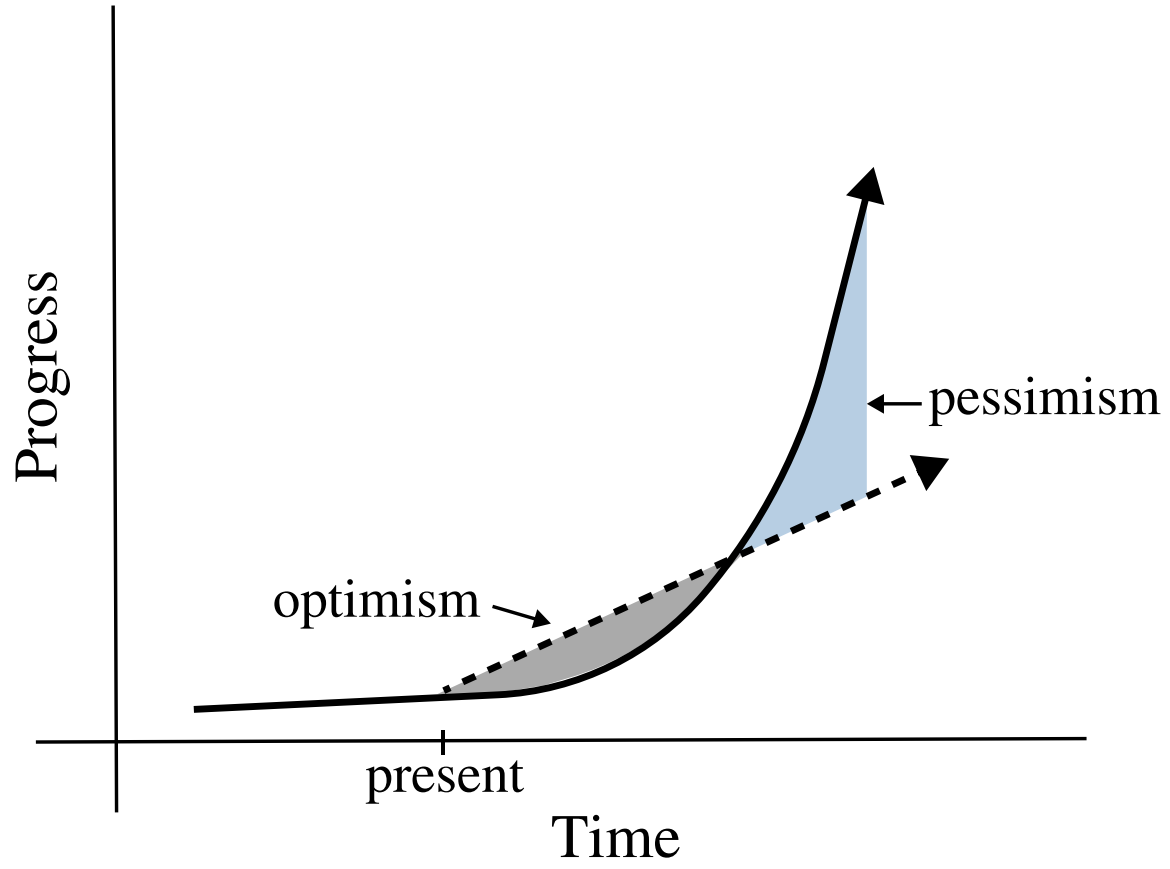
Predicting the Future

Richard Hamming in *The Art of Doing Science and Engineering:*

- **Practice:** Reserved Friday afternoon for “great thoughts”
- **Fundamentals:** Understand current trends and rapidly assimilate new knowledge
- **Open mind:** An “open mind leads to the open door, and the open door tends to lead to the open mind”

Predicting the Future

Short-term overly optimistic, long-term overly pessimistic



Self-supervised learning

Prediction: Self (and semi) supervised learning will be a big part of speech recognition applications.

Why:

- Works and is already widespread for text applications
- Reduces the need for labelled data

Self-supervised learning

Challenges:

- Scale

Research implications:

- Sparsity for lighter-weight models
- Optimization for faster training
- Prior knowledge for sample efficiency

On-device Inference and Training

Prediction: Most speech recognition will happen on device or at the edge.

Why:

- Privacy
- Latency
- 100% availability

On-device Inference and Training

Challenges:

- Low-memory and low-energy footprint models
- Labels for training

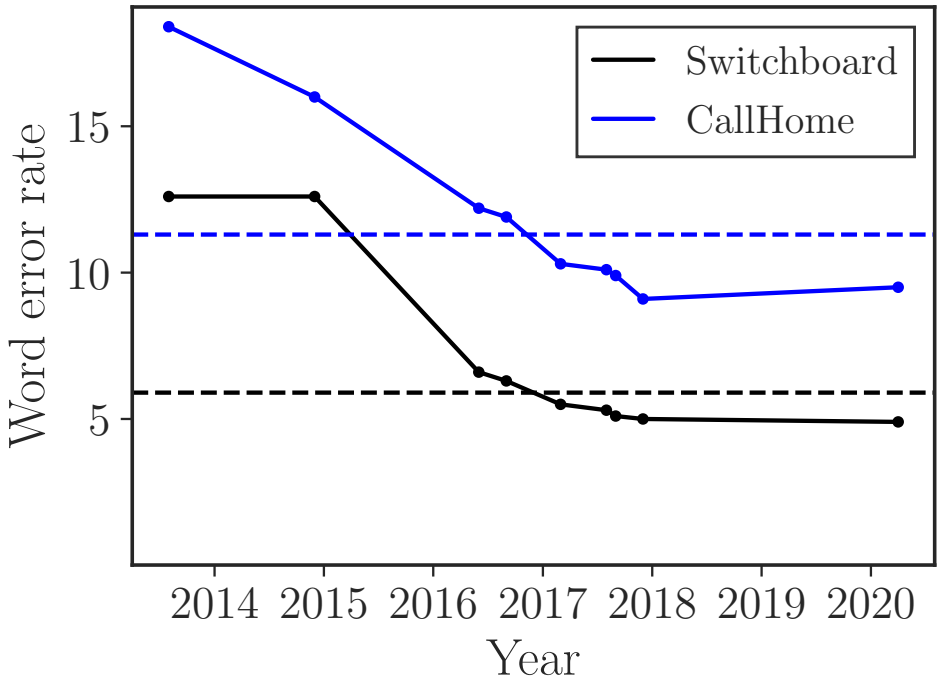
Research implications:

- Sparsity
- Weak supervision

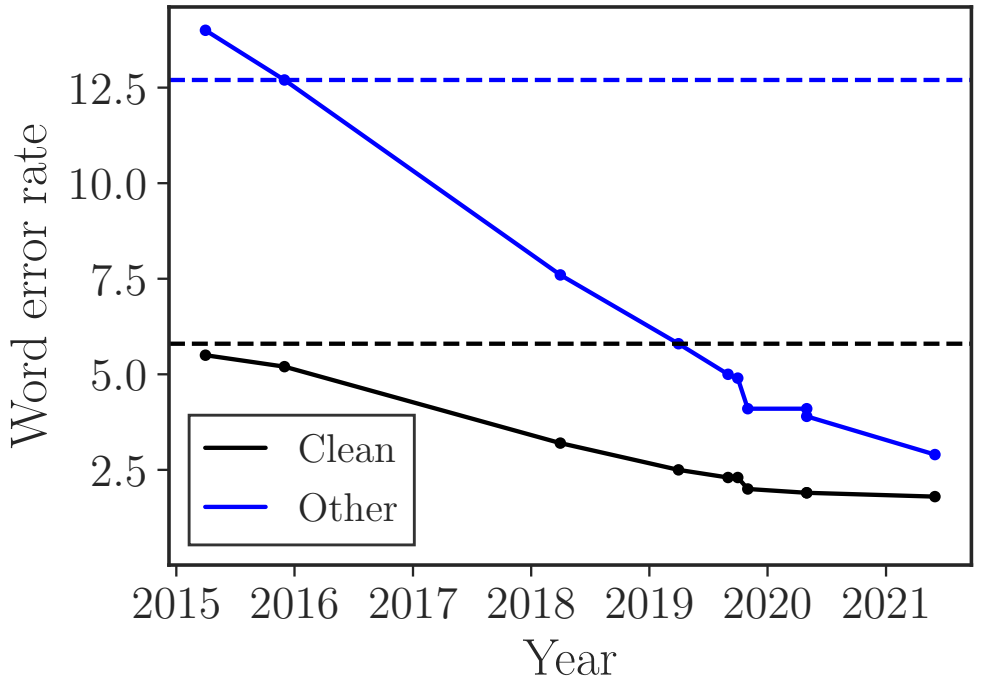
Word Error Rates

Prediction: No more “improved word error rate on benchmark X with model architecture Y” publications.

Switchboard



LibriSpeech



Richer Representations

Prediction: For downstream tasks, transcriptions will be replaced by richer representations.

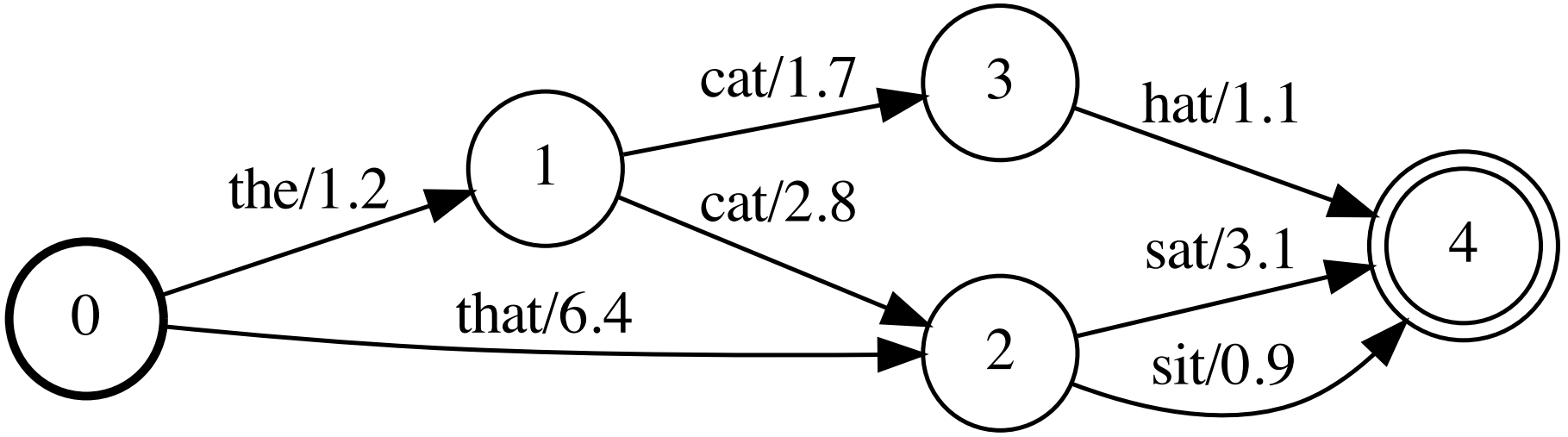
Why:

- Verbatim transcript often not needed
- Semantic correctness is more important
- Uncertainty is useful

Richer Representations

Research implications:

- Efficient representation of hypotheses with uncertainty
- Models which can take as input a lattice of hypotheses



Personalization

Prediction: Speech recognizers will be personalized to individual users.

Why: Context is essential

- User's contacts
- Topic of conversation
- Conversation history
- Accent and other speech idiosyncrasies

Personalization

Challenges: On-device training

Research implications:

- Light-weight models (sparsity)
- Weak supervision
- Models which optimally incorporate context

Application Predictions

Transcription services: 99% automated

Voice Assistants: Only incremental improvements

- Speech recognition is no longer the bottleneck
- Open domain question answering
- Conversational AI

Summary

Prediction
Self-supervised learning and pretrained models are here to stay.
Most speech recognition (inference) will happen at the edge.
On-device model training will be much more common.
Sparsity will be a key research direction to enable on-device inference and training.
Improving word error rate on common benchmarks will fizzle out as a research goal.
Speech recognizers will output richer representations (graphs) for use by downstream tasks.
Personalized models will be commonplace.
Most transcription services will be automated.
Voice assistants will continue to improve, but incrementally.

Thanks

The best way to predict the future is to invent it.

- Alan Kay