

---

# 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
2. Audio output
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

---

- Evaluation measures:
  - accuracy: WER
  - transaction success
  - time taken to complete transaction
  - user satisfaction

## Evaluation

---

- How to perform the evaluation:
  - real system
  - Wizard-of-Oz
  - usability testing
  - pilot tests

## 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
3. Virtual characters (ex: FOCAL Virtual Advisers)
  - input: microphone (vs telephone)
  - integration of multi-modal information (e.g. visual, gesture)
  - output: lip-synching

## FOCAI Virtual Advisers

---



## SLDSs: Conclusions

---

- development
- research
- commercialisation
- user acceptance