Speech Processing 11-492/18-492

Review

ASR, TTS, Dialog,
S2S, VC, SID and CALL
Speech Overview

- **ASR**
  - Automatic Speech Recognition (AM and LM)

- **TTS**
  - Text to speech: unit selection and statistical parametric synthesis

- **Dialog**
  - Spoken dialog systems: direct and mixed initiative dialogs, non-task oriented
Speech Overview

- **VC**
  - Voice conversion, transformation, morphing

- **SID**
  - Speaker ID, Speaker recognition

- **CALL**
  - Computer Aided Language Learning

- **S2S**
  - Speech to Speech translation
Acoustic models

- Acoustic models (usually HMMs)
- Modeling all ways to say each phoneme

Language models

- Modeling word sequence likelihoods
- Tri-grams and grammars
• ASR and Bayes rule

\[ P(W \mid O) \]

By Bayes rule

\[ \frac{P(W)P(O \mid W)}{P(O)} \]

Acoustic model

\[ P(O \mid W) \]

Language model

\[ P(W) \]
ASR Evaluation

- WER
  - Word error rate vs Accuracy
- What is the expected/acceptable WER of
  - Dictation
  - Dialog systems
  - Speech IR
  - Conversational speech with a far field microphone with multiple overlapping non-native speakers (who know each other) with heavy vehicle traffic in the background and other bystanders chatting (and a brass band playing in the background)
- **Text analysis**
  - Homographs, symbol, expansion

- **Linguistic analysis**
  - Pronunciation lexicons
  - Prosody: breaks, intonation, duration

- **Waveform synthesis**
  - Formant synthesis, concatenative synthesis, statistical parametric synthesis
Waveform Synthesis

- **Diphones**
  - Mid-phone to mid-phone speech units

- **Unit selection**
  - Selecting appropriate sub-word units from large databases of natural speech

- **Statistical Parametric Speech**
  - Build speech model of “averages” of similar speech

- **Limited domain synthesis**
  - Targeted synthesis
TTS Evaluation

- **Yes that sounds like a robot**
- **Human listening tests**
  - MOS scale for “likable”
  - SUS sentences for understandability
  - Human personal preferences.
Spoken Dialog Systems

- **System types:**
  - System initiative
  - Mixed initiative
  - HMIHY (How may I help you)
  - Non-task oriented
Spoken Dialog System Evaluation

- Task completion
- Call length
- Number of turns
- (Number of Calls)
- Break down by
  - New/repeat callers
  - Different usage types
New Languages

- **Text examples**
  - For finding nice prompts
  - For building language models
- **Phoneme definitions**
- **Pronunciation lexicon**
- **Recordings**
  - Lots for ASR, one good one for TTS
Speech to Speech

- Real time
- Targeted/wide vocabulary
- Speech not text
- Often resource limited target language
  - Need a written form, and collect own data
Voice Conversion

- **Convert source text to target speaker**
  - Small amount to target speaker (e.g. 30 utts)

- **GMM-based models**

- **Uses**
  - Speaker conversion
  - Style conversion
  - Cross lingual voice conversion
  - De-identification

- **Evaluation**
  - Listening
  - Speaker ID systems
Speaker ID

- **Speaker recognition**
- **Who is speaking**
  - Security, passwd access
  - Diarization (who is speaking in a meeting)
  - Speaker, language, dialect, style ID
- **Techniques**
  - GMM and Phone based techniques
- **Computer aided language learning**
- **Reading tutors**
  - *First and second language Learners*
- **Second language learners**
  - *Pronunciation trainers*
  - *Fluency practice*
  - *Interactive scenario experience*
Final Notes

- Don’t forget to fill in Faculty Course Evaluation
- Final exam
  - Sun 18th 8:30am-11:30am WEH 5302
  - CLOSED BOOK