Role of Speech User Interface

Tomas Macek
It is more than 65 years ago when US Department of defense begun funding the first speech processing project.
- What are the reasons for “slow” progress?
- Is the speech really as big thing in UI as originally expected?
- What are the current trends and techniques?
Why speech

- Speech is fast (large lists, dates, times)
- Speech is natural and intuitive
- Speech input device is small
- Capturing emotional state
- Determining speaker identity
Speech interfaces can be just a burden if not designed properly!
Problems with speech

- Speech is transient (no history on the screen)
- Speech is “serial”.
- Limited short term memory of the user
- Real time apps (speech is slow)
- Problems with noisy environment
- Other modalities can be more effective in some cases
- Privacy
Application areas

- Large list selections, dates and times.
- Hands busy situations
- Embedded systems with no keyboard or screen
- Telephony
- Pervasive systems – Car, Home environments
Speech recognition is not the same as speech understanding!
NLP Technologies

- TTS (Text To Speech)
- ASR (Automatic Speech Recognition)
- NLU (Natural Language Understanding)
- Dialog management

- Speaker ID, speaker verification
- Voice detection and location
- Language detection
ASR

- **Where it is done?**
  - Remote (on server)
  - Local (on a client device)
  - Hybrid (both)

- **Output**
  - Recognized sentence, N-best or lattice
  - Confidence
  - Annotation
ASR – Automatic Speech Recognition

- **Who can speak?**
  Speaker independent
  Speaker dependent
  Speaker adaptation

- **What can I say?**
  List of phrases
  Grammar
  Dictation
  unlimited vocabulary or domain specific
ASR

- **When to speak?**
  PTS – Push To Speak
  PTA – Push To Activate, Silence detection
  Always Speak Mode, Trigger words
  Barge in

- **How does it work?**
  Acoustic models
  Language models
NLU Natural Language Understanding

- Rule based, silent word based
- Statistical

- Input – recognized phrase or N-best
- Output – action and attributes
TTS-Text To Speech
Formant synthesis

Small size
Low quality

(a) A Parallel Formant Synthesiser
Concatenate synthesis

• Connecting PCM
• High processing power and memory requirements
• Prosody
• Coarticulation
• Emotions
• Voice morphing
Dialog

- Directed dialog
- Mixed initiative
- Believe state modeling
- POMDP (Partially Observable Markov Decision processes)
Trends

- Multimodal systems
- Pervasive systems
- Natural Dialog
- Audio-Visual speech recognition
- Domain knowledge utilization
Some hints how to write the speech application

- Indicate that user speaks to the machine
- Keep in mind short term memory of the user.
- Provide “what can I say” option through the app.
- Provide “go back” option throughout the app.
- Build in an error correction mechanism
Pervasive systems - Paradigm Shift

- Applications started to reach out of PC boxes, they blend and become part of the environment, the users will be living in applications.
- This process started already in automotive industry, the home is the next.
- Interaction model needs new interaction means, mouse and keyboards will no longer suffice.
- Speech recognition and computer vision can help.
Standards

- Open source engines
  TTS, ASR, NLU
- Markup languages
  VoiceXML,
  SSML
  X+V, SALT
- JSAPI, SMAPI
<?xml version="1.0" encoding="UTF-8"?> <vxml xmlns="http://www.w3.org/2001/vxml" version="2.1">
  <form id="get_address">
    <field name="citystate">
      <grammar type="application/srgs+xml" src="citystate.grxml"/>
      <prompt> Say a city and state.</prompt>
    </field>
    <field name="street">
      <grammar type="application/srgs+xml" src="citystate.grxml"/>
      <prompt> What street are you looking for?</prompt>
    </field>
    <filled>
      <prompt> You chose <value expr="street"/> in <value expr="citystate"/> </prompt>
    </filled>
  </form>
</vxml>
<?xml version="1.0"?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <voice gender="female">Mary had a little lamb,"</voice>
  <!-- now request a different female child's voice -->
  <voice gender="female" variant="2">Its fleece was white as snow.</voice>
  <!-- processor-specific voice selection -->
  <voice name="Mike">I want to be like Mike."</voice>
</speak>
Architecture
In car UI
Some trends

- Intelligent room
- Audio-visual recognition
- Taking notes
- Person tracking, person recognition
- Situation modeling
- Question answering
Thank you!

tomas_macek@cz.ibm.com