Human Communication by Speech

Communication channels:

- Direct channels
  - verbal communication (written/spoken communication)
  - non-verbal communication (facial expressions, body movement)
- Indirect channels (body language)

Speech as HCI

Speech input

- Automatic speech recognition (ASR)
  - dictation, subtitles, call centres, translation, meeting minutes, computer games, · · ·
- Speaker recognition (identification, verification)
- Disease diagnosis, pronunciation evaluation

Speech output

- Text-to-speech synthesis (TTS)
  - system response, screen reader for the blind, car navigation systems, directory assistance, toys, computer games, · · ·

Speech as HCI (cont. 2)

Speech input-output

- Voice conversion
- Speech to speech (language) translation
- Spoken dialogue system
- Education / therapy
- Noise cancelling / speech enhancement
Speech tech. – the ideal and the real

How good / feasible are those state-of-the-art technologies?

- automatic speech recognition (ASR)
- text-to-speech synthesis (TTS)
- spoken dialogue system
- audio + visual system

Dream ASR systems

- Many examples can be found in films and animations
  - HAL2001, C3PO
- Many national big ASR projects have been organised since 1970s

MIT Oxygen project (image video)

Robots in the StarWars movies

Real ASR systems

- Video clip + Demo
- Speech translation system (ATR)

Real ASR systems (cont. 2)

video clips
Real ASR systems (cont. 1)

There are many reasons that ASR goes wrong in the real world.

ASR systems are not robust against:
- noises (background music/voices, reverberation)
- change of microphones, room acoustics
- change of speakers, speaking styles
- change of the topics spoken
- disfluencies (e.g. fillers)
- unknown words

Text-to-Speech Synthesis

- Concatenative approach (recorded audio based approach, unit-selection approach)
  - e.g. Festival, CerVoicce
- Model based approach
  - Formant synthesis
  - Articulatory synthesis
  - Statistical model based synthesis (HMM synthesis)

Progress of ASR

Text-to-Speech Synthesis

- Concatenative approach (recorded audio based approach, unit-selection approach)
  - e.g. Festival, CerVoicce
- Model based approach
  - Formant synthesis
  - Articulatory synthesis
  - Statistical model based synthesis (HMM synthesis)

HMM-based synthesis is gaining attention, because
- Easy to synthesise arbitrary speaker’s voice with a small amount of training data.
- Easy to control emotions in speech
- Easy to control speaking style

( http://www.nist.gov/speech/history/ )
**Text-to-Speech Synthesis (cont. 2)**

HTS Demo
- Application to many languages
- Multi-dialectal HTS

**Speech applications (cont. 2)**

**Industries**

- SpinVox (www.spinvox.com)
  Voice message $\Rightarrow$ ASR(+human) $\Rightarrow$ text / email (SMS)
- VoxGen (www.voxgen.com)
  Ring’n Sing $^{TM}$:
  ‘Karaoke’ singing contest (game) over the telephone

**Speech applications**

- Meeting summarisation (AMI project)

**Speech + Animation = Talking Head**

- More natural HCI, human-like communication
- Useful for spoken dialogue systems to reflect system inner status to users.

- August [J. Gustafson et al., KTH 1999]
- Hyumu [Dosaka, NTT 2000]
- Baldi [Massaro (UCSC), Sutton (OGI) 1998]
Speech + Animation = Talking Head
(cont. 2)

- Different agents having the same personality?
- Agents should have different personalities from each other

Needs a mechanism to generate an unlimited number of personalities

Giving a personality to an agent?
(cont. 2)

- Rule-based approach
- Handcrafted model based approach
- Statistical approach / Corpus-based approach
  (Machine learning approach)
  Background:
  - ASR: speech/text corpus → HMMs, n-grams
  - TTS: speech/text corpus → concatenative synthesis, HMM-based synthesis
- Facial animation: facial pictures + video database
  Data of a real person

Giving a personality to an agent?

Talking-head Demo
- Nina: a flight booking system
- head motion control
- Androids

The Uncanny Valley Hypothesis
Summary

- Speech technologies are still under development
- Feasibility of speech interface should be taken into account when designing HCI
- ASR researchers are looking for killer applications.

References

- Speech Recognition
- Festival TTS system
  http://www.cstr.ed.ac.uk/projects/festival/
- HMM-based speech synthesis
  http://hts.sp.nitech.ac.jp/
- EMIME project (Effective Multilingual Interaction in Mobile Environments)
  http://www.emime.org/
- AMI project
  http://www.amiproject.org/
- SpinVox
  http://www.spinvox.com/
- Ring’n Sing
  http://www.voxgen.com/products.asp

References (cont. 2)

- Valdi (a talking head)
  http://mambo.ucsc.edu/
- HUMAINE project
  http://emotion-research.net/
- SEMAINE project
  http://www.semaine-project.eu/
- SSPNet project
  http://sspnet.eu/
- Valdi (a talking head)
  http://mambo.ucsc.edu/
- HUMAINE project
  http://emotion-research.net/
- SEMAINE project
  http://www.semaine-project.eu/
- SSPNet project
  http://sspnet.eu/