Speech Processing 11-492/18-492

Speech Synthesis
Evaluation

Evaluating Speech Synthesis

- *How good is the voice?*
  - *This voice is a 45.67*
- *Is voice X better than voice Y*
- *Why?*

Evaluation

- *Objective measures*
  - *Run a program and get a number*
- *Subjective measures*
  - *Have human listeners extract a score*
- *Do Object and Subjective scores correlate*

Human Tests

- *Synthesis people are warped*
  - *The more you listen the better it becomes*
  - *They hear things others don't*
- *Non-synthesis people are warped*
  - *People very sensitive to listening conditions*
  - *What question do you ask*
  - *What hardware you play it on*
- *There are (at least) two orthogonal scales*
  - *Understandability*
  - *Naturalness*

Standard Tests

- DRT: diagnostic rhyme tests
  - Test confusable phones
  - “bat” vs “pat”
  - Good for identifying phone errors
  - Sometimes in carrier sentences
    - → Now we will say pat again.
  - Unit selection
    - → Just include the standard works in the database

Standard Tests

- SUS: Semantically unpredictable sentences
  - Det adj noun verb det adj noun
  - Automatically filled in with low frequency words
    - → The parklike holders threw the vague vegetables
    - → The simplistic consonants swam the episcopal quartet
    - → The dark geniuses woke the humane emptiness.
    - → The masterly serials withdrew the collaborative brochure
- Test for understandability
  - Ask users to type in what they hear
  - Good as discrimination
  - Very hard for even fluent non-natives

Standard tests

- *MOS: mean opinion scores*
  - *1-5 quality, naturalness, "like it"*
  - *Take average score*

Some experimental problems

- Order of presentation
- Other aids change perception
  - Showing the text makes it much easier
  - Having a talking head “improves” the synthesis
- Hardware quality
  - Some voices better on the telephone
  - Loud speaker quality (headphone quality)
  - Room acoustics
  - Volume
- Understandability
  - Harder if doing other task
- Personal preference
  - Voice is full understandable but “creepy”
  - Voice is incomprehensible but “funny”
  - Sounds like my grade school teacher

TTS Evaluation

- *How good are your ears?*

SUS Sentences

- sus_00005 
- sus_00012 
- sus_00017 
- sus_00022 

SUS Sentences

- *The sorrowful premieres sang the ostentation gymnast*
- *The temperamental gateways forgave the weatherbeaten finalist*
- *The disruptive billboards blew the sugary endorsement*
- *The serene adjustments foresaw the acceptable acquisition*

TTS Evaluation

TTS Evaluation

- *In mud eels are, in mud none are*
- *A 1918 state constitutional amendment made Massachusetts one of 23 states where citizens can enact laws by plebiscite.*
- *Which is which*
  - *The numbers are 25 and 34.*
  - *The numbers 20 5 and 34.*
- *What is the temperature in Pittsburgh*

Objective Synthesis Tests

- Text analysis
  - How well do you cover NSWs
  - How well do you cover homographs
- Lexical coverage
  - How often do you see a new word
- Lexical correctness
  - How correct are pronunciations
  - For unseen words
  - For seen words
- Phonetic intelligibility
  - DRT tests
- Semantic intelligibility
  - SUS tests

Blizzard Challenge

- *Annual Event from 2005 (15 years plus)*
- *Distribute large databases of speech*
- *Participants*
  - *Build a voice*
  - *Synthesize a set of sentences*
- *Listeners*
  - *Listen and grade results*

Blizzard Challenge

- *2005: US English synthesis, 4 voices, 1 hour each*
  - *4 teams plus "Studio" (human speech)*
- *2006: US English: 1 voice: 6 hours and 1 hour*
  - *12 teams*
- *2007: US English: 1 voice: 9 hours and 1 hour*
  - *14 teams*
- *2008: UK English: 15 hours: Mandarin 5 hours*
  - *19 teams*
- *2009: UK English: 15 hours: Mandarin 5 hours*
- *2010: UK English 18 hours: Mandarin 6 hours*
- *2010- Audio Books, Indian Languages, Speaking in Noise*
- *Split between industry and academia*
- *Split between Asia, Europe, America (mostly Europe and Asia).*

Listeners

- *Three sets of listeners*
  - *Speech experts (participants)*
  - *Paid undergrads (native speakers)*
  - *Volunteers*
- *Types of tests*
  - *MOS tests (1-5)*
  - *SUS tests*
  - *DRT tests*
- *About 300 listeners in total*

Listening

- *Web based*
  - *So everyone did it in a different environment*
  - *But we got access to more people*
  - *Asked to do it in quiet office with headphone*
  - *Could listen multiple times*

Blizzard Challenge Results

- *Speech Experts*
  - *Like synthesis better*
  - *Understand synthesis better*
- *Volunteers don't always finish tests*
- *Undergrads sometimes finish tests*
  - *(or put in filler answers)*
- *Results were correlated over different subgroups*

Application Tests

- *How does it work *in* the application*
- *With real application data*
- *A good voice is not noticed*
- *Have *real* users evaluate it*
- *Give them a choice (even if artificial)*
  - *CEO chooses the one they like!*

Clearer Spoken Output

- *In Let's Go Bus Domain*
- *Lexical Choice*
  - *The next bus is at 10:23*
  - *The next bus is in 11 minutes*
- *Prosodic variation*
  - *The next bus is at 10:23*
  - *The next bus is at, 10:23.*
- *Spectral variation*
  - *Clear articulation (when asked to repeat)*
  - *The next bust is at, 10:23.*

Summary

- *TTS Evaluation is hard*
  - *But not impossible*
  - *Clear ways (that are consistent) are available*
    - → *MOS scores*
    - → *SUS*
    - → *Application based testing*