Evaluation of Speech Separation, Corpus Development:

The Speech Recognition Experience

Alex Acero

Microsoft Research

Speech Separation and Comprehension in Complex Acoustic Environments by Humans and Machines

Why Evaluation?

- So that we can track progress
- We need objective measurements
- Progress in ASR (2004-2009)?
  - Error rates in Switchboard goes down by 50% ☺
- Progress in Source Separation (2004-2009)?
  - 12 barn owls lost sense of direction ☹
  - Measured transfer function of 35 bathrooms ☹
  - Published 890 papers
  - Hearing aids: 80% of users prefer 2009 aids ☺
  - ASR error rates go down by 30% in cocktail parties ☺

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

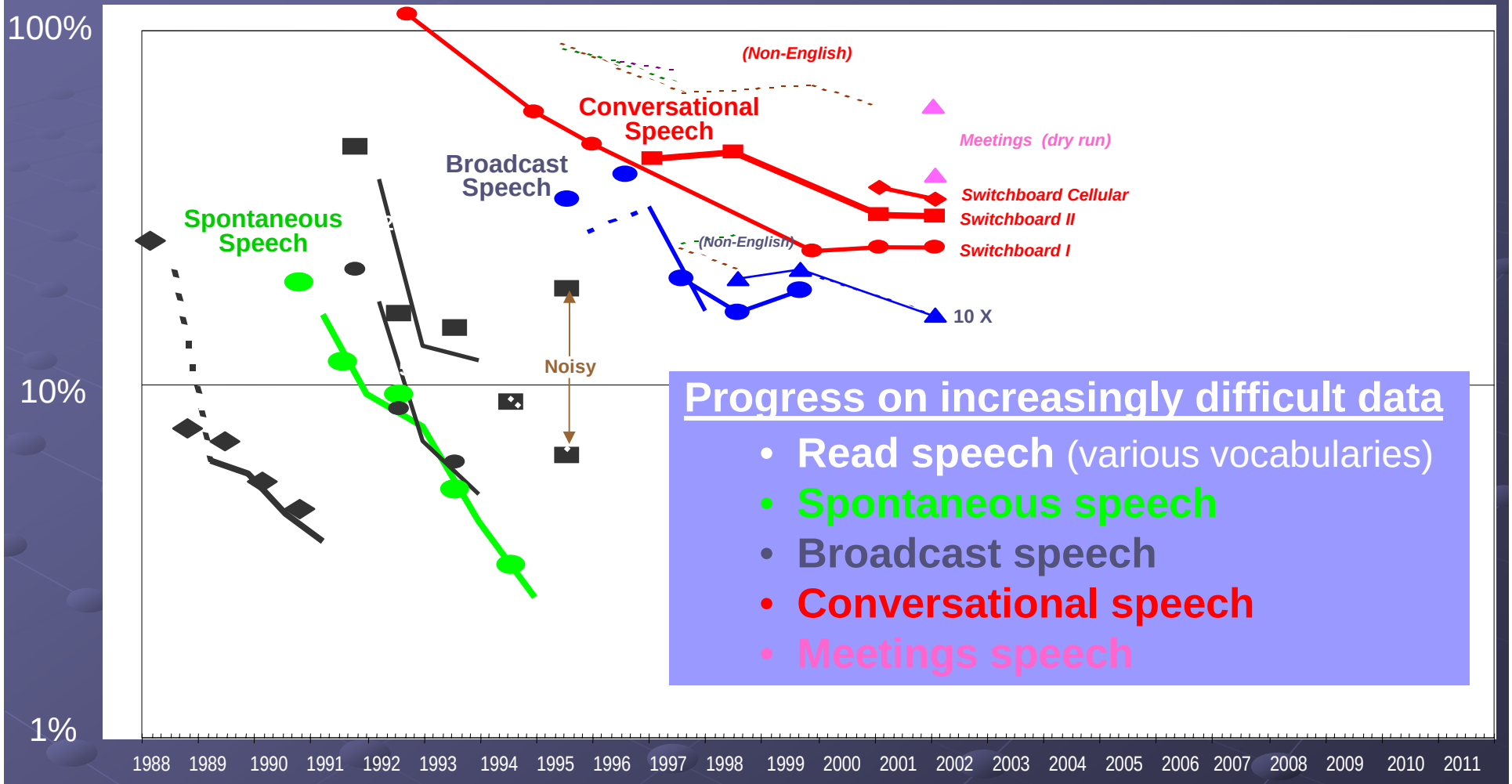
The DARPA ASR Program

- From 1970 till today
- Program goals: lower error rates
- Common tasks:
  - Training set, dev set, test set, vocabulary
- Only techniques that improve accuracy are used
- Data-driven:
  - “There is no data like more data”
- Annual workshops:
  - Sharing algorithmic advances
- Requires large teams:
  - ASR systems are complex

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

ASR Historical Progress

Word Error Rates for Speaker-Independent Speech-to-Text



Progress on increasingly difficult data

- **Read speech** (various vocabularies)
- **Spontaneous speech**
- **Broadcast speech**
- **Conversational speech**
- **Meetings speech**

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

The DARPA ASR Program

- Effective ☺:
  - Error rates halve every 5-7 years
- Little diversity ☹:
  - All systems are similar
- EARS Program (2002-2006):
  - Traditional evaluation
  - Novel approaches
- Mostly clean speech

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

Noise robustness in ASR: Aurora

- Aurora Goals:
  - Compare noise robust front-ends for ASR
  - Fast experiment turnaround => digit recognition
  - Simple => ASR system as black box (HTK based)
- Aurora2: Noise is added digitally
- Aurora3: Speech recorded in a noisy car
- Aurora4: WSJ speech with additive noise
- Over 20 papers per Eurospeech/ICSLP
- Great progress in technology
- Small labs can play!

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

NIST Meeting Transcription Task

- Meetings recorded at ICSI, CMU and NIST
- From 3 to 8 participants
- Several microphones:
  - Reference: close-talking
  - Lapel microphone per person (CMU)
  - Far field microphones on table (ICSI, NIST)
- Over 100 hours transcribed
- Evaluation in 2003 and 2004
- Best system in 2004 had 45% error rate ☹
- No funding => few participants

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

Human simulating a machine?

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$

Summary

- Evaluation is key to progress
- Need to define metrics
- Build systems that work mimicking the human auditory system or not

Thank you

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}$$