Internet Applications

Lecture 11 – Voice User Interface Design for Telephone Applications

Steve Young, Feb 2003
Modified by Jason D. Williams, Feb 2004 & Feb 2005

Cambridge University Engineering Department
Machine Intelligence Lab
Outline: IA Lectures 11-14

• Lecture 11
  • The Voice Browser & relevant properties
  • Why designing speech applications for the telephone is difficult
  • Typical Voice User Interface (VUI) design approaches

• Lectures 12-13
  • VoiceXML
  • Speech Recognition Grammar Specification (SRGS)

• Lecture 14
  • Worked example (Design & VoiceXML implementation)
Welcome to the Entropic Cinema Information Line for the week ending tomorrow. Which town or city would you like Cinema information for? 
Reading 
Cinema information for Reading. What information would you like?
Information on the Odeon Cinema please. Please confirm or correct me, do you want times for the Matrix at the Odeon today?
No, I would like times for Star Wars please.
Do you want times for Star Wars the Phantom Menace at the Odeon today?
Yes please
It is on at 12.30 and 5.40. ...
Voice Browser Architecture (1/3)

- Voice Browser encapsulates:
  - Call control
  - Speech recognition
  - Audio playback
  - Provides layer of abstraction between these services and the content
  - Enables re-use of web server infrastructure
Some key assumptions: Voice browser

- Dialogue pages (and audio, and grammars) can be downloaded without perceptible delay
- Acceptable human-machine dialog can be scripted using a “Dialogue Mark-up Language”

Addressing Latency

- where possible, pre-compile and cache (compiled) grammars
- where possible, pre-fetch and cache audio
- where possible, pre-fetch and scan control pages
- carefully consider amount of computation on “page” (i.e., voice browser) vs. on page server

A suitable mark-up language – VoiceXML – next lecture
Inside the Voice Browser

Voice Browser

- Text-to-Speech Synth
- Prompt playback
- HTTP interface
- Silence Detector/Barge-in
- Recogniser
- Call control
- Telephony Interface

Voice Browser connects to the server through
www.info.com

Server contains:
- Dialogue control pages
- Prompt audio
- Reco grammar
- Data

Very simplified!
Basic Question/Answer Cycle

(1) Load Grammar

travel.gram

$toplace = to $place
$fromplace = from $place
$place = London | Manchester | ....
$route = $toplace [ $fromplace ] ...

(2) Output Prompt

Prompt playback

Text-to-Speech

(3) Start Listening for speech

Silence Detector

(4) Start Recognising

Recogniser

route($toplace(London), $fromplace(Manchester) )

(4b) Cut prompt

(5) Return results

Telephony Interface
Why VUI design is hard (1/2)

Each question/answer cycle has 3 possible outcomes…

1. Obtained a recognition result (Reco)
2. Recognition result was rejected because confidence score was below a confidence threshold
   • Also called a “no-match” (NM)
3. Endpoint never found speech
   • Also called “timeout” / “no-input” (NI)

… but these “outcomes” give incomplete – or even misleading – information about what the user really said/did!

The human/computer speech recognition interface can be viewed as a “noisy channel”
Endpoint issues (1/2)

- Barge in – user must be able to interrupt the prompt
- Echo Cancellation – must remove prompt audio output (enables barge-in)
- Silence detection – need to know reliably when user speaks
### Endpointer issues (2/2)

#### End-pointer...

<table>
<thead>
<tr>
<th>Triggered</th>
<th>Didn’t trigger</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OK</strong></td>
<td><strong>Bad</strong></td>
</tr>
<tr>
<td>(Hope recognition is right!)</td>
<td>Talk louder?</td>
</tr>
<tr>
<td><strong>Bad</strong></td>
<td><strong>OK</strong></td>
</tr>
<tr>
<td>Background noise?</td>
<td>Inattention/busy?</td>
</tr>
<tr>
<td>Hope recognition rejects this</td>
<td>Doesn’t know what to say?</td>
</tr>
<tr>
<td>• Move to quieter place</td>
<td>Didn’t know it was time to talk?</td>
</tr>
<tr>
<td>• Speaker phone?</td>
<td>Use DTMF alternatives?</td>
</tr>
<tr>
<td>• Use DTMF alts?</td>
<td></td>
</tr>
</tbody>
</table>

#### User…

- **Spoke**
- **Didn’t Speak**

#### Outcome:
- Recognition attempt
- No-input
Recognition issues (1/2)

• Currently deployed systems use conventional HMM-based recognition.

• Language models
  • Most use finite-state (or limited recursion CFG) based grammar constraints
  • Some (more cutting edge) systems use N-Grams (especially for call routing)

• Both techniques produce a confidence score

• Score is compared with the confidence threshold to “accept” or “reject”
# Recognition issues (2/2)

## Recognizer behaviour

<table>
<thead>
<tr>
<th>In-grammar</th>
<th>Utterance</th>
<th>Outcome:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept (correct)</td>
<td>Accept (wrong)</td>
<td>Reject</td>
</tr>
<tr>
<td><strong>CAin</strong></td>
<td><strong>FAin</strong></td>
<td><strong>FRin</strong></td>
</tr>
<tr>
<td>OK</td>
<td>Very bad!</td>
<td>Bad!</td>
</tr>
<tr>
<td></td>
<td>- Confirm: give caller the chance to catch mis-rec</td>
<td>- But user just needs to repeat themself</td>
</tr>
<tr>
<td><strong>CAout</strong></td>
<td><strong>FAout</strong></td>
<td><strong>CRout</strong></td>
</tr>
<tr>
<td>Bonus!</td>
<td>Very bad!</td>
<td>Fine</td>
</tr>
</tbody>
</table>
| | - Confirm: give caller the chance to catch mis-rec | - Caller confusion  
| | | - Wrong expectation  
| | | - Speech quality  
| | | - End-pointer error? |

**Recognition Result obtained**

**No-match**

12
Why VUI design is hard (2/2)

More problems…

- How does the user know what options are available?
- (Very) limited persistence of memory (3 items?)
- Turn-taking model unnatural
- Widely varying expectations in general population

Conversational interface, but not a conversation

- DTMF systems were often (poorly) designed by IT groups, not human-factors people ( "Voice-mail jail" )
- Bad VUI design in a speech app can make it impossible to use
VUI guidelines (1/2)

Structural

- Every recognition result could be a false-accept
  - Use confirmations and/or explain how to back-up
  - Confirm early – correcting more than 1-2 items at a time is hard
- Give the caller some control with global commands
  - Help, Repeat, Go back, start over, operator
- Consider all reasons why a no-match or no-input occurs
  - Confusion, Inattention, Out of grammar, False-reject
- Collapse needless interactions
- Avoid long output – divide into “chunks”
- Where possible, provide DTMF fall-back
- Focus on “the successful first-time user”
  - Provide shortcuts & hints for repeat callers
- Prioritize for task completion & reco accuracy first (not speed)
VUI guidelines (2/2)

Wordings

- Problem of short term memory
- Keep lists of options short (3?) (one can be “more options”)
- Introduce commands (like globals) when needed (not before)
- Make the explanation of the command *equal to the command*
- Use tailored, escalating no-match and no-input problems
  - These are a form of help, not errors
- Give examples instead of explanations
- Consistency counts for a lot
- Use conversational norms wherever possible
# VUI guidelines examples (1/3)

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
</table>
| **System:** What's your account number?  
**User:** 223344  
**System:** Sorry, that is an invalid account number. Please try again. What’s your account number? | **System:** What’s your account number?  
**User:** 223344  
**System:** Hmm… I heard 223345, but I don't recognize that account. Please tell me your account number again, or say OPERATOR. |
| **System:** What extension would you like to call?  
**User:** [silent – doesn’t know the ext of John Smith]  
**System:** Please speak louder. What extension would you like to call? | **System:** What extension would you like to call?  
**User:** [silent]  
**System:** Sorry, I didn’t hear anything. Please tell me an extension number or say HELP and I’ll give you a list. |
| **System:** Ok, I’ve completed your order.  
If you want to start over, say START OVER; if not, to end the call, say GOODBYE.  
**User:** Start over  
**System:** We’re back at the beginning… | **System:** Ok, I’ve completed your order.  
If you’re finished, feel free to hang up. Otherwise, hang on a moment, and I’ll take you back to the beginning. [pause] We’re back at the beginning… |
### VUI guidelines examples (2/3)

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System:</strong> To delete, say REMOVE; to hear again, say REPEAT.</td>
<td><strong>System:</strong> You can say REPEAT or REMOVE</td>
</tr>
<tr>
<td><strong>User:</strong> Repeat again</td>
<td><strong>User:</strong> Repeat</td>
</tr>
<tr>
<td><strong>System:</strong> To reach an extension, please tell me the four-digit</td>
<td><strong>System:</strong> You can say an extension number like 3324, or a name like</td>
</tr>
<tr>
<td>extension number, or say a person’s first and last name. So, how can</td>
<td>SUSAN SMITH. So, how can I direct your call?</td>
</tr>
<tr>
<td>I direct your call?</td>
<td><strong>System:</strong> Welcome to Acme. <strong>[pause]</strong> Your important number is</td>
</tr>
<tr>
<td><strong>System:</strong> Welcome to Acme. At any point you say REPEAT to hear</td>
<td>2234. Remember, you can always say REPEAT.</td>
</tr>
<tr>
<td>something again. Let’s get started.</td>
<td>[pause]** Now, do you want to do something else?</td>
</tr>
<tr>
<td><strong>System:</strong> Your important number is 2234. Do you want to do</td>
<td><strong>User:</strong> Repeat</td>
</tr>
<tr>
<td>something else?</td>
<td><strong>User:</strong> Repeat</td>
</tr>
<tr>
<td><strong>User:</strong> Not yet, can you say that again (?)</td>
<td></td>
</tr>
<tr>
<td><strong>System:</strong> Thanks for calling Acme. You’ve reached our a speech</td>
<td><strong>System:</strong> Thanks for calling Acme. You’ve reached our a speech</td>
</tr>
<tr>
<td>recognition system. How can I help you?</td>
<td>recognition system. I can understand things like WHAT’S MY BALANCE or</td>
</tr>
<tr>
<td><strong>User:</strong> Uhh... Account balance request status</td>
<td>I NEED TO PAY A BILL. So, how can I help you?</td>
</tr>
<tr>
<td></td>
<td><strong>User:</strong> I need to check my balance</td>
</tr>
</tbody>
</table>
### VUI guidelines examples (3/3)

#### Before

<table>
<thead>
<tr>
<th>System:</th>
<th>Are you calling about a <em>new</em> booking?</th>
</tr>
</thead>
<tbody>
<tr>
<td>User:</td>
<td>No</td>
</tr>
<tr>
<td>System:</td>
<td>Ok, so you're calling about an <em>existing</em> booking. To change that, say GO BACK. Otherwise…</td>
</tr>
<tr>
<td>User:</td>
<td>Go back… ?</td>
</tr>
<tr>
<td>System:</td>
<td>To access your balance, transfer funds, or for other services on your current accounts, say ACCOUNTS. To report a lost or stolen credit card or ATM card, say LOST CARD. For branch hours and information, say BRANCH INFORMATION. And for additional choices, say MORE OPTIONS.</td>
</tr>
<tr>
<td>User:</td>
<td>More options</td>
</tr>
<tr>
<td>System:</td>
<td>If you'd like to change your PIN, say PIN CHANGE. Or if you're a merchant calling to verify a check, please say MERCHANT CHECK VERIFICATION. And to hear a brief tutorial about how to use this system, you can say INSTRUCTIONS. If you want to hear the first list of options again, just say GO BACK.</td>
</tr>
</tbody>
</table>

#### After

<table>
<thead>
<tr>
<th>System:</th>
<th>Are you calling about a <em>new</em> booking?</th>
</tr>
</thead>
<tbody>
<tr>
<td>User:</td>
<td>No</td>
</tr>
<tr>
<td>System:</td>
<td>Ok, so you’re calling about an <em>existing</em> booking. If I got that wrong, say GO BACK. Otherwise, please hold the line.</td>
</tr>
<tr>
<td>User:</td>
<td>[silent]</td>
</tr>
<tr>
<td>System:</td>
<td>Here’s a list of choices; when you hear the one you want, say it. MY ACCOUNTS, LOST CARD, BRANCH INFORMATION, or MERCHANT CHECK VERIFICATION.</td>
</tr>
</tbody>
</table>
Prompt Generation

A variety of methods are used

a) play back pre-recorded utterance  simple canned response
b) use slot and filler techniques  limited domain eg weather forecasts
c) use full text-speech synthesis  open domain eg news reports

Slot and filler approaches can give excellent results if fillers reflect prosodic context appropriately.
Prompt Generation - examples

“Thanks for calling Cambridge Pizza. What would you like to order?”

First time asking - falling

“Sorry, I didn’t catch that. What would you like to order?”

Second time asking - rising

“You have ordered two regular pizzas, and one seafood pizza.”

rising falling

“You have ordered two regular pizzas, and one seafood pizza.”

rising falling

“You have ordered two regular pizzas, and one seafood pizza.”

mid rising mid falling

“You have ordered two regular pizzas, and one seafood pizza.”

mid rising mid falling
An example call

System: Hi, I’m … a virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? [pause] You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano

An example call

System: Hi, I’m a … virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano


Most callers are repeat callers.

Keep opening brief; don’t provide instructions for every caller.
An example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call?  [pause] You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano


...but this caller is new & says just the first name. This results in a **misrecognition!**

For low confidence scores, use explicit confirmation
An example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano

An example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard **Tom McCurry**. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? You can say things like, **Suzanne Smith**, the **sales department**, **Extension 20**, or just say **operator**. To hear all your choices, say **help**.

Caller: Larry Piano

System: Transferring to **Larry Piano**. Say stop to cancel. **Connecting**.

Experienced caller (still) knows how to respond – novice caller more likely to pause

Short pause
An example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? [pause] You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano


1 misrec + waited through short pause... caller probably needs help

Give some examples & mention help command
An example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Uh, Larry?

System: I heard Tom McCurry. Is that right?

Caller: No

System: Oh, I’m sorry – How can I direct your call? [pause] You can say things like, Suzanne Smith, the sales department, Extension 20, or just say operator. To hear all your choices, say help.

Caller: Larry Piano


Hi-confidence score this time

Use implicit confirmation, but make it clear how to correct the system.
Another example call

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: I heard Larry Piano, home office. Is that right?

Caller: Yes

System: Please hold while I transfer your call. Sorry, but I don’t have that phone number for Larry Piano. We’re back at the beginning – how can I direct your call?

Caller: ???
Another example call: Problem

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: I heard Larry Piano, home office. Is that right?

Caller: Yes

System: Please hold while I transfer your call. Sorry, but I don’t have that phone number for Larry Piano. We’re back at the beginning – how can I direct your call?

Caller: ???

Dialog was built from a tech flow chart:

1) Get & confirm phone number
2) Initiate transfer
3) Catch errors from transfer

Use sample dialogues to design the flow of the call, not flow charts.
Another example call: Improved (1/2)

System: Hi, I’m a … virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: I heard Larry Piano, home office. Is that right?

Caller: Yes

System: Sorry, but I don’t have that phone number for Larry Piano. Try saying something like Larry Piano, cell phone. How can I direct your call??

Caller: Larry Piano, cell phone
Another example call: Improved (1/2)

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: I heard Larry Piano, home office. Is that right?

Caller: Yes

System: Sorry, but I don’t have that phone number for Larry Piano. Would you like to try his… CELL PHONE, OFFICE PHONE, or NEITHER OF THESE.

Caller: Cell phone

Clearly a better user experience…

… but can we do better yet?
Another example call: Improved (2/2)

System: Hi, I’m a ... virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: Hmm… I heard Larry Piano, home office, but I don’t have that phone number. Try saying something like Larry Piano, cell phone. How can I direct your call?

Caller: Larry Piano, cell phone
Another example call: Improved (2/2)

System: Hi, I'm a ... virtual auto-attendant. How can I direct your call?

Caller: Larry Piano at his home office

System: Hmm... I heard Larry Piano, home office, but I don't have that phone number. Try saying something like Larry Piano, cell phone. How can I direct your call?

Caller: Larry Piano, cell phone

2 possible problems:
1) Don’t have the phone number
2) Misrecognition

This approach addresses both & saves a step
### Summary

- Key idea is that voice dialogs can be represented by a scripting language.
- This allows recognition, synthesis, and telephony to be abstracted away from application implementation, using Web standards (HTTP).
- In addition to usual quality/accuracy issues, recognition processing must encompass robust silence detection and barge-in capability.
- We can mitigate these limitations with successful VUI design:
  - Consider all reasons for No-match and No-input events.
  - Use brief prompts to accommodate persistence of memory limitations.
  - Enable caller to correct mistakes & take (some) initiative.
  - Reflect conversational norms (e.g., prosody).
APPENDIX

Students are not responsible for new material beyond this point
Voice Web Architecture

- Dialog Pages
- www.info.com
- http requests
- Internet
- Dialog Process
- Browser
- Telephone Exchange
- PSTN
- Dialog Server
- Media Server
- Etherswitch
Silence Detection

Can distinguish speech from silence using simple energy detector, however

- some speech sounds are typically very low amplitude eg fricatives
- many sources of false triggering eg line noise, coughs, laughter

Best silence detector is recogniser itself, hence basic strategy is

- buffer 1-2 seconds of speech
- use energy based detector with low threshold ie allow false triggering
- when energy detector triggers, start recognising from beginning of buffer
- include silence models in recogniser, if no speech recognised then reset recogniser

When listening for speech, always include a time-out since user may speak and be unheard. Typical response to time-out is

“Sorry, I didn’t hear that – please speak louder”
Key recognition issues:

- Efficiency – need at least 10 recognition channels per cpu
- Memory – need to limit per channel memory to a few Mbyte
- Noise robust – ideally should suppress background noise and transients caused by network errors

Typical techniques

- keep vocabularies and grammars as small as possible
- use highly optimised search routines
- implement in integer arithmetic and store model parameters as shorts (or even bytes)
- share model sets and dictionaries
- run each recogniser in its own thread and keep an active pool to minimise start-up time.
Barge-In and Echo Cancellation

Barge-in requires “listening” to be enabled as soon as a prompt is issued. The prompt often “reflects back” into the input channel and this creates two main problems:

1. silence detector can be falsely triggered by the prompt
2. recogniser can be confused by prompt since it appears as background noise

Solution to (1) is to calibrate during opening prompt by building an energy histogram of energy levels in each 20msec segment

S: Hello, this is the Acme Company web site. <Pause>
What service do you want?
U: Stock information, please.

Set Threshold
Solution to (2) is to apply adaptive equalisation.

Adaptive filter is variable delay line with filter coefficient adjusted on-line to minimise energy in combined response plus delayed prompt signal. More sophisticated systems can also adapt to spectral distortion applied to prompt.
Noise Compensation

Available computer power and limited adaptation data limits noise compensation to fairly simple schemes

- running average cepstral mean subtraction
- multi-style model training
- spectral subtraction
- feature vector duplication when frames are dropped (mobile only)
11.1 Voice Browser Architecture (2/3)

"Premise-based" solution

- Re-use web infrastructure
- Standards help protect application investments, skills required, etc.

PSTN

Voice Browser

Intranet

Acme, Ltd.

- Dialogue control pages
- Prompt audio
- Reco grammar
- Data
11.1 Voice Browser Architecture (3/3)

“Voice Service Provider” (VSP) solution

- Two approaches:
  1. “Outsourced telephony service”: Answer the phone for a company; outsource speech recognition service
  2. “The Voice Web”: Enable a caller to surf from site to site

- Note importance of standards!
Some additional assumptions for a VSP-based “Voice Web”

- Dialogues can be made sufficiently uniform that “surfing” from web site to web site will be possible
- Speech recognition/synthesis is expensive, hence provide shareable server farms eg in telephone exchange
- A single speech recogniser can provide sufficient accuracy to cover all possible applications

some of these assumptions are problematic, hence current systems are limited to very specific domains known a-priori

the future: PDAs?
11.8 Business Issues: Voice Browser

In general, the voice browser architecture has been positively received in industry.

- vendor-independent languages (e.g., VoiceXML)
- re-use existing web infrastructure, staff skills
- for VSP-based solutions, allows companies to switch between VSPs rapidly, or outsource on an ad-hoc basis
- for premise-based solutions, allows buyers to “mix-and-match” components (ASR, TTS, etc.)

Some possible concerns:
- More complicated than traditional (e.g., DTMF) systems
- Larger hardware investment
11.9 Business Issues: VSPs

Ultimately, success of VSP and voice web concept depends on ability to generate revenues. Business analysis raises some important issues.

- will 3rd party recognisers be good enough?
- how will the output voice sound to my customers?
- will there be excessive delays within the dialog?

VSP outsourcing:
- is a VSP cost effective vs. premise-based?
- how to integrate with my call centre?
- will the telephony costs actually be higher?

“The Voice Web”:
- who pays for telephone calls?
- how much of my revenue is the VSP taking?
- how can I retain customer loyalty?
11.10 Future directions – Aurora project

The Aurora Project – a multi-vendor consortium designing a standard noise robust front end to go inside a mobile phone.
Aurora enables architectural variations

Distributed-script Model

Distributed-audio Model

Distributed audio addresses some corporate concerns about VSPs, Voice Browser, etc.