Developing IVVR Applications for Teaching and Customer Self Service

K. W. (Bill) Scholz
NewSpeech, LLC
Agenda

- Introduction – the IVVR mobile paradigm
- Application development and delivery
  - Dialog flow specification
  - Animation specification
  - Application generation
- Deployment and distribution
Introduction

- Overview: interactive voice-video interface to support interactive learning and customer self-service distributed by 3G and IP network
- Network-delivered IVVR presented on mobile handsets can be effective for:
  - driving individual learning, particularly language learning
  - providing customer self-service
  - providing access to information that can be structured as a simple computer-human dialog
    - Device / appliance trouble shooting and repair
    - Situation management (e.g., managing a fender-bender)
    - Public institution access (e.g., airport check-in)
### Development & Delivery

**Lesson Builder**
- Lesson
- Relational Database

**Application Server**
- Lesson
- Video Application
- Servlet Container

- VoiceXML Interpreter
- Speech Recognizer
- Speech Synthesizer
- Video Manager

**Developer enters the application dialog and animation triggers**

**Lesson builder creates application packaged as a Web ARchive (WAR) file**

**Application is unpacked and deployed**

**Arriving call triggers VoiceXML interpreter to request VoiceXML stream; manages ASR, TTS, Video.**

**SIP / RDP data delivered to users via network**

**IP or 3G Network**
From Dialog flow to application

- Specify dialog flow: the text for all prompts and responses is entered
- Simulate the dialog flow to help find flaws:
  - Off-line simulator on developer’s computer
  - On-line execution using TTS, no video or recordings
- Generate recording script (text, filenames, characters)
- Record audio; optionally preview on line
- Generate video by mapping phoneme strings to lip sync and gestures and motions to keywords
- Build application with completed VoiceXML, speech grammars, and Audio-Video files
As a call flow diagram....

I understand you are interested in investing some money in the stock market

Yes I am interested in mutual funds

Yes I have some specific stocks in mind

Actually, I’m quite conservative and would prefer money market funds.

I see. What are your investment preferences?

OK. Which stocks are you thinking about?

Ok, are you interested in commercial paper, or bank CDs?
Dialog Flow Specification

As a call flow chart....

Turn Q2Start
Prompt I understand you are interested in investing some money in the stock market.
Response 1 Yes, I am interested in mutual funds.
Response 2 Yes I have some specific stocks in mind.
Response 3 Actually, I am quite conservative and would prefer money market funds.

Turn Label_A
Prompt I see. What are your investment preferences?
Response 1 I prefer growth stocks. But not too risky.

Turn Label_B
Prompt OK. Which stocks are you thinking about?
Response 1 Well I am thinking about some of the oil stocks, the majors like shell or ekson mobile.

Turn Label_C
Prompt I understand. Are you looking at commercial paper or at bank CDs?
Response 1 Commercial paper is ok.
Video Generation

- Recording script generated to manage audio recording
- Avatar images and background settings chosen for each character
- Keywords in prompts & responses mapped to video snippets of gestures, moods, and eye motions
- Recognizer identifies key phonemes requiring mouth movements and maps these to video snippets
- Sound dubbed onto audio track
- Completed SWF master converted and compressed for shipping via 3G network
**Recording script**

**Prompt and response text, filenames, avatar characters**

<table>
<thead>
<tr>
<th>PROMPT TEXT</th>
<th>AVI Filename</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>I understand you are interested in investing some money in the stock market.</td>
<td>MB0002_C102</td>
<td>M10_Professional</td>
</tr>
<tr>
<td>I see. What are your investment preferences?</td>
<td>MB0002_C103</td>
<td>M10_Professional</td>
</tr>
<tr>
<td>Any interest in biotech? Or how about international plays?</td>
<td>MB0002_C104</td>
<td>M10_Professional</td>
</tr>
<tr>
<td>Yes, I am interested in mutual funds.</td>
<td>MB0002_U102</td>
<td>F8_Sporty Casual</td>
</tr>
<tr>
<td>Yes I have some specific stocks in mind.</td>
<td>MB0002_U202</td>
<td>F8_Sporty Casual</td>
</tr>
<tr>
<td>Actually, I am quite conservative and would prefer money market funds.</td>
<td>MB0002_U302</td>
<td>F8_Sporty Casual</td>
</tr>
<tr>
<td>I prefer growth stocks. But not too risky.</td>
<td>MB0002_U103</td>
<td>F8_Sporty Casual</td>
</tr>
<tr>
<td>etc.</td>
<td>etc.</td>
<td>etc.</td>
</tr>
</tbody>
</table>
Lip sync – text to mouth movement

Text strings processed by speech recognizer into phoneme strings

....soon....

Phonemes mapped to artist renderings of mouth movements

Mouth movements video snippets merged with avatar
**Gestures, moods, eye motions**

- **Flow chart snippet used to guide avatar construction**

<table>
<thead>
<tr>
<th>a_characters</th>
<th>a_eyemotions</th>
<th>a_holdmood (Preset Mood)</th>
<th>Mood Code</th>
<th>a_mood</th>
<th>Hold Gestures Code</th>
<th>a_holdgestures (Preset Hold Gestures)</th>
<th>Gestures Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1_business professional</td>
<td>eyes look up for a moment</td>
<td>normal_mood</td>
<td>mood_normal</td>
<td>normal_mood</td>
<td>H_L_01_hold</td>
<td>H01_LEFT finger upper left_hold</td>
<td></td>
</tr>
<tr>
<td>F2_business professional</td>
<td>eyes look down for a moment</td>
<td>02_pleasant_mood</td>
<td>mood_01</td>
<td>01_very pleased_mood</td>
<td>H_R_01_hold</td>
<td>H01_RIGHT finger upper right_hold</td>
<td></td>
</tr>
<tr>
<td>F3_airline/rentalcar/uniform</td>
<td>eyes look left for a moment</td>
<td>04_bad mood</td>
<td>mood_02</td>
<td>02_pleasant_mood</td>
<td>H_L_02_hold</td>
<td>H02_LEFT finger upper right_hold</td>
<td></td>
</tr>
<tr>
<td>F4_airline/rentalcar/uniform</td>
<td>eyes look right for a moment</td>
<td>05_attentive/interested_mood</td>
<td>mood_03</td>
<td>03_curious_expression</td>
<td>H_R_02_hold</td>
<td>H02_RIGHT finger upper left_hold</td>
<td></td>
</tr>
<tr>
<td>F5_business professional</td>
<td>eyes looks left and hold w/ blink state.</td>
<td>08_frustrated_mood</td>
<td>mood_04</td>
<td>04_bad mood_mood</td>
<td>H_03_hold</td>
<td>H03_finger pointing oneself_hold</td>
<td></td>
</tr>
<tr>
<td>F6_business professional</td>
<td>eyes looks right and hold w/blink state.</td>
<td>09_angry_mood</td>
<td>mood_05</td>
<td>05_attentive/interested_mood</td>
<td>H_04_hold</td>
<td>H04_finger up elbow down_hold</td>
<td></td>
</tr>
<tr>
<td>F7_senior business professional</td>
<td>eyes looks up and hold w/blink state</td>
<td>13_indifferent/cool mood</td>
<td>mood_06</td>
<td>06_apologetic_expression</td>
<td>H_05_hold</td>
<td>H05_finger to object in front_hold</td>
<td></td>
</tr>
</tbody>
</table>

- Etc.
Keyword-triggered animation

<table>
<thead>
<tr>
<th>Greeting</th>
<th>hi mr barrett i'm hank leno from digital international</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword:</td>
<td>1: hi 2: international</td>
</tr>
<tr>
<td></td>
<td>hi hank it's a pleasure to finally meet you</td>
</tr>
<tr>
<td></td>
<td>1: meet 2: pleasure</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label 1</th>
<th>me too thanks</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword:</td>
<td>1: thanks</td>
</tr>
<tr>
<td></td>
<td>i heard many good things about you</td>
</tr>
<tr>
<td></td>
<td>1: many</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label 2</th>
<th>oh thank you very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword:</td>
<td>1:</td>
</tr>
<tr>
<td></td>
<td>as i mentioned in my email you have a very interesting technology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Animation Features - Windows Internet Explorer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword</strong> international</td>
</tr>
<tr>
<td><strong>Mood 1</strong> 01_curious_expression</td>
</tr>
<tr>
<td><strong>Gesture 1</strong> HB1_LEFT finger upper left_up</td>
</tr>
<tr>
<td><strong>Eye Motion 1</strong></td>
</tr>
<tr>
<td><strong>Background 1</strong></td>
</tr>
</tbody>
</table>
Lesson Builder Template Processing

^1 Lesson Name
^2 Lesson Start Form Title
^3 Sensitivity Default Status
^4 Barge-in Default Status
^5 Greeting Prompt Text
^6 Document Title

Generated VoiceXML Code
Example: teaching ESL

- Video – student practicing English with mobile handset
Summary

- The Interactive Voice-Video Response paradigm
- Application development
  - Dialog flow specification
  - Prompt recording and lipsync generation
  - Animation specification, keyword-triggers
  - Generate vxml, grammars, and avi
  - Build into industry-standard web archive
- Delivery: J2EE app server, VXML Media Server, SIP/RTP to client handset
Thank You.

NewSpeech, LLC.
Advanced Speech Solutions
Karl W. 'Bill' Scholz, Ph.D.
President
203 Vassar Circle
Stratford, PA
19087
cell: 610-212-8016
drive: 610-989-9089
bill.scholz@comcast.net
http://www.newspeechsolutions.com/