Speech and Dialog

Tomas Macek,
tomas_macek@cz.ibm.com
Overview

- IBM Prague R&D
- NLP summary
- Voice UI design
- Dialog systems
IBM Prague R&D Lab, Watson Dialog Services
NLP technologies
Natural Language Processing

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

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

Prompt:
You mean Mary Lou at 9am tomorrow, right?
action=new_meeting
time = 9 (am | pm)
invitees = her

schedule a meeting with her at nine
Voice

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 and why not opt for speech interfaces

**It is GOOD and BAD**

- Speech is fast (large lists, dates, times)
- Speech is natural and intuitive
- Speech input device is small
- Capturing emotional state
- Determining speaker identity

- 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

Speech recognition is not the same as speech understanding!
ASR – Automatic Speech Recognition

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

- **Where?**
  - Remote (on server)
  - Local (on a client)
  - Hybrid (both)

- **What can s/he say?**
  - List of phrases
  - Grammar
  - Dictation

- **Output**
  - sentence, annotated sentence
  - N-best or lattice
  - Confidence
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
TTS Text To Speech

Formant synthesis

- Small size
- Low quality

Concatenate synthesis

- Connecting PCM
- High processing power and memory requirements
- Prosody
- Coarticulation
- Emotions
- Voice morphing
Concatenate synthesis

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

- UI: Text, GUI, Multimodal
- Rule based
- Statistical

Directed dialog, mixed initiative dialog, turn taking, Believe state modeling, Deep learning, Anaphora resolution, turn taking, POMDP - Partially Observable Markov Decision Processes
**Cognitive Avatar**

**Customer:** Technology Exploration Center, Software Group

**Description:**
- Talking head with synchronized lips and numerous face expressions
- Body gesture recognition
  - User waves to get attention
  - User moves head forward to mute the system
- Implements six selected dialogue domains (small talk, weather, time, name days, local space navigation, education).
- WDS backend component permits fast authoring and maintenance
- Grammar + Remote dictation; Dictation + NLU
- Remote microphone, microphone array, techniques of opening microphone based on noise and state of dialogue
- Situation awareness (number of people and ambient noise level considered)
- Proactive attention request activities

(no people around) It is 12:00, time for lunch
(shows menu on the screen)
(person passing by) Hi, do you know ..
(stops talking when no attention is drawn)

Hi, this is John
Hi John
What is the weather forecast
It is going to be sunny in low 30?
And next day?
It will rain Tomorrow.
Where can I find a rest room?
The rest room is at the end of the corridor, do not forget your batch (shows location on the screen)
Probabilistic Dialog Management

A data-driven approach to dialog management, consisting of a belief state tracking component and an action planning component.

Challenges:
- State-action space can be very large
- Learn mapping from belief states to actions
- Action planning requires large amounts of user interaction, which is not always practical

Approach:
- Approximately represent the distribution of dialog states
- Reinforcement Learning
- Developing alternative approaches such as user simulation
Bootstrapping a dialog application

- Observations by Beau Cronin characterizing many untapped opportunities in AI
  - Data is inherently small
  - Data cannot be interpreted without a sophisticated model
  - Data cannot be pooled across customers

- This is the world we are targeting for our dialog applications
  - Often no existing dialog transcripts or data
  - Desired dialog flow is tacit and known only to a subject matter expert
  - Customer wants something running immediately!

- Our basic approach
  - Start with a scripted dialog system
    - Rapidly assembled with the expert of the chief stakeholder
    - Use in-house dialog modeling languages
  - Transition to a POMDP-based approaches as data becomes available
IBM Bluemix
The Digital Innovation Platform

Build your apps, your way.
Use a combination of the most prominent open-source compute technologies to power your apps. Then, let Bluemix handle the rest.

- Instant Runtimes
  App-centric runtime environments based on Cloud Foundry.

- IBM Containers
  Portable and consistent delivery of your app without having to maintain it.

- Virtual Machines
  Get the most flexibility and control over your environment with VMs.
- Service vs application,
- credentials
Text to Speech

The Text to Speech service understands text and natural language to generate synthesized audio output complete with appropriate cadence and intonation. It is available in 10 voices across 5 languages.

Resources:
- API Reference
- Documentation
- Fork and Deploy on Bluemix
- Fork on Github

Input Text

The text language must match the selected voice language: Mixing language (English text with a Spanish male voice) does not produce valid results. The synthesized audio is streamed to the client as it is being produced, using the HTTP chunked encoding. The audio is returned in the Ogg Opus format which can be played using VLC and Audacity players.

Would you like to help make this service better?
- Allow Watson to learn from this session
- Opt out

Text

SSML

Conscious of its spiritual and moral heritage, the Union is founded on the indivisible, universal values of human dignity, freedom, equality and solidarity; it is based on the principles of democracy and the rule of law. It places the individual at the heart of its activities, by establishing the citizenship of the Union and by creating an area of freedom, security and justice.
Speech to Text

The IBM Watson Speech to Text service uses speech recognition capabilities to convert English, Spanish, Brazilian Portuguese and Mandarin speech into text.

Resources:

- API Reference
- Documentation
- Fork on Github
- Fork and Deploy on Github

Transcribe Audio

Use your microphone (compatible only with Google Chrome and Mozilla Firefox), upload pre-recorded audio (WAV containing uncompressed audio or FLAC file formats), drag and drop recorded audio onto the page, or use the audio samples provided. The returned text includes metadata that provides the timestamps for start and end times of individual words, confidence scores behind those words, and alternative hypotheses for phrases.

Would you like to help make this service better?

- Allow Watson to learn from this session
- Opt out

US English broadband model (16KHz)
Dear Tomas! Would you like to find a movie that's now playing or coming soon?

Movie now playing please

Are you in the mood for a specific genre?

comedy, some nice comedy

Do you prefer a certain movie rating?

yes

Which one?

R
WDS - applet
Selected trends

- Cognitive systems
- Pervasive systems
- Natural Dialog
- Multimodal systems
- 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
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
VoiceXML

```xml
<?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"
xsi:schemaLocation="http://www.w3.org/2001/10/synthesis http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
xml:lang="en-US">
  <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