Using Speech with Computers

Alan W Black
Overview

- Practical and Theory:
  - Understand Concepts, Implement Solutions
- Speech Synthesis
  - Text to Speech
- Speech Recognition
  - Speech to Text
- Spoken Dialog Systems
  - Interaction with machines
MWF 3:30-4:20

GHC 4101

Lecturer: Alan W Black (awb@cs.cmu.edu)

TA: Sai Krishna Rallabandi
srallaba@andrew.cmu.edu

tts.speech.cs.cmu.edu/courses/11492/
Course Details

- Three lectures a week
- Various readings
- 4 Homeworks
  - Speech Recognition
    - Train a speech recognition system
  - Speech Synthesis
    - Build a synthesizer from your own voice
  - Spoken Dialog System
    - With competition between your dialog systems
- Other
- Final Exam
Homeworks

- **(Mostly) Practical**
  - **Build something that talks/can be spoken to**
  - **Software and speech data will be provided**
    - Will run on Windows/Linux or OSX
    - Access to Linux servers if required
  - **Written description of what you did**
**Schedule Details**

- **Week 1**
  - Applications, Human and Computer Speech Processing

- **Week 2-4 Speech Recognition**
  - Signal representation, acoustic modeling
  - Language modeling, applications
  - Tuning, evaluation, expectations
Week 5-7 (Oct) Speech Synthesis
- Text processing, prosody, waveform synthesis
- Building voices, evaluations, voice conversion

Week 8 Multilinguality
- Supporting new languages efficiently

Week 9-11 Dialog Systems
- VoiceXML, Mixed initiative, barge-in
- Design, installation and tuning.
Course Details

- **Week 12**
  - Speech to Speech translation
  - Language support, tight integration
- **Week 13**
  - Evaluation and expectations
- **Week 14**
  - Speaker ID, Silent Speech, Conversion
  - What still needs to be done.
- **Week 15 (Dec)**
  - Exam
Why Speech

- **Most natural way to communicate**
  - *(For Humans)*

- **Not ideal for everything**
  - *Graphics and text can be better (sometimes)*

- **Doesn’t compress well**

- **Hard to search**
Compression

- **Alice in Wonderland**
  - **Text**
    - 150K uncompressed
    - 43K compressed
  - **Speech (2hrs 20mins)**
    - 270M uncompressed
    - 600K compressed (mp3, 24KBS)
Searching

- **Find all NPR broadcasts mentioning Tesla**
  - *Listen to them all*

- **From lecture recordings**
  - *Find all occurrences of “this will be in the exam”*

- **So listen to it faster …**
  - *Normal*
  - *2x  4x  8x*
Eyes/Hands Free

- **Interaction when driving**
  - Look at screen to see next turnoff
  - “In 200 yards turn right onto Murray Ave.”

- **Blind users/ Assistive technology**
  - Text isn’t very useful

- **Alerts**
  - “Self-destruct in 10 seconds” vs blinking light

- **Telephone dialog systems**
Speech Applications

- **Command and Control**
- **Information Agents**
  - IVR Telephone Dialog Systems
  - Agents: Siri, avatars, question/answering
- **Speech to Speech Translation**
- **Speech summarization**
  - Lecture or Meeting summarization
- **Transcription/Dictation**
- **Speaker Identification**
  - emotion/dialect/language
- **Language Learning**
“Hot” Commercial Applications

- **Location-based services:**
  - Yahoo
  - Google Maps
  - Microsoft Live Search
  - Android/iPhone

- **Spoken Assistants**
  - Apple’s Siri, Google NOW, Microsoft Cortana, Echo

- **All phone/pda based**
  - Use speech-in
  - Directions speech-out
Other Speech uses

- **Spoken Dialog Systems**
  - Amazon Alexa, Apple Siri, Microsoft Cortana, Google Home

- **Assistive Technologies**
  - Screen readers
  - Augmentitive and assistive communication devices

- **On-line Personalization**
  - Blogcasts (your voice, or appropriate voice)
  - Game character customization

- **Talking Heads**
  - CMU’s roboceptionist

- **Singing Synthesis**
  - XML interface for song specification