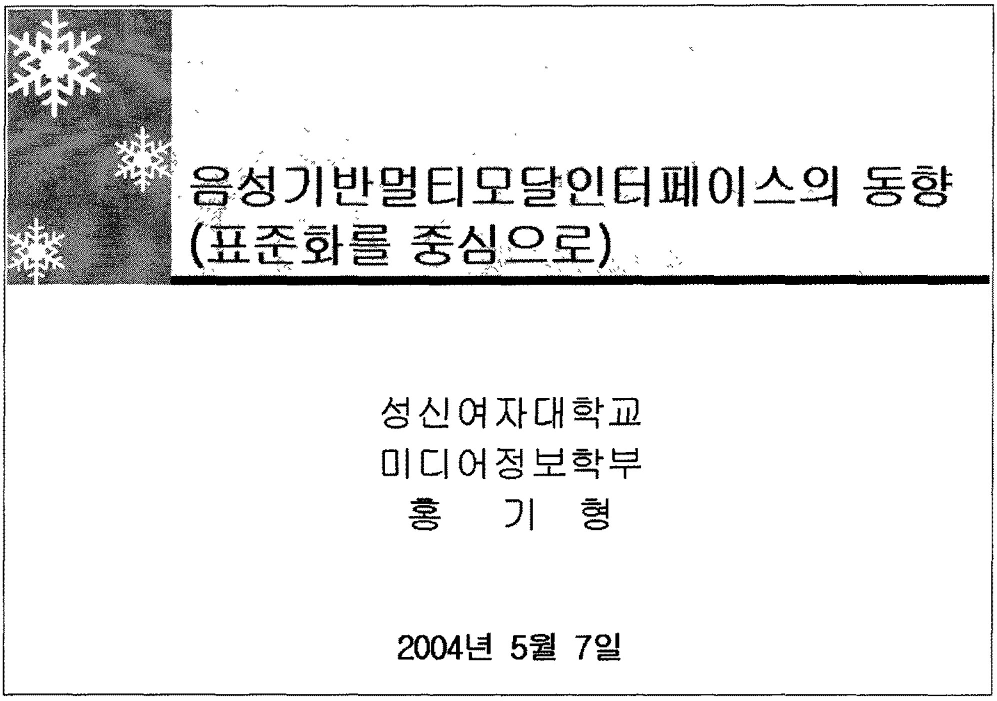
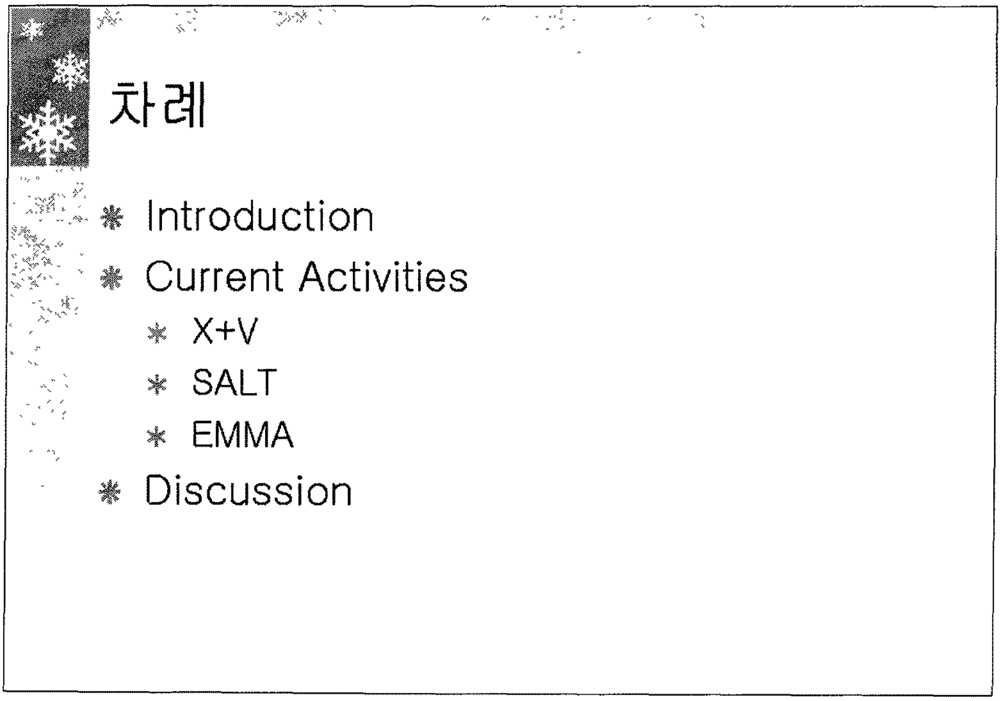


음성기반멀티모달인터페이스의 동향 (표준화를 중심으로)

성신여자대학교
미디어정보학부
홍기형

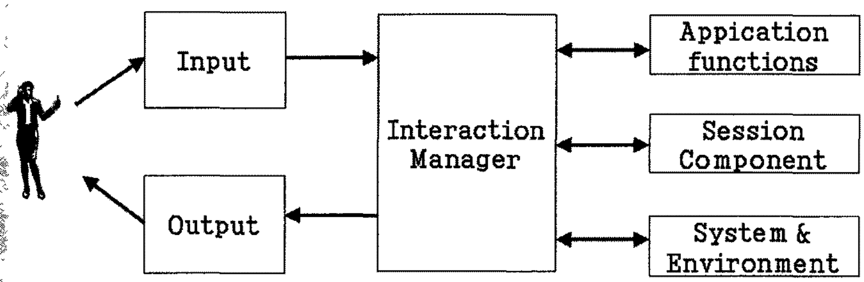
2004년 5월 7일



차례

- Introduction
- Current Activities
  - X+V
  - SALT
  - EMMA
- Discussion

Interaction



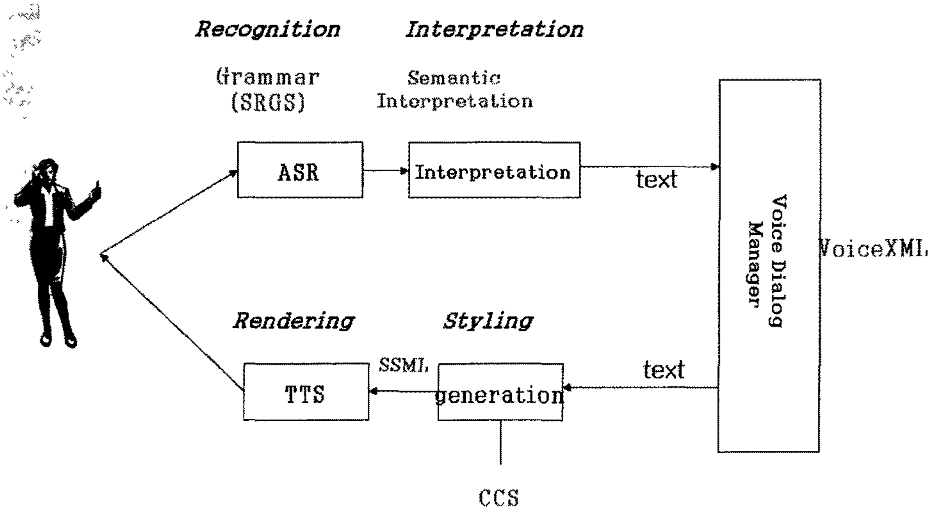
Session component:
state management, and temporary and persistent session supporting

System & Environment :
device capabilities, user preferences and environmental conditions

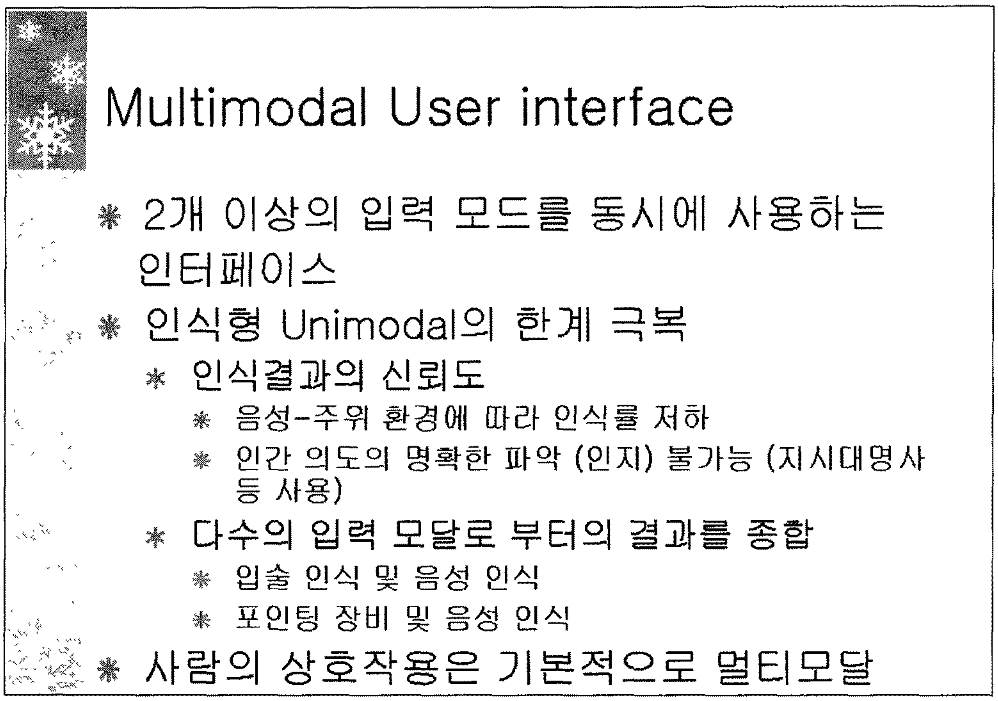
3

Unimodal Interface

- VoiceXML, SSML, SRGS

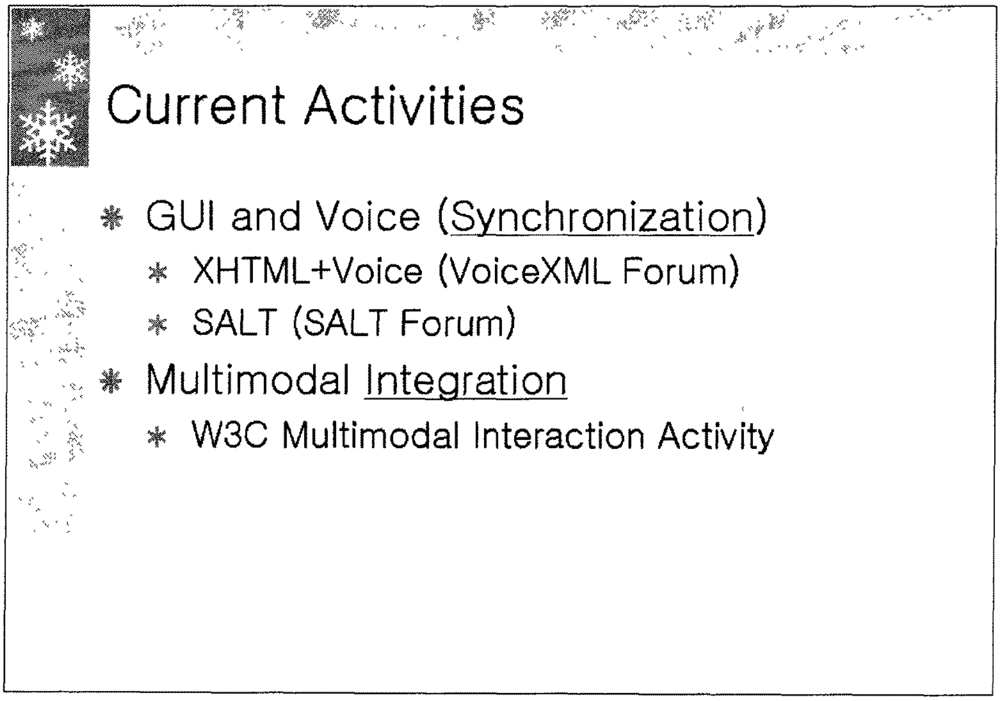


4



Multimodal User interface

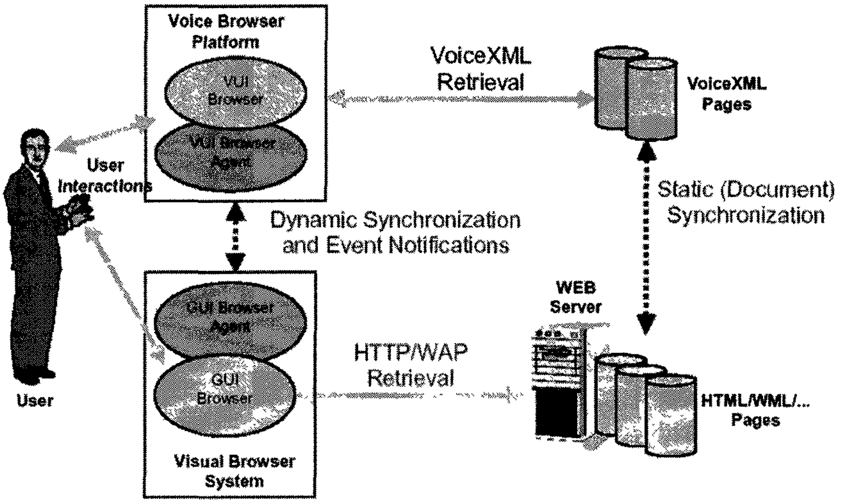
- 2개 이상의 입력 모드를 동시에 사용하는 인터페이스
- 인식형 Unimodal의 한계 극복
  - 인식결과의 신뢰도
    - 음성-주위 환경에 따라 인식률 저하
    - 인간 의도의 명확한 파악 (인지) 불가능 (지시대명사 등 사용)
  - 다수의 입력 모달로 부터의 결과를 종합
    - 입술 인식 및 음성 인식
    - 포인팅 장비 및 음성 인식
- 사람의 상호작용은 기본적으로 멀티모달



Current Activities

- GUI and Voice (Synchronization)
  - XHTML+Voice (VoiceXML Forum)
  - SALT (SALT Forum)
- Multimodal Integration
  - W3C Multimodal Interaction Activity

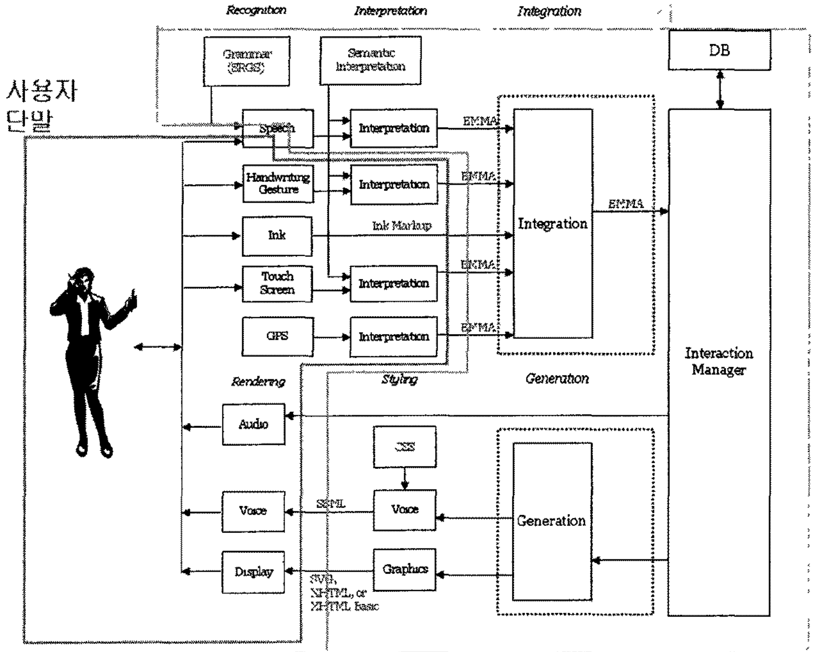
Multi-Modal Access Architecture (X+V and SALT)



7

(Siemans, 2001.11.26)

Multimodal Interface Framework

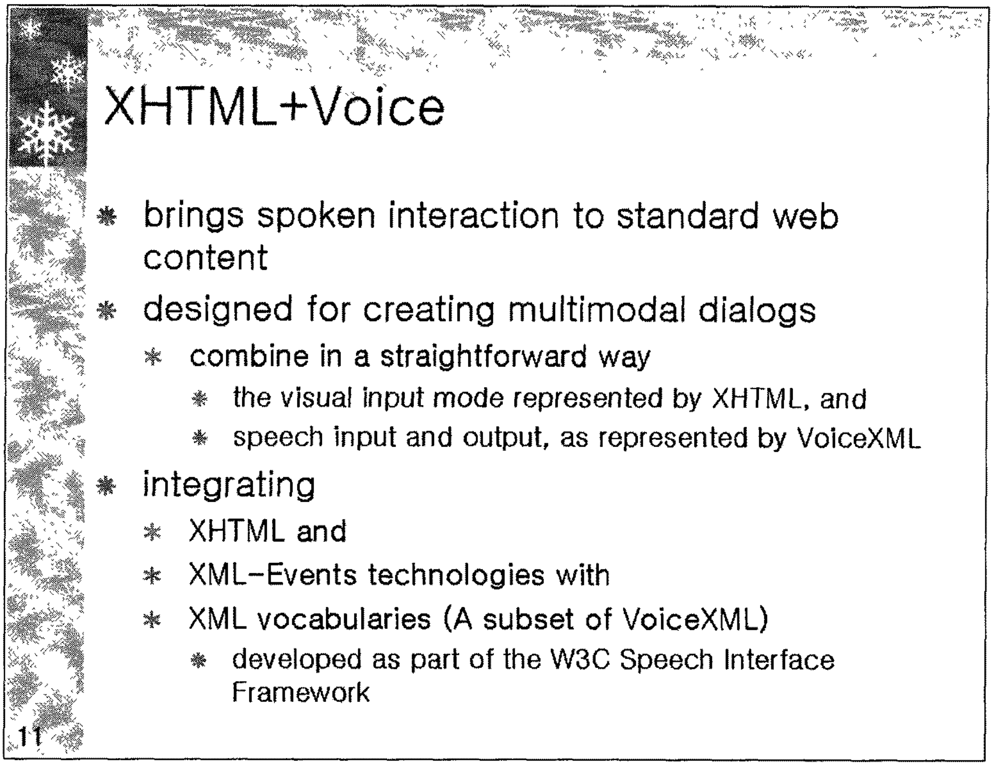


8

X+V : XHTML+Voice

소개

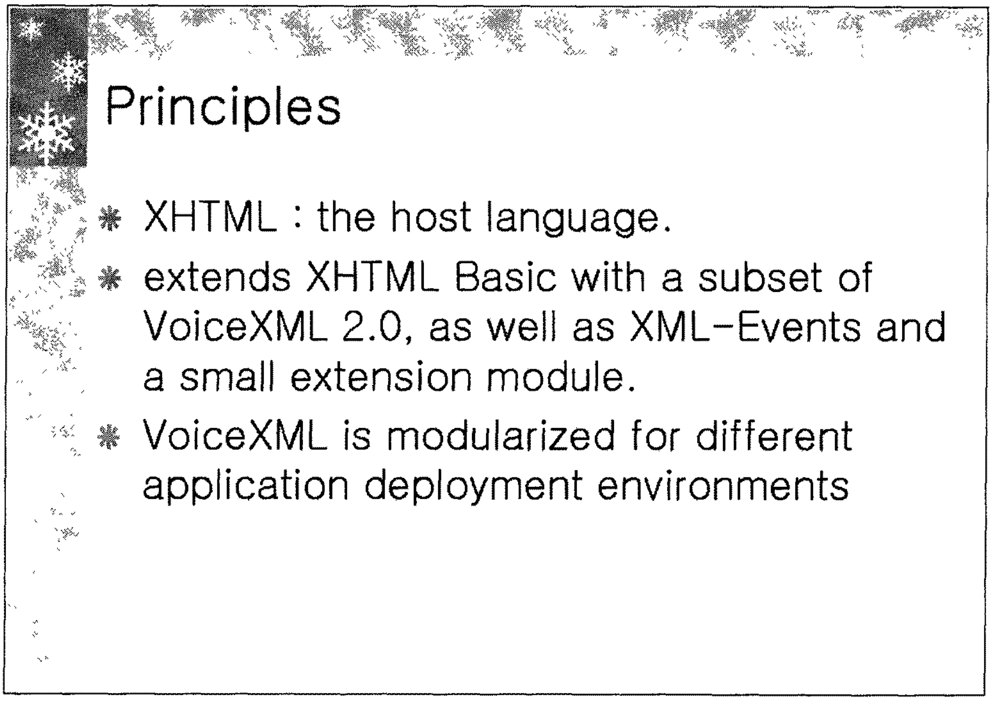
- IBM, Motorola, and Opera Software
  - VoiceXML forum
- XHTML+Voice Profile 1.0
  - http://www.w3.org/TR/xhtml+voice
  - 21 December 2001
- XHTML+Voice Profile 1.1
  - http://www.ibm.com/software/pervasive/multimodal/x+v/11/spec.htm
  - 28 January 2003



XHTML+Voice

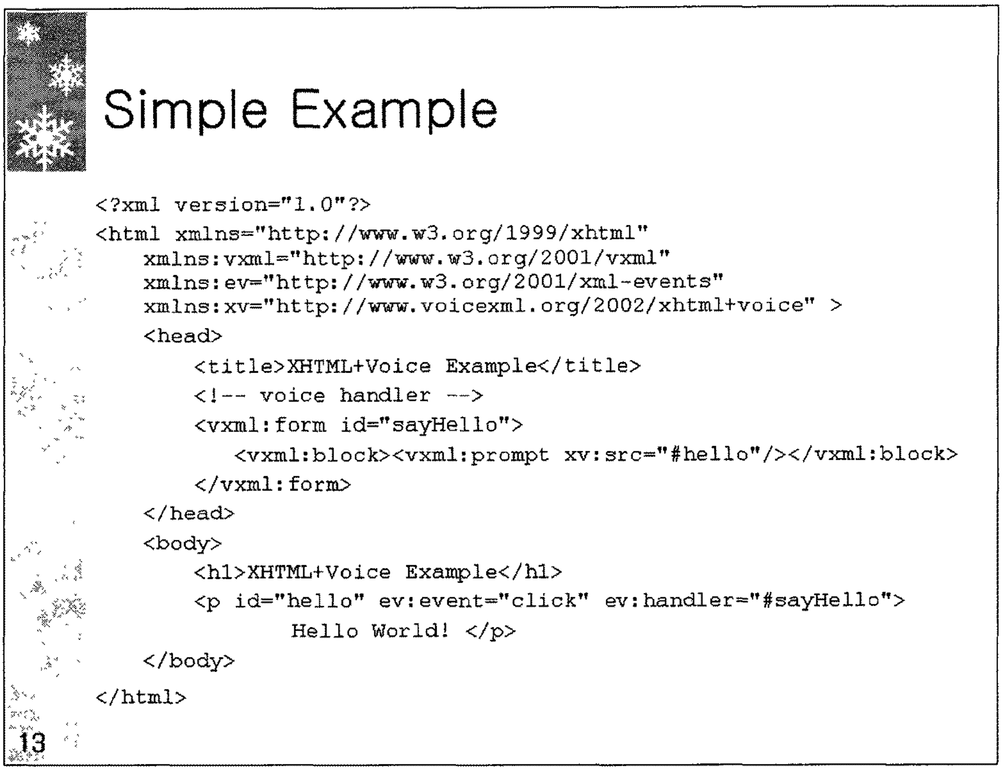
- brings spoken interaction to standard web content
- designed for creating multimodal dialogs
  - combine in a straightforward way
    - the visual input mode represented by XHTML, and
    - speech input and output, as represented by VoiceXML
- integrating
  - XHTML and
  - XML-Events technologies with
  - XML vocabularies (A subset of VoiceXML)
    - developed as part of the W3C Speech Interface Framework

11



Principles

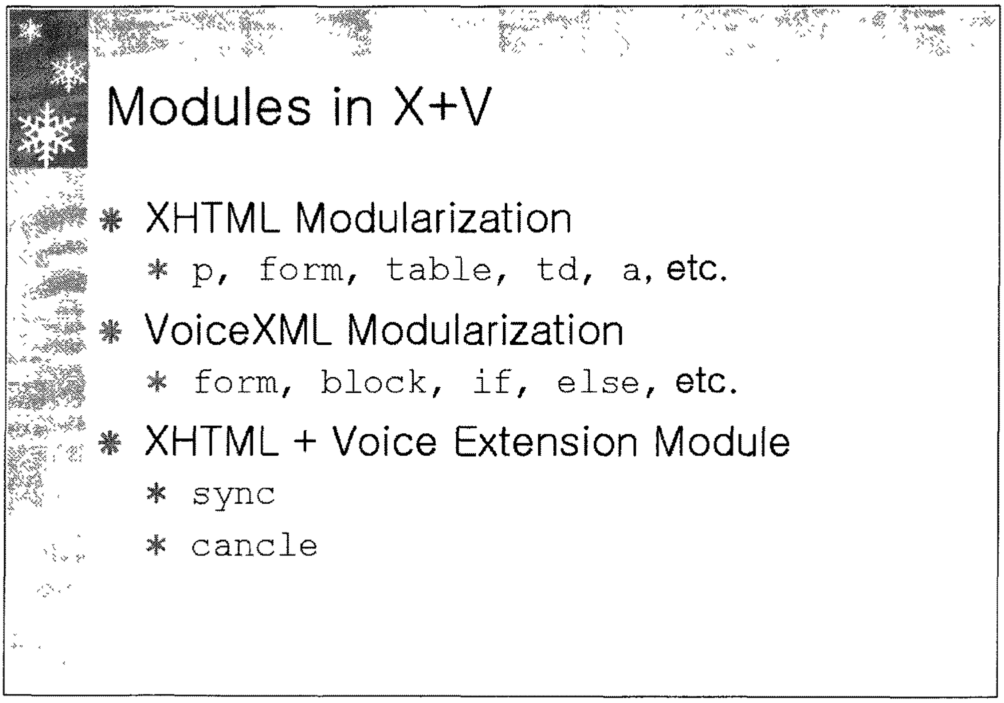
- XHTML : the host language.
- extends XHTML Basic with a subset of VoiceXML 2.0, as well as XML-Events and a small extension module.
- VoiceXML is modularized for different application deployment environments



Simple Example

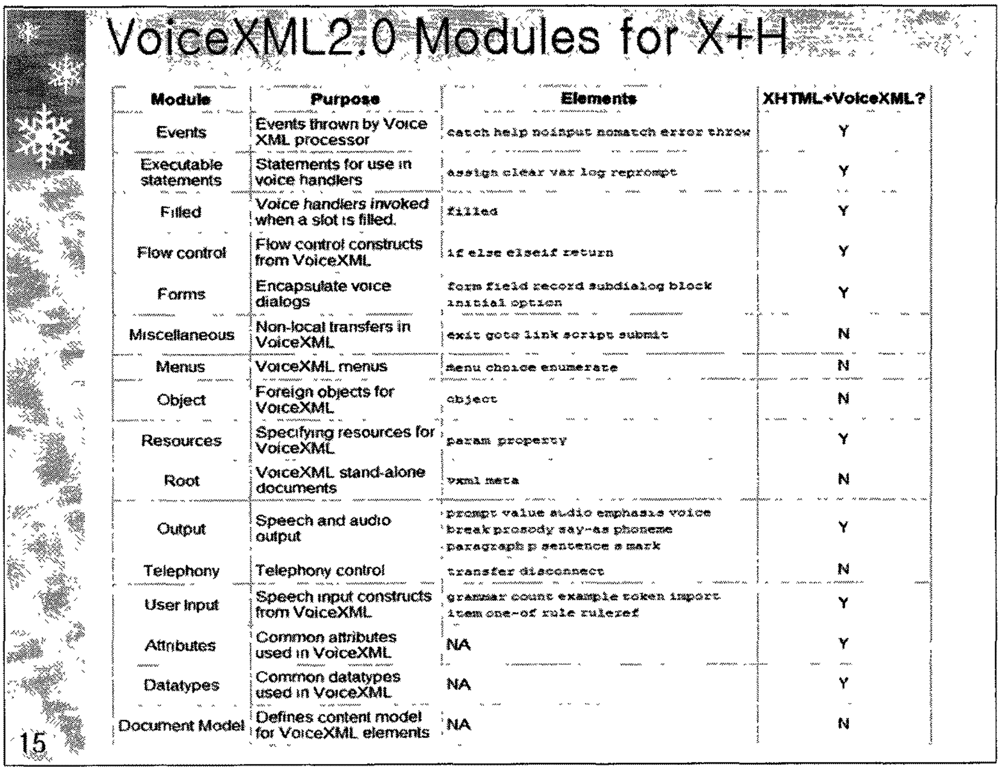
```
<?xml version="1.0"?>
<html xmlns="http://www.w3.org/1999/xhtml"
   xmlns:vxml="http://www.w3.org/2001/vxml"
   xmlns:ev="http://www.w3.org/2001/xml-events"
   xmlns:xv="http://www.voicexml.org/2002/xhtml+voice" >
   <head>
      <title>XHTML+Voice Example</title>
      <!-- voice handler -->
      <vxml:form id="sayHello">
         <vxml:block><vxml:prompt xv:src="#hello"/></vxml:block>
      </vxml:form>
   </head>
   <body>
      <h1>XHTML+Voice Example</h1>
      <p id="hello" ev:event="click" ev:handler="#sayHello">
             Hello World! </p>
   </body>
</html>
```

13



Modules in X+V

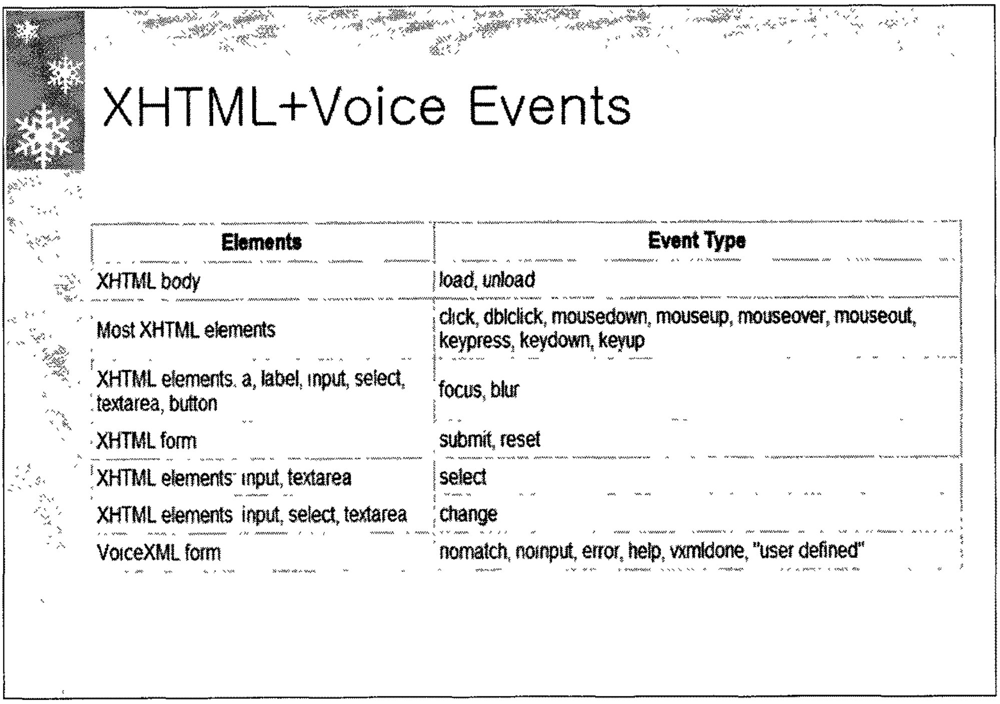
- XHTML Modularization
  - p, form, table, td, a, etc.
- VoiceXML Modularization
  - form, block, if, else, etc.
- XHTML + Voice Extension Module
  - sync
  - cancle



VoiceXML2.0 Modules for X+H

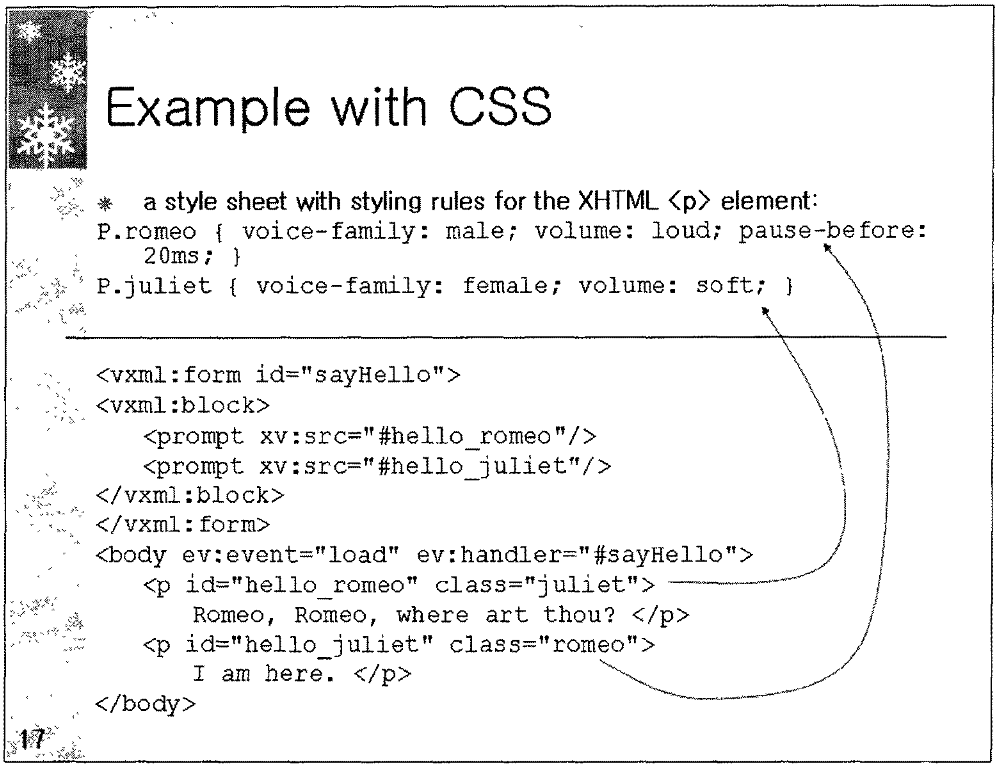
Module	Purpose	Elements	XHTML+VoiceXML?
Events	Events thrown by Voice XML processor	`catch help noinput nomatch error throw`	Y
Executable statements	Statements for use in voice handlers	`assign clear var log reprompt`	Y
Filled	Voice handlers invoked when a slot is filled.	`filled`	Y
Flow control	Flow control constructs from VoiceXML	`if else elseif return`	Y
Forms	Encapsulate voice dialogs	`form field record subdialog block initial option`	Y
Miscellaneous	Non-local transfers in VoiceXML	`exit goto link script submit`	N
Menus	VoiceXML menus	`menu choice enumerate`	N
Object	Foreign objects for VoiceXML	`object`	N
Resources	Specifying resources for VoiceXML	`param property`	Y
Root	VoiceXML stand-alone documents	`vxml meta`	N
Output	Speech and audio output	`prompt value audio emphasis voice break prosody say-as phoneme paragraph p sentence s mark`	Y
Telephony	Telephony control	`transfer disconnect`	N
User Input	Speech input constructs from VoiceXML	`grammar count example token import item one-of rule ruleref`	Y
Attributes	Common attributes used in VoiceXML	NA	Y
Datatypes	Common datatypes used in VoiceXML	NA	Y
Document Model	Defines content model for VoiceXML elements	NA	N

15



XHTML+Voice Events

Elements	Event Type
XHTML body	load, unload
Most XHTML elements	click, dblclick, mousedown, mouseup, mouseover, mouseout, keypress, keydown, keyup
XHTML elements. a, label, input, select, textarea, button	focus, blur
XHTML form	submit, reset
XHTML elements input, textarea	select
XHTML elements input, select, textarea	change
VoiceXML form	nomatch, noinput, error, help, vxmldone, "user defined"



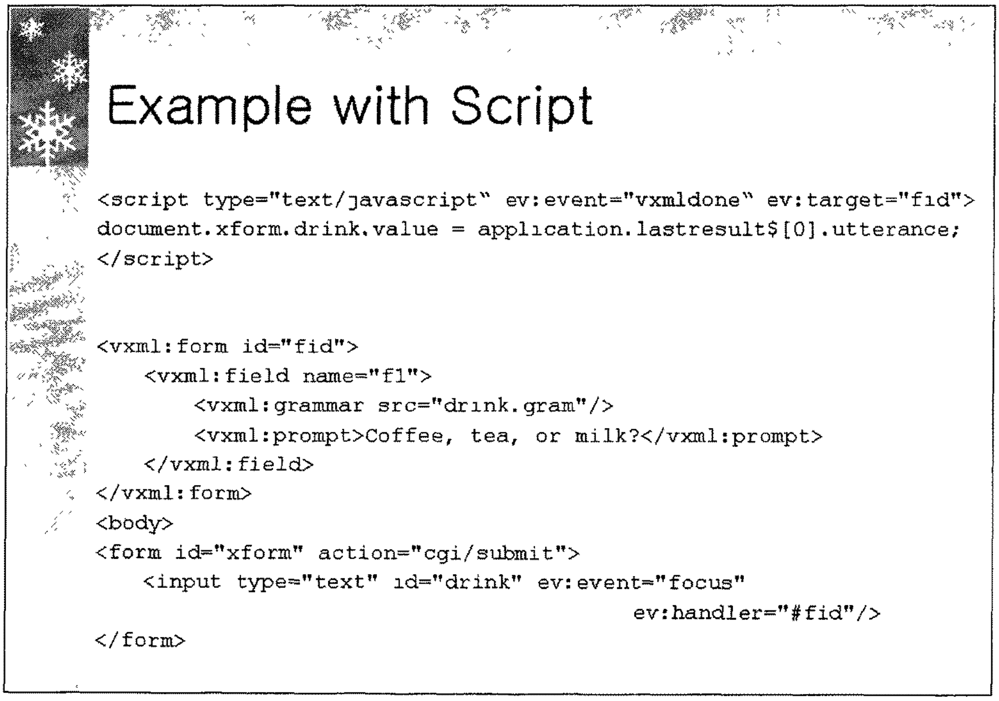
Example with CSS

⁕ a style sheet with styling rules for the XHTML <p> element:

```
P.romeo { voice-family: male; volume: loud; pause-before:
   20ms; }
P.juliet { voice-family: female; volume: soft; }
```

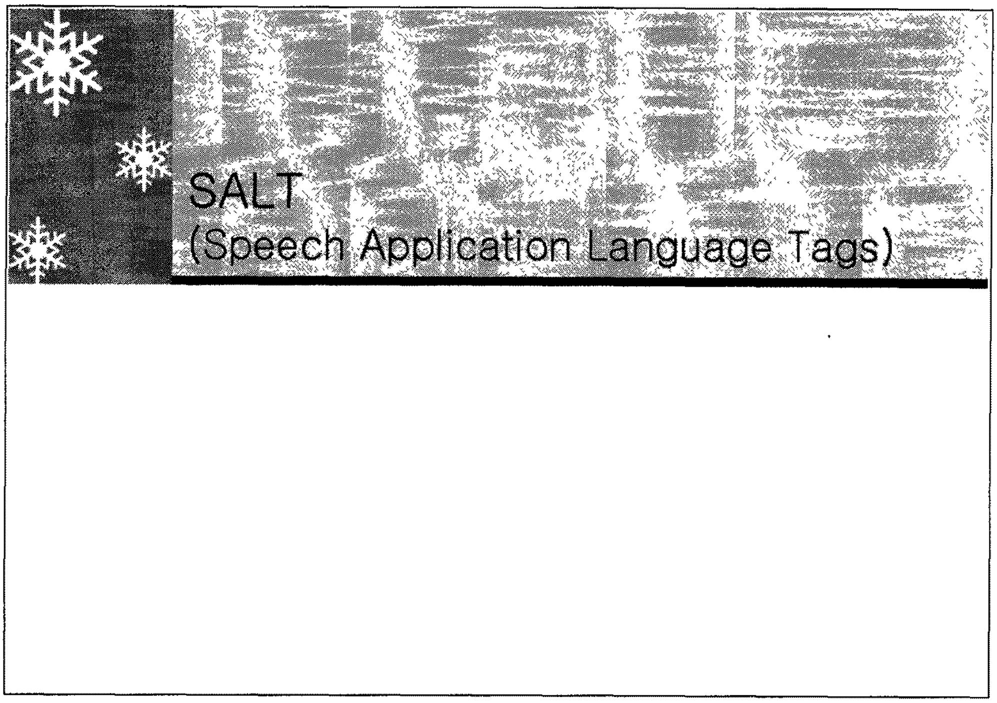
```
<vxml:form id="sayHello">
<vxml:block>
   <prompt xv:src="#hello_romeo"/>
   <prompt xv:src="#hello_juliet"/>
</vxml:block>
</vxml:form>
<body ev:event="load" ev:handler="#sayHello">
   <p id="hello_romeo" class="juliet">
      Romeo, Romeo, where art thou? </p>
   <p id="hello_juliet" class="romeo">
      I am here. </p>
</body>
```

17

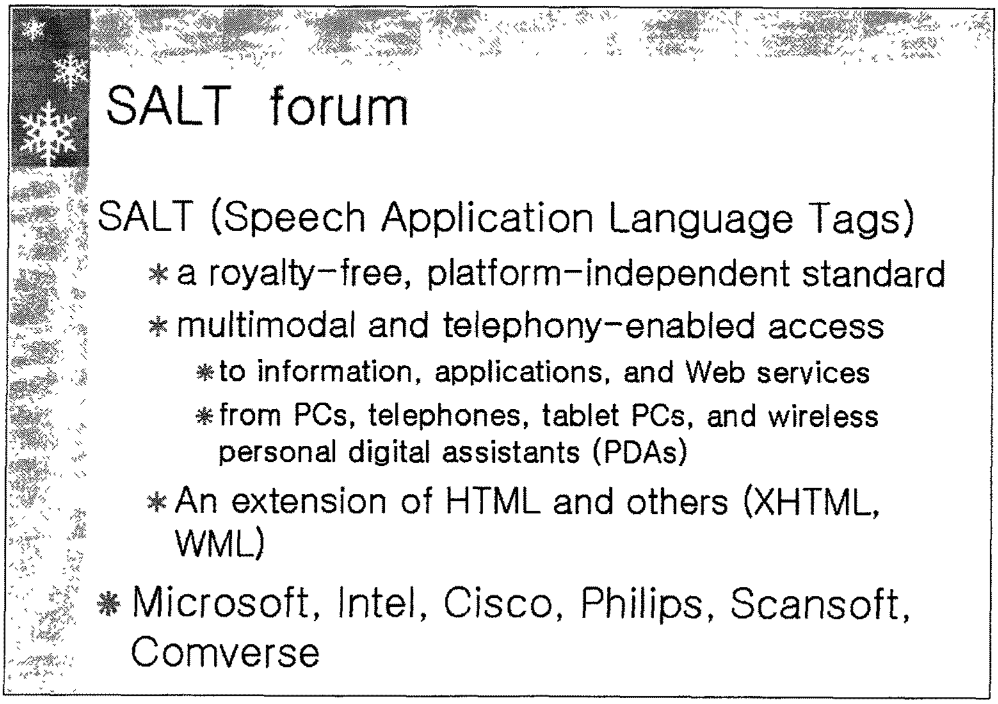


Example with Script

```
<script type="text/javascript" ev:event="vxmldone" ev:target="fid">
document.xform.drink.value = application.lastresult$[0].utterance;
</script>


<vxml:form id="fid">
   <vxml:field name="f1">
      <vxml:grammar src="drink.gram"/>
      <vxml:prompt>Coffee, tea, or milk?</vxml:prompt>
   </vxml:field>
</vxml:form>
<body>
<form id="xform" action="cgi/submit">
   <input type="text" id="drink" ev:event="focus"
                                             ev:handler="#fid"/>
</form>
```



SALT
(Speech Application Language Tags)



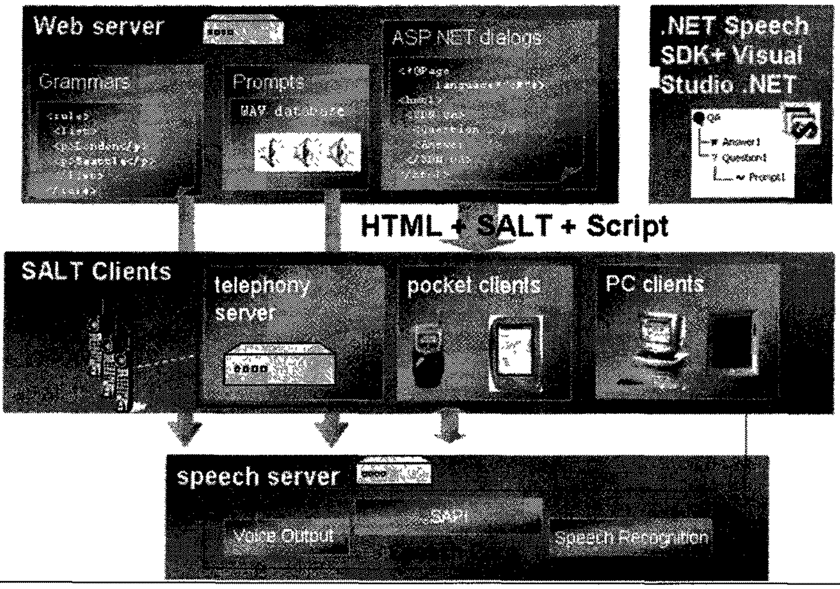
SALT forum

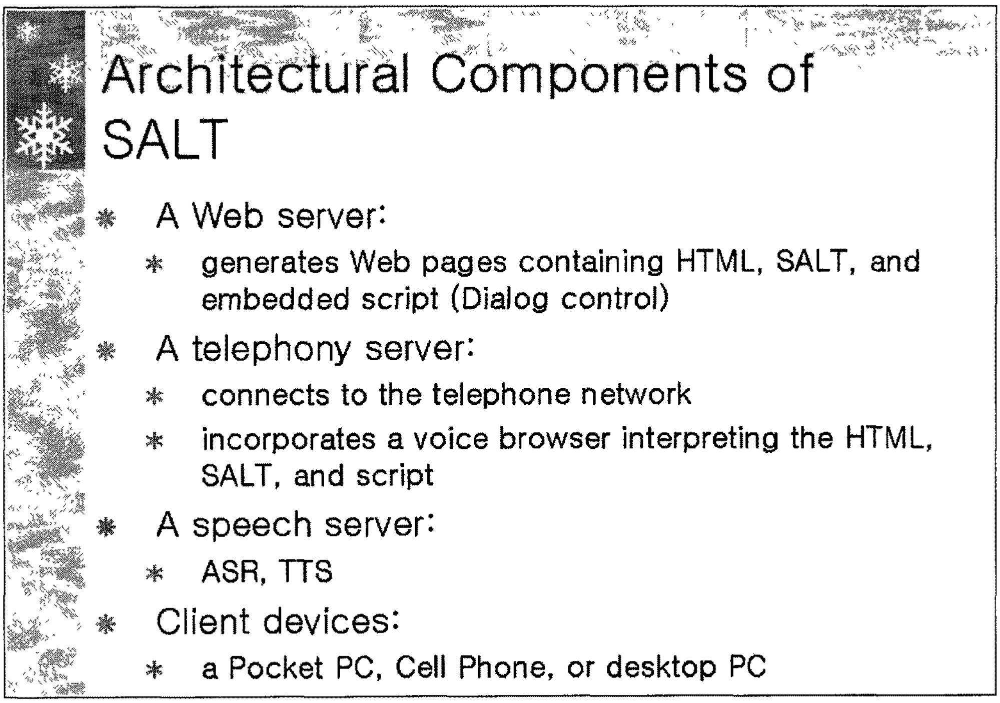
- SALT (Speech Application Language Tags)
  - * a royalty-free, platform-independent standard
  - * multimodal and telephony-enabled access
    - * to information, applications, and Web services
    - * from PCs, telephones, tablet PCs, and wireless personal digital assistants (PDAs)
  - * An extension of HTML and others (XHTML, WML)
- * Microsoft, Intel, Cisco, Philips, Scansoft, Comverse

SALT (Speech Application Language Tags)

- A small set of XML elements
  - Attributes
  - DOM object properties,
  - Events
  - Methods
- An Embedded Markup Language
  - in conjunction with a source markup document (HTML, WML, XHTML)
  - to apply a speech interface to the source page

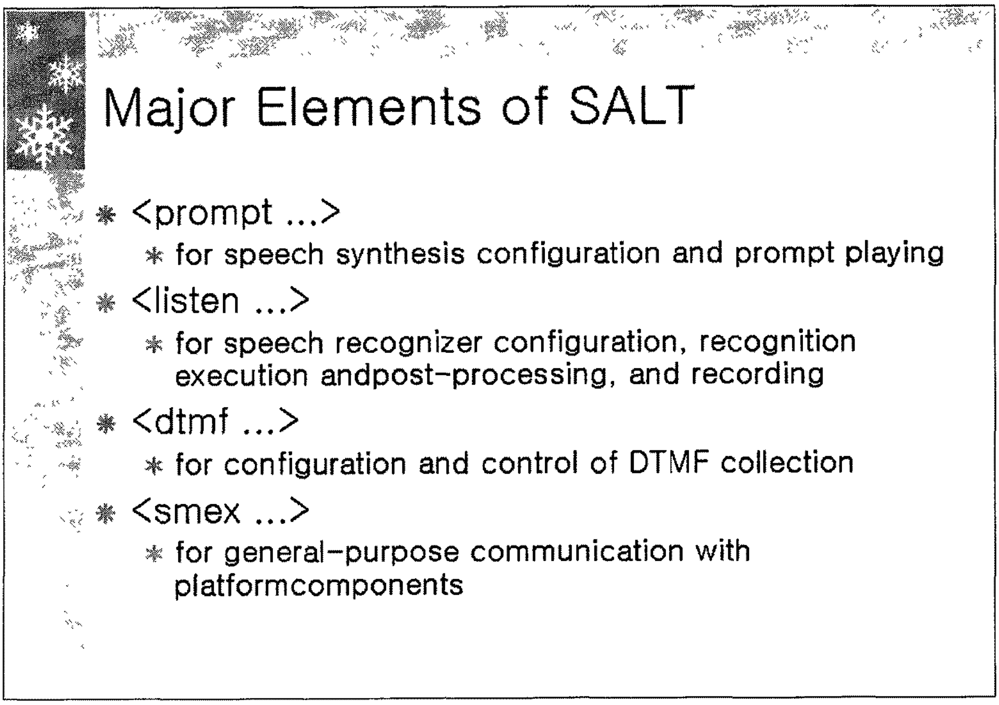
SALT architecture





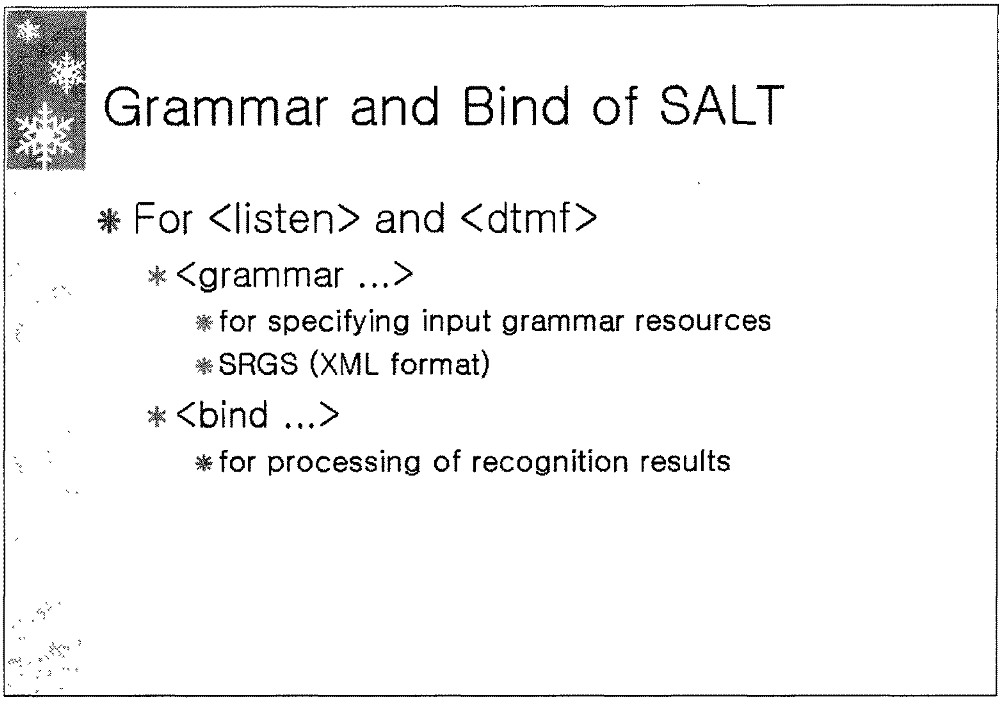
Architectural Components of SALT

- A Web server:
  - generates Web pages containing HTML, SALT, and embedded script (Dialog control)
- A telephony server:
  - connects to the telephone network
  - incorporates a voice browser interpreting the HTML, SALT, and script
- A speech server:
  - ASR, TTS
- Client devices:
  - a Pocket PC, Cell Phone, or desktop PC



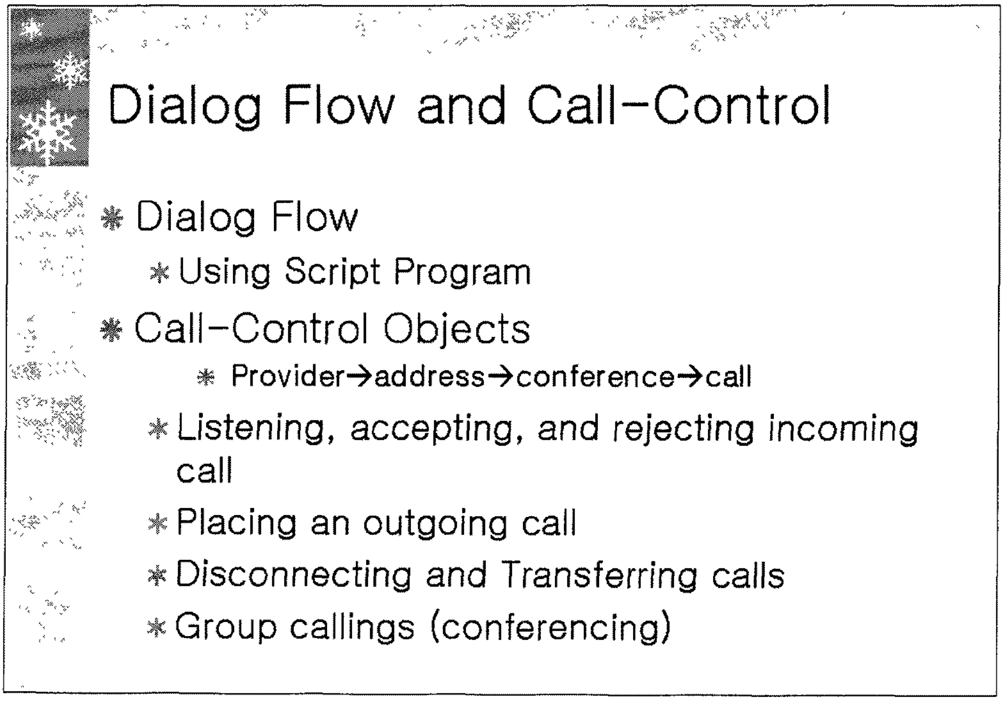
Major Elements of SALT

- <prompt ...>
  - for speech synthesis configuration and prompt playing
- <listen ...>
  - for speech recognizer configuration, recognition execution andpost-processing, and recording
- <dtmf ...>
  - for configuration and control of DTMF collection
- <smex ...>
  - for general-purpose communication with platformcomponents



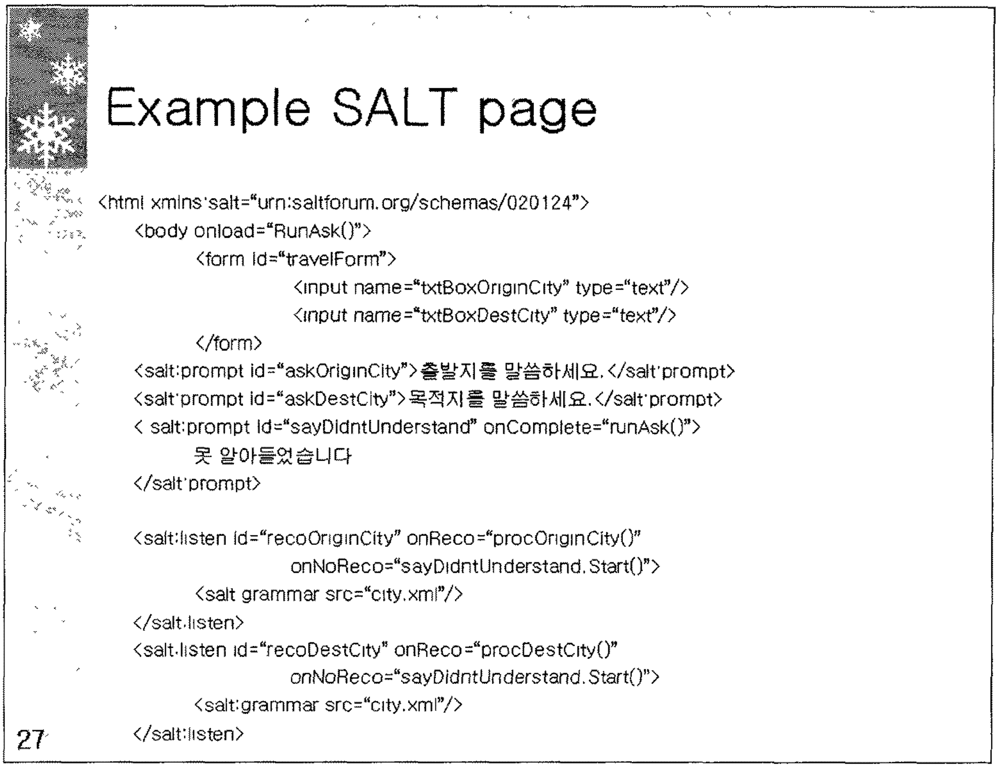
Grammar and Bind of SALT

- For <listen> and <dtmf>
  - <grammar ...>
    - for specifying input grammar resources
    - SRGS (XML format)
  - <bind ...>
    - for processing of recognition results



Dialog Flow and Call-Control

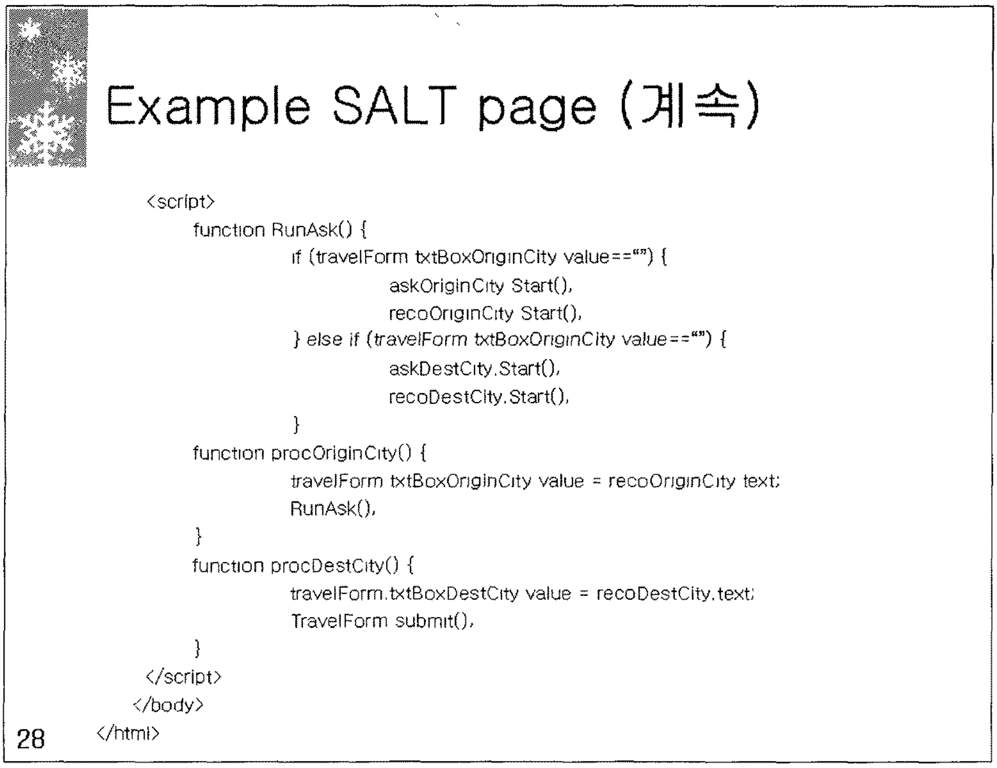
- Dialog Flow
  - Using Script Program
- Call-Control Objects
    - Provider→address→conference→call
  - Listening, accepting, and rejecting incoming call
  - Placing an outgoing call
  - Disconnecting and Transferring calls
  - Group callings (conferencing)



Example SALT page

```
<html xmlns'salt="urn:saltforum.org/schemas/020124">
    <body onload="RunAsk()">
        <form id="travelForm">
                <input name="txtBoxOriginCity" type="text"/>
                <input name="txtBoxDestCity" type="text"/>
        </form>
    <salt:prompt id="askOriginCity">출발지를 말씀하세요.</salt'prompt>
    <salt'prompt id="askDestCity">목적지를 말씀하세요.</salt'prompt>
    < salt:prompt id="sayDidntUnderstand" onComplete="runAsk()">
        못 알아들었습니다
    </salt'prompt>

    <salt:listen id="recoOriginCity" onReco="procOriginCity()"
                onNoReco="sayDidntUnderstand.Start()">
        <salt grammar src="city.xml"/>
    </salt.listen>
    <salt.listen id="recoDestCity" onReco="procDestCity()"
                onNoReco="sayDidntUnderstand.Start()">
        <salt:grammar src="city.xml"/>
    </salt:listen>
```

27



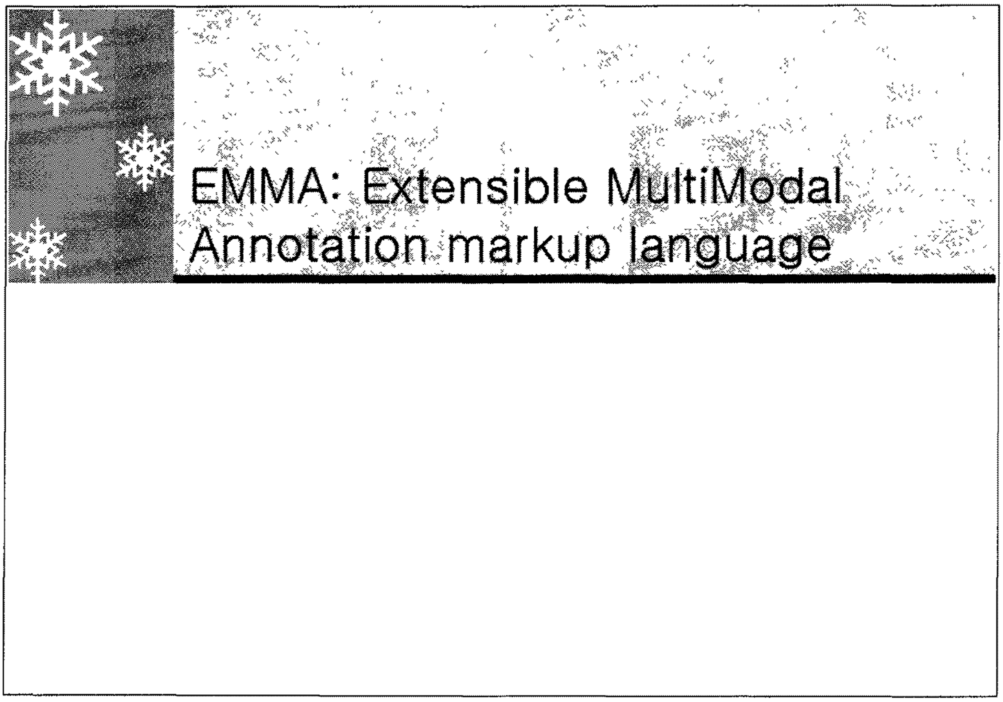
Example SALT page (계속)

```
     <script>
          function RunAsk() {
                    if (travelForm txtBoxOriginCity value=="") {
                              askOriginCity Start(),
                              recoOriginCity Start(),
                    } else if (travelForm txtBoxOriginCity value=="") {
                              askDestCity.Start(),
                              recoDestCity.Start(),
                    }
          function procOriginCity() {
                    travelForm txtBoxOriginCity value = recoOriginCity text;
                    RunAsk(),
          }
          function procDestCity() {
                    travelForm.txtBoxDestCity value = recoDestCity.text;
                    TravelForm submit(),
          }
     </script>
    </body>
</html>
```

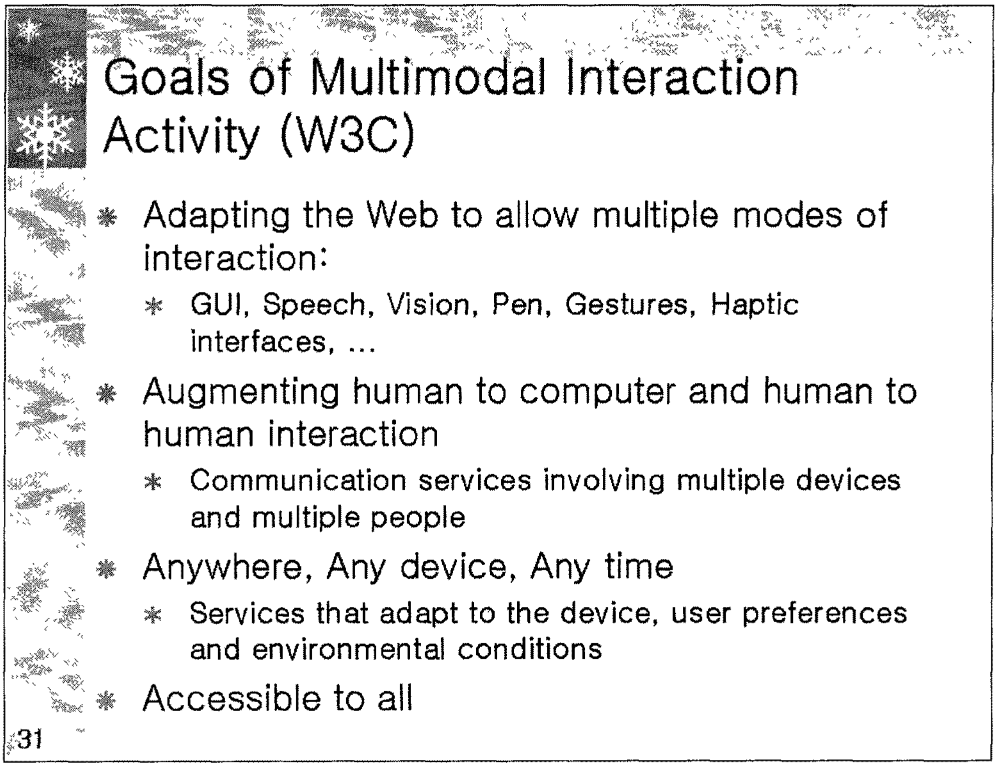
28

현재 상태

- SALT 1.0 Spec.
  - July 2002
- W3c의 Multimodal Activity에 제출



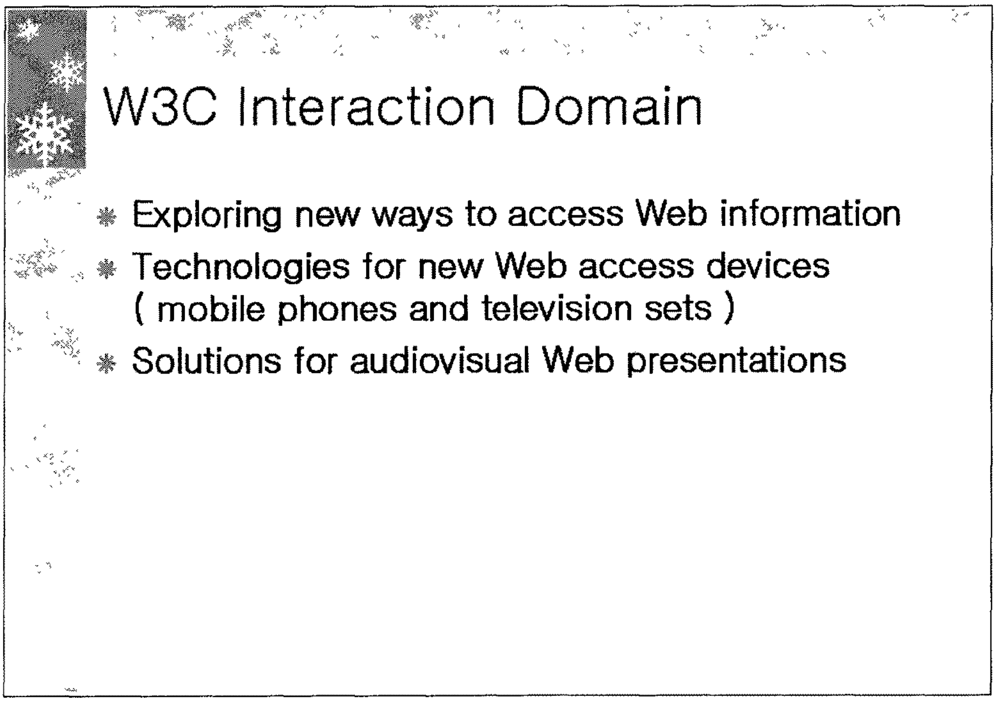
EMMA: Extensible MultiModal Annotation markup language



Goals of Multimodal Interaction Activity (W3C)

- Adapting the Web to allow multiple modes of interaction:
  - GUI, Speech, Vision, Pen, Gestures, Haptic interfaces, ...
- Augmenting human to computer and human to human interaction
  - Communication services involving multiple devices and multiple people
- Anywhere, Any device, Any time
  - Services that adapt to the device, user preferences and environmental conditions
- Accessible to all

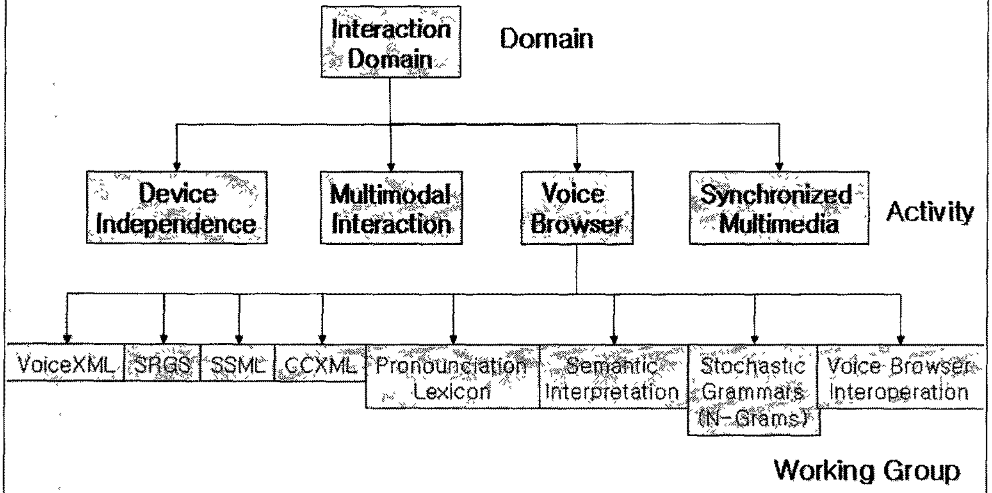
31



W3C Interaction Domain

- Exploring new ways to access Web information
- Technologies for new Web access devices (mobile phones and television sets)
- Solutions for audiovisual Web presentations

Interaction Domain



Interaction Domain

Domain

Device Independence

Multimodal Interaction

Voice Browser

Synchronized Multimedia

Activity

VoiceXML

SRGS

SSML

CCXML

Pronounciation Lexicon

Semantic Interpretation

Stochastic Grammars (N-Grams)

Voice Browser Interoperation

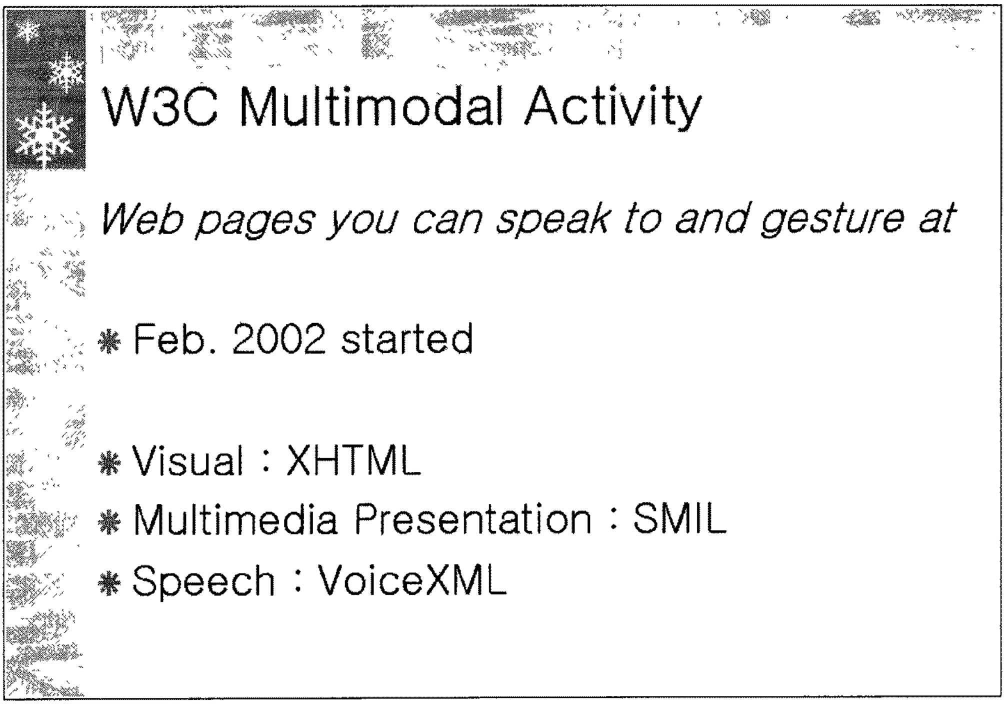
Working Group

33

Domain Activities

- **Device Independence**
  - a seamless Web for all access devices
  - from desktop PCs to in-car computers, TV, digital cameras, and cellular phones
- **Multimodal Interaction**
  - extending the Web user interface to allow multiple modes of interaction
  - Input: Voice, key pad, keyboard, mouse, stylus or other input device
  - Output: spoken prompts and audio, and to view information on graphical displays
- **Synchronized Multimedia**
  - design of a language for scheduling multimedia presentations
  - Synchronized Multimedia Integration Language (SMIL)
- **Voice Browser**

34

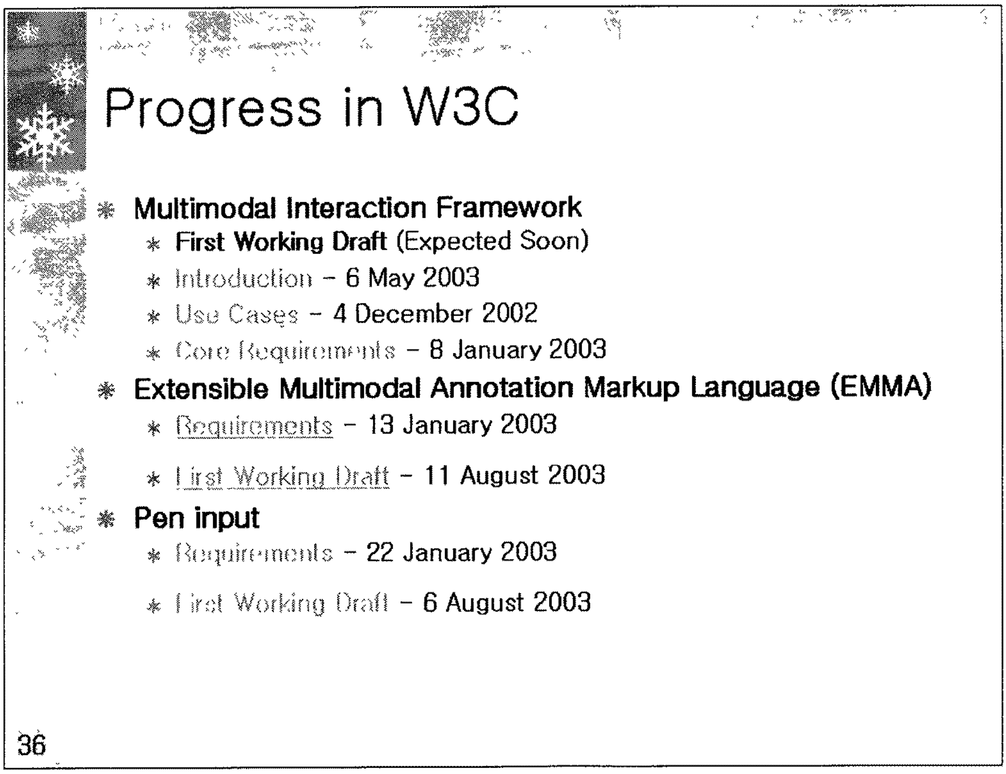


W3C Multimodal Activity

Web pages you can speak to and gesture at

- ❋ Feb. 2002 started

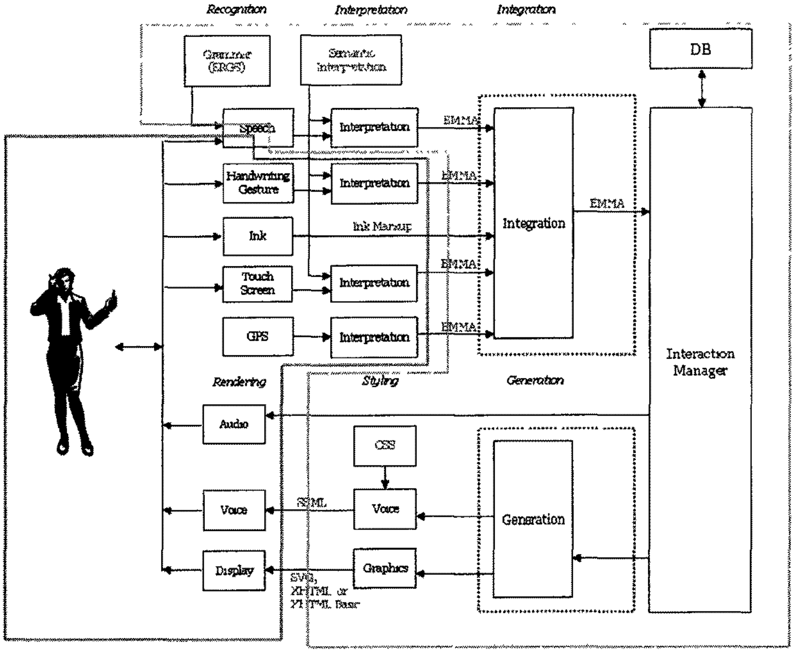
- ❋ Visual : XHTML
- ❋ Multimedia Presentation : SMIL
- ❋ Speech : VoiceXML



Progress in W3C

- ❋ **Multimodal Interaction Framework**
  - ⁎ **First Working Draft** (Expected Soon)
  - ⁎ Introduction – 6 May 2003
  - ⁎ Use Cases – 4 December 2002
  - ⁎ Core Requirements – 8 January 2003
- ❋ **Extensible Multimodal Annotation Markup Language (EMMA)**
  - ⁎ Requirements – 13 January 2003
  - ⁎ First Working Draft – 11 August 2003
- ❋ **Pen input**
  - ⁎ Requirements – 22 January 2003
  - ⁎ First Working Draft – 6 August 2003

36

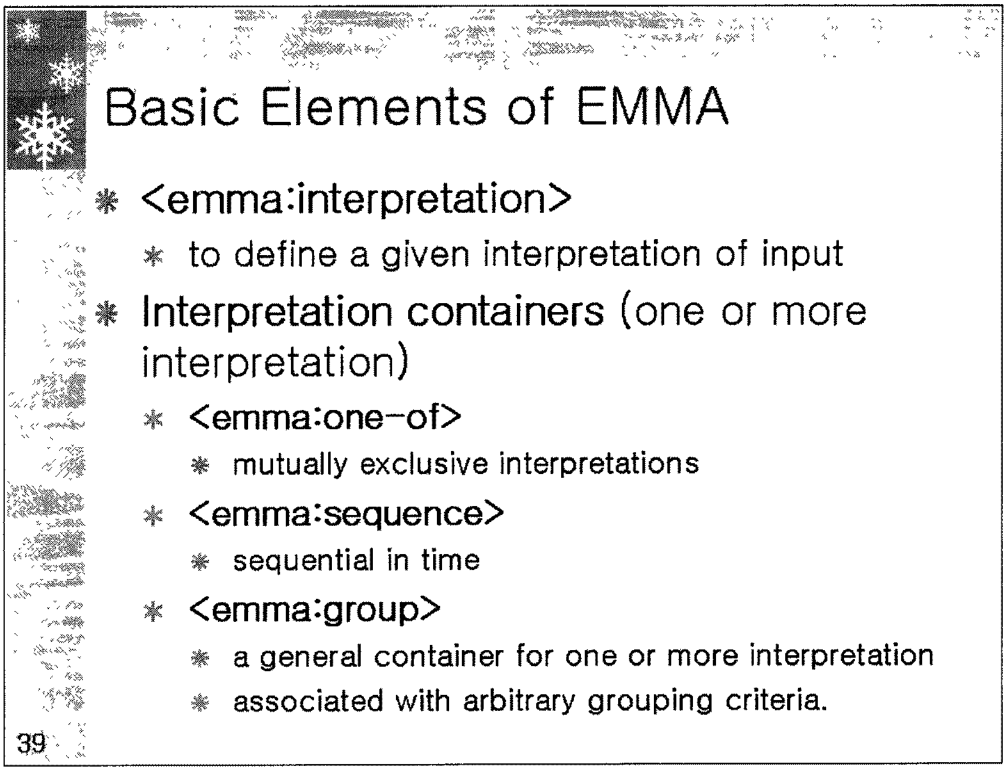

Recognition
Interpretation
Integration
Grammar (SRGS)
Semantic Interpretation
DB
Speech
Interpretation
EMMA
Handwriting Gesture
Interpretation
EMMA
Ink
Ink Markup
EMMA
Integration
Touch Screen
Interpretation
EMMA
GPS
Interpretation
EMMA
Interaction Manager
Rendering
Styling
Generation
Audio
CSS
Voice
SSML
Voice
Generation
Display
Graphics
SVG, XHTML or XHTML Basic

37

EMMA

- An Annotation for
  - (Multimodal) User input
  - the result of (Multimodal) recognitions
- Producer
  - Speech recognizers
  - Handwriting recognizers
  - Natural language understanding engines
  - Other input media interpreters (e.g. DTMF, pointing, keyboard)
  - Multimodal integration component
- Consumer:
  - Interaction manager
  - Multimodal integration component

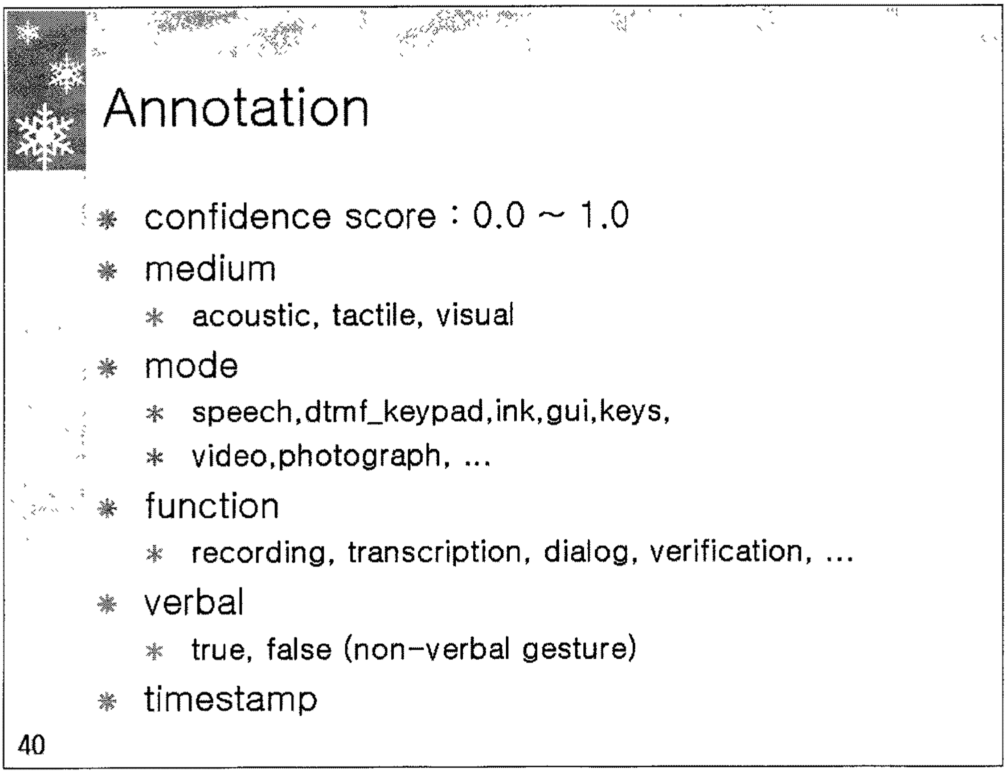
38



Basic Elements of EMMA

- <emma:interpretation>
  - to define a given interpretation of input
- Interpretation containers (one or more interpretation)
  - <emma:one-of>
    - mutually exclusive interpretations
  - <emma:sequence>
    - sequential in time
  - <emma:group>
    - a general container for one or more interpretation
    - associated with arbitrary grouping criteria.

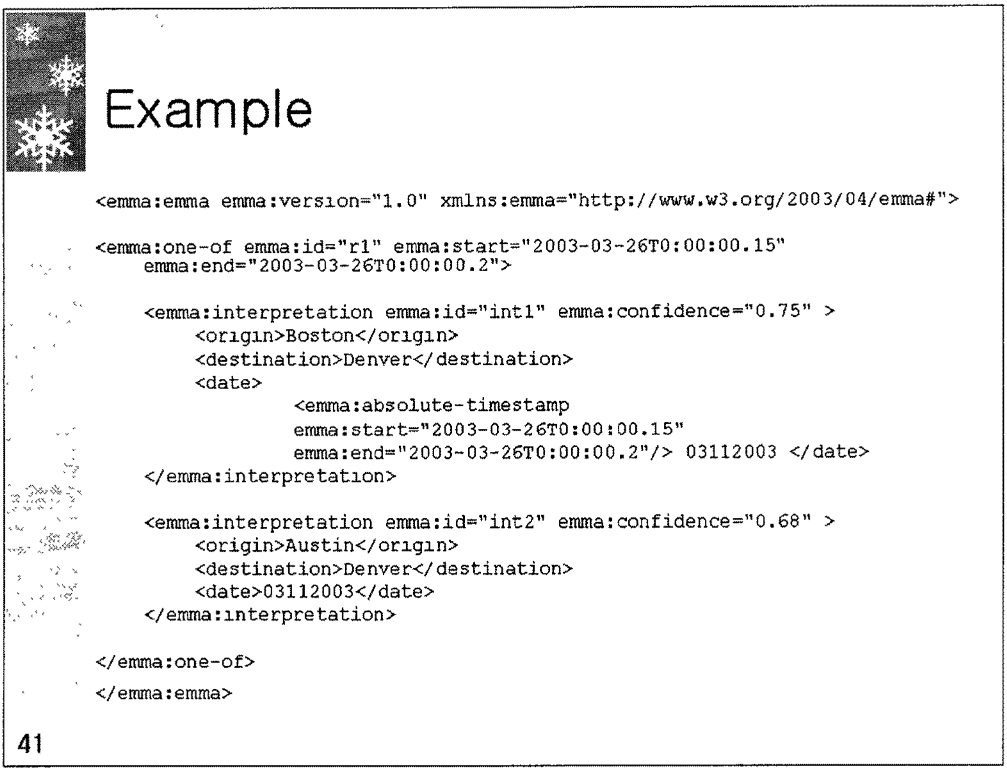
39



Annotation

- confidence score : 0.0 ~ 1.0
- medium
  - acoustic, tactile, visual
- mode
  - speech,dtmf_keypad,ink,gui,keys,
  - video,photograph, ...
- function
  - recording, transcription, dialog, verification, ...
- verbal
  - true, false (non-verbal gesture)
- timestamp

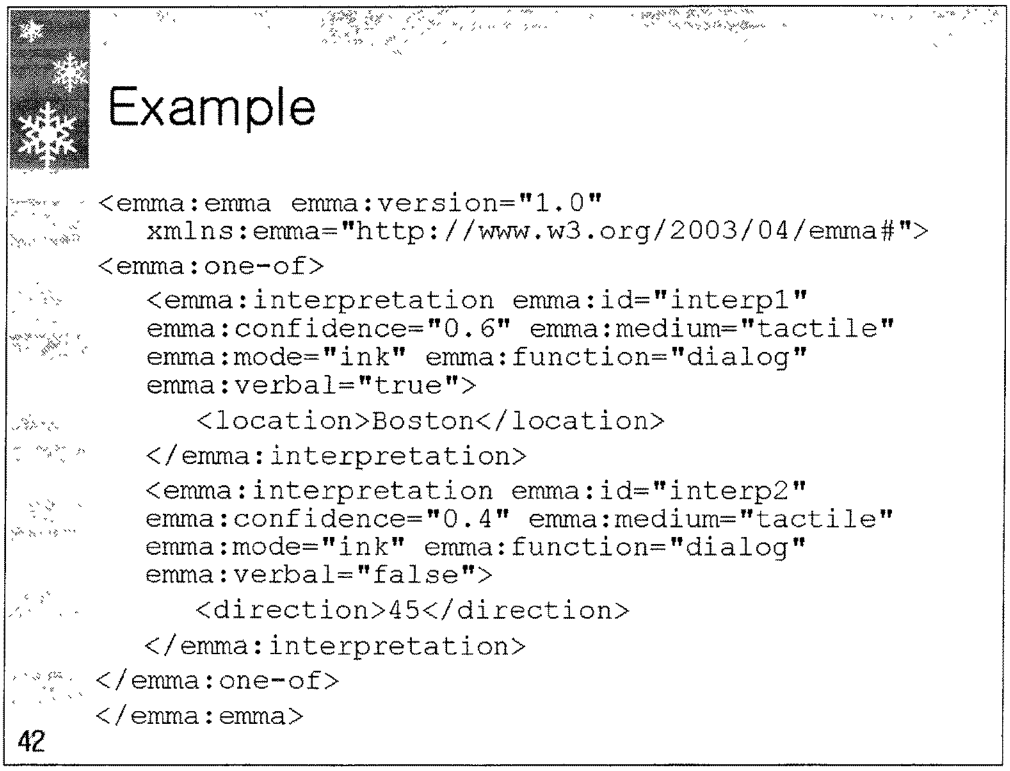
40



Example

```
<emma:emma emma:version="1.0" xmlns:emma="http://www.w3.org/2003/04/emma#">

<emma:one-of emma:id="r1" emma:start="2003-03-26T0:00:00.15"
    emma:end="2003-03-26T0:00:00.2">

    <emma:interpretation emma:id="int1" emma:confidence="0.75" >
        <origin>Boston</origin>
        <destination>Denver</destination>
        <date>
                <emma:absolute-timestamp
                emma:start="2003-03-26T0:00:00.15"
                emma:end="2003-03-26T0:00:00.2"/> 03112003 </date>
    </emma:interpretation>

    <emma:interpretation emma:id="int2" emma:confidence="0.68" >
        <origin>Austin</origin>
        <destination>Denver</destination>
        <date>03112003</date>
    </emma:interpretation>

</emma:one-of>
</emma:emma>
```

41

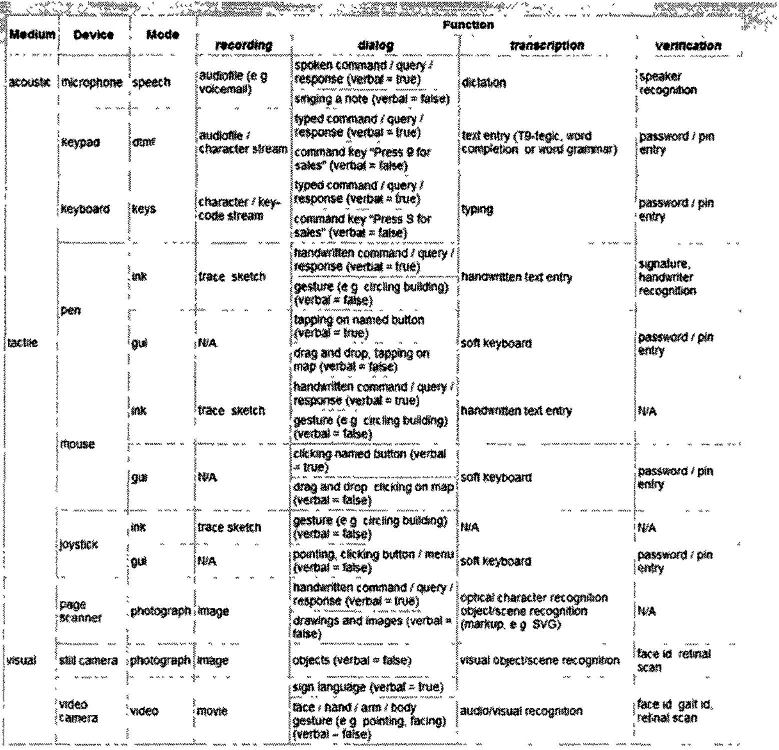


Example

```
<emma:emma emma:version="1.0"
   xmlns:emma="http://www.w3.org/2003/04/emma#">
<emma:one-of>
   <emma:interpretation emma:id="interp1"
   emma:confidence="0.6" emma:medium="tactile"
   emma:mode="ink" emma:function="dialog"
   emma:verbal="true">
      <location>Boston</location>
   </emma:interpretation>
   <emma:interpretation emma:id="interp2"
   emma:confidence="0.4" emma:medium="tactile"
   emma:mode="ink" emma:function="dialog"
   emma:verbal="false">
      <direction>45</direction>
   </emma:interpretation>
</emma:one-of>
</emma:emma>
```

42

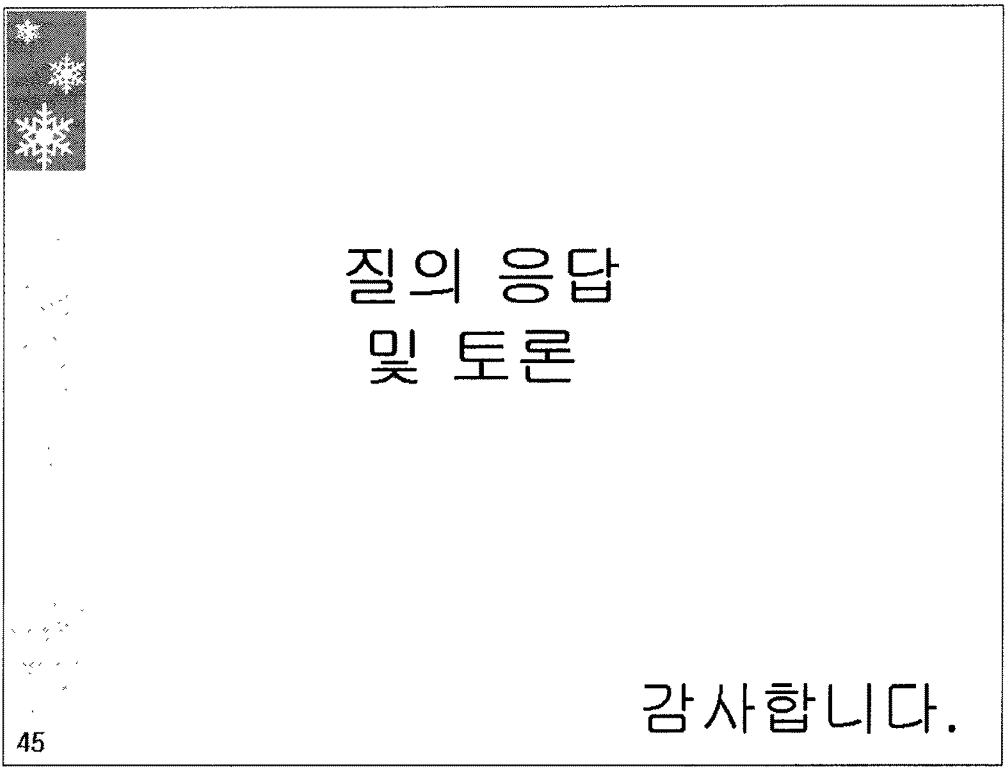
EMMA Modes



Medium	Device	Mode	Function			
			recording	*dialog*	*transcription*	*verification*
acoustic	microphone	speech	audiofile (e g voicemail)	spoken command / query / response (verbal = true) singing a note (verbal = false)	dictation	speaker recognition
	keypad	dtmf	audiofile / character stream	typed command / query / response (verbal = true) command key "Press 9 for sales" (verbal = false)	text entry (T9-tegic, word completion or word grammar)	password / pin entry
	keyboard	keys	character / key-code stream	typed command / query / response (verbal = true) command key "Press S for sales" (verbal = false)	typing	password / pin entry
tactile	pen	ink	trace sketch	handwritten command / query / response (verbal = true) gesture (e g circling building) (verbal = false)	handwritten text entry	signature, handwriter recognition
		gui	N/A	tapping on named button (verbal = true) drag and drop, tapping on map (verbal = false)	soft keyboard	password / pin entry
	mouse	ink	trace sketch	handwritten command / query / response (verbal = true) gesture (e g circling building) (verbal = false)	handwritten text entry	N/A
		gui	N/A	clicking named button (verbal = true) drag and drop clicking on map (verbal = false)	soft keyboard	password / pin entry
	joystick	ink	trace sketch	gesture (e g circling building) (verbal = false)	N/A	N/A
		gui	N/A	pointing, clicking button / menu (verbal = false)	soft keyboard	password / pin entry
	page scanner	photograph	image	handwritten command / query / response (verbal = true) drawings and images (verbal = false)	optical character recognition object/scene recognition (markup, e g SVG)	N/A
visual	still camera	photograph	image	objects (verbal = false)	visual object/scene recognition	face id retinal scan
	video camera	video	movie	sign language (verbal = true) face / hand / arm / body gesture (e g pointing, facing) (verbal = false)	audio/visual recognition	face id gait id, retinal scan

References

- W3C Interaction Domain
  - http://www.w3.org/Interaction/
- W3C Voice Browser Activity
  - http://www.w3.org/Voice/
- W3C Multimodal Interaction Activity
  - http://www.w3.org/2002/mmi/
- VoiceXML forum
  - http://www.voicexml.org/
- SALT forum
  - http://www.saltforum.org/



질의 응답
및 토론

감사합니다.

45