
ALTA Summer School 2003 Dialog Systems Session 4: The bigger picture

Dominique Estival
Dominique.Estival@dsto.defence.gov.au

The bigger picture

Topics covered:

1. VoiceXML (see also other standards, e.g. SALT)
2. Audio output: TTS and recording
3. Evaluation:
4. Other technologies
5. Non-telephone dialogue systems
6. Conclusions

VoiceXML

- special-purpose programming language for developing speech interfaces
- VoiceXML Forum formed in 1999 by AT&T, IBM, Lucent Technology and Motorola: make Internet content available by phone and voice
- VoiceXML 1.0 released in March 2000
- VoiceXML 2.0 in February 2003
- W3C Voice Browser Working Group: <http://www.w3.org/Voice> will define new versions of VoiceXML

VoiceXML

- simplifies the development of voice-enabled Web sites
- basic elements are verbal forms and menus
- VoiceXML documents can be static or dynamically generated
- VoiceXML documents can use the same business logic and databases as the visual Web

VoiceXML examples: <http://www.w3c.org/Voice/Guide/>

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<vxml version="2.0" lang="en">
<form>
<block>
<prompt bargein="false">Welcome to Travel Planner!
<audio src="http://www.adline.com/mobile?code=12s4"/>
</prompt>
</block>
</form>
</vxml>
```

VoiceXML examples: <http://www.w3c.org/Voice/Guide/>

```
<?xml version="1.0"?> <vxml version="2.0">  
<menu>  
  <prompt> Say one of: <enumerate/> </prompt>  
  <choice next="http://www.sports.example/start.vxml"> Sports </choice>  
  <choice next="http://www.weather.example/intro.vxml"> Weather </choice>  
  <choice next="http://www.news.example/news.vxml"> News </choice>  
  <noinput>Please say one of <enumerate/> </noinput>  
</menu>  
</vxml>
```

VoiceXML examples: <http://www.w3c.org/Voice/Guide/>

Computer: Say one of: Sports; Weather; News.

Human: Astrology

Computer: I did not understand what you said.

(a platform-specific default message.)

Computer: Say one of: Sports; Weather; News.

Human: Sports

Computer: *(proceeds to <http://www.sports.example/start.vxml>)*

Audio output: TTS vs recorded prompts

1. naturalness vs costs for recordings
14 hrs of recording for an average app (Kotelly 2003)
2. availability of voices for TTS
e.g. Australian accent, choice of male/female voice
3. pre-recorded prompts sound better, but less natural when segments must be spliced
4. TTS better for dynamically changing information, e.g. news, weather

Audio output: TTS

- Types of TTS
 - Formant TTS
 - sounds the worst
 - needs only low computer power, disk space, or memory
 - Concatenative TTS
 - can sound very good
 - requires fast computers and much more disk space and memory
- Fine-tuning TTS
 - pauses
 - prosody

Other technologies

1. speaker identification

- select speaker among pool of candidates (tens or hundreds)

- individual enrolment, background model

2. speaker verification

- ensure speaker belongs to the pool of candidates (thousands)

- language model, background model

3. language ID, gender ID

- similar techniques

Evaluation measures

- accuracy: WER
 - can be misleading: not all words are equally important
 - need to measure content
- transaction success
 - from the point of view of the system or of the user?
- time taken to complete transaction
 - can count absolute time or number of turns
- user satisfaction
 - how do you measure it?

How to perform the evaluation

- real system
 - reveals real problems
 - obviously need to have the system: may be too late to correct problems
- Wizard-of-Oz
 - double-edged sword: can ignore system performance
 - but doesn't show what really happens with system (e.g. DB access)
- usability testing, during development (with real users)
- pilot tests (a few hundred calls)
- partial deployment (<10,000 calls)

Non-telephone dialogue systems

1. VoIP (VoiceXML)
2. voice interaction with personal computers:
 - ex: dictation, data voice entry, personal digital assistant
 - can use speaker-dependent ASR
 - multi-media (e.g. maps, charts, etc.)
3. Virtual characters (ex: FOCAL Virtual Advisers)
 - input: microphone (vs telephone)
 - integration of multi-modal information (e.g. visual, gesture)
 - output: lip-synching

FOCAL Virtual Advisers



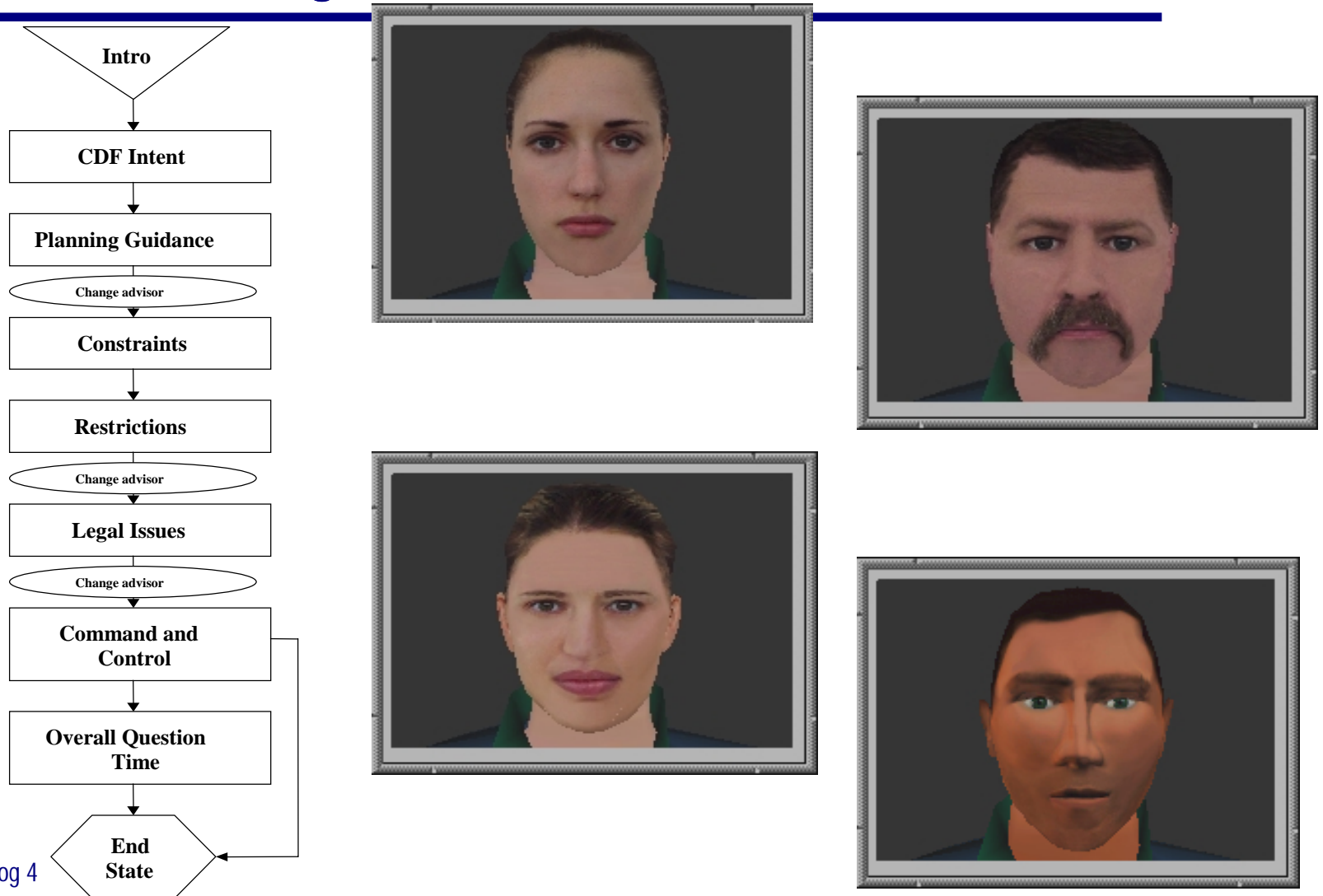
Virtual Advisers in FOCAL

provide information and planning advice to the human users

- act as a natural interface between users and complex information systems (e.g. can launch and run other apps)
- virtual participant in the planning and analysis process
- VAs as domain specialists, presenting different types of info
- multiple VAs advisers at the same time, e.g. UN rep, allied forces (US/UK...), Intelligence, Logistics ...

⇒ future: attend to the on-going discussion between users and intervene to present appropriate information, if needed

Commander's initial guidance to the TPG



Attitude Dialogue Agents

- *Speaker*
 - receives AV pairs from NL component (Regulus/Nuance)
 - produces a corresponding Attitude expression
- *Conductor*:
 - receives Attitude expression from *Speaker*
 - forwards it on to *IS* agents; waits for responses
 - sends lists of Attitude expressions to be presented to users
- *MMP* (Multimedia Presenter)
 - iterates through list of expressions sent by *Conductor*
 - chooses the media to be used to present each expression
- *NLG* (Natural Language Generator)
 - uses templates to transform Attitude expressions into English
- *IS*: wrappers to Information Source, or Domain specialists

Example

"What are the restrictions on accessing PNG airspace?"

NL: (**question** whatquestion **concept** restriction **obj1** png airspace)

Speaker: (comm_act (restriction png airspace ?restrictions) from speaker type whquestion in_response_to null)

Conductor to all IS: (restriction png airspace ?restrictions)

Restrictions:

I believe (restriction png airspace (active 2010 MOA)) in ?kb 0.0

Conductor tells MMP to present:

(restriction png airspace (active 2010 MOA))

NLG: (create_template 0.1 (restriction png airspace (active 2010 MOA)) whanswer "....")

TTS: *"The 2010 MOA with PNG remains extant. It permits access to PNG airspace."*

SLDSs: Conclusions

- development
- research
- commercialisation
- user acceptance

Some References

Calder, Jo, Klein, Ewan, Moens, Mark, Zeevat, Henk. (1986). "Problems of Dialogue Parsing" in *Syntactic Parser for Dialogue*. Edinburgh: Centre for Cognitive Science. ACORD Deliverable T2.1.

Carberry, Sandra, Lambert, Lynn. (1999). "A Process Model for Recognizing Communicative Acts and Modeling Negotiation Subdialogues". *Computational Linguistics*, vol. 25, no.1, pp. 1-53.

Estival, Dominique, Broughton, Michael, Zschorn, Andrew, Pronger, Elizabeth. (2003). "Spoken Dialogue for Virtual Advisers in a semi-immersive Command and Control environment". *Proceedings of the SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.

Johnston, Michael, Bangalore, Srinivas, Vasireddy, Gunaranjan, Stent, Amanda, Ehlen, Patrick, Walker, Marilyn, Whittaker, Steve, Maloor, Preetam. (2002). "MATCH: an Architecture for Multimodal Dialogue Systems". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pp-376-383.

Some References

Kotely, Blade. (2003). *The Art and Business of Speech Recognition: Creating the Noble Voice*. Boston: Addison-Wesley.

McTear, Michael. (2002). "Spoken Dialogue Technology: Enabling the Conversational User Interface". *ACM Computing Survey*, vol. 34, no.1, pp.90-169.

Moore, Johanna, Paris, Cécile. (1993). "Planning texts for advisory dialogues: Capturing intentional and rhetorical information". *Computational Linguistics*, vol.19, no.4, pp.661-694.

Prasad, Rashmi , Walker, Marilyn. (2002). "Training a Dialogue Act Tagger for Human-Human and Human-Computer Travel Dialogues". Proceedings of 3rd SIGDIAL. U. Pennsylvania. pp.162-173 .