Robust Speech Recognition in the Real Word

Panel Discussion Presentation at ICASSP 2004, Montreal Canada

Ramesh Gopinath
Manager, Pervasive Speech Technologies
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
May 17, 2004
rameshgl@us.ibm.com
The Main Problem

• The Problem
  – Designing, developing, deploying and maintaining robust speech applications expensive
    • ASR technologies are fragile
    • Skills required are expensive

• Key Ingredients of a Solution to this problem
  – Excellent, robust, out-of-the-box performance of ASR technologies
  – Good Programming Model
    • VoiceXML, SALT, etc.
    • Reduce the skill-level to author speech applications
  – Good Tool-set
    • Technology tools, Application Development tools, Deployment tools, Reporting and Maintenance tools
  – Open standards
    • Future-proof investment, increase choices
  – Managing customer and end-user expectations
    • Can’t have guaranteed levels of performance for all users
ASR Technology – Fragility Everywhere!

• When to listen?
  – Voice-activity detection
  – Echo-cancellation and barge-in

• What to listen to and recognize?
  – Quality of ASR (1-best \textit{and} N-best/lattices)
    • Too many variations in the acoustic input (speaker, channel, noise, accent etc.)
    • Lack of data for Grammar/LMs for new domains
  – Reliability of ASR – confidence scores
    • Credibility of confidence scores e.g., after model adaptation

• What to ignore?
  – Noises, non-speech sounds and out-of-grammar utterances

• How to treat users? Be kinder, gentler?
  – End users wonder: “\textit{What can I say?”}
ASR Robustness – There is no Data like More Data!

• Deployments – what have we learned?
  – Dramatic improvements with customization and tuning with real data from deployments
    • Nearly 30-50% improvement
    • Lexicon, Grammars/LM and Acoustic Model adaptation

• DARPA Program – what have we learned?
  – Dramatic progress in every new task in first 1-2 years – data a big part
  – Slow progress after that
What should we do?

• **Increase Data & Computing**
  – 10-fold to 100-fold increase in training data – address coverage issue; requires more computing
  – Unsupervised learning from large volumes of data

• **Invent/Investigate new ASR paradigms**
  – E.g., learn from the robustness of human speech recognition – the *acoustics-first* approach advocated by Jont Allen

• **Move beyond the Word Error Rate (WER) metric**
  – Quality of N-best and lattices
  – Reliability of ASR confidence scores
  – Rejection
  – Signal-level problems (robust VAD, echo-cancellation)