Sistemi Multimodali: scenari, architetture, tecnologie

Stefano Puglia
stefano.puglia@w-lab.it
puglia@dis.uniroma1.it
Roma, 01/06/2005

Università di Roma Tre – Dipartimento di Informatica ed Automazione

Sommario

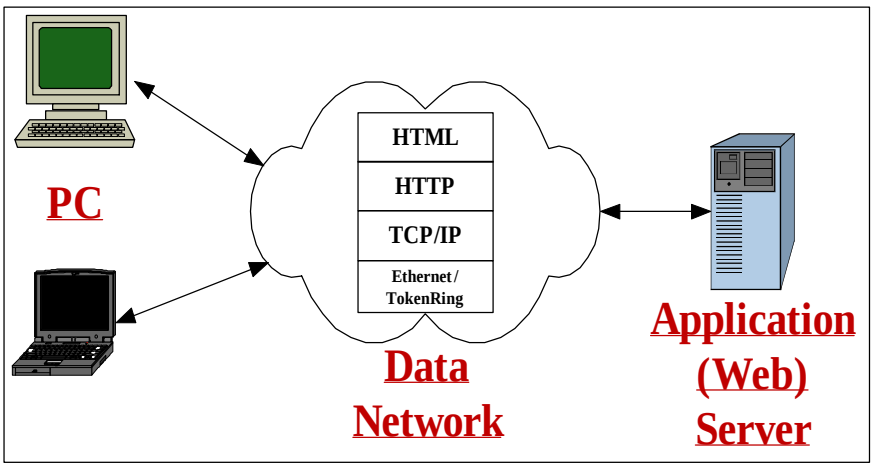
- Motivazioni
  - cosa, perché ed in quali contesti
- Multicanalità
  - caratteristiche e limiti
- Multimodalità
  - architetture, sincronizzazione e scenari
- Tecnologie per la multimodalità
  - linguaggi

Un obiettivo dichiarato…([MMI])

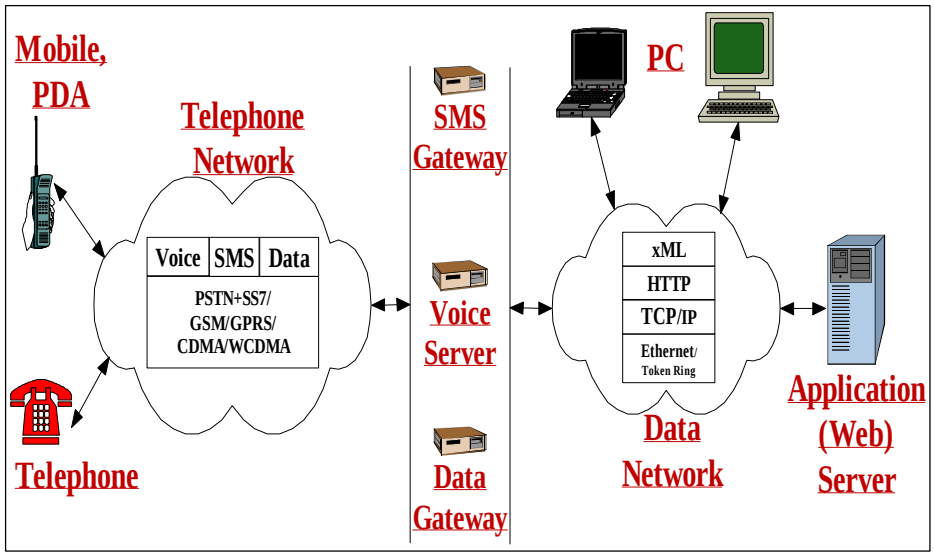
"Multimodal applications should be able to adapt to changing device capabilities, user preferences and environmental conditions."

W3C – MMI (MultiModal Interaction Activity Group)

Da uno scenario tradizionale…



…ad uno più evoluto!



Alcune nuove necessità

- **Ampliare i contesti d'uso** (supportando la crescente richiesta di accesso fisso e mobile, in diversi "ambienti d'uso"- vedi anche [DI])
- Garantire una **continuità d'interazione** (senza dover riprendere dall'inizio nella migrazione da un dispositivo – modo – ad un altro)
- **Ed inoltre:**
  - **Ampliare il parco d'utenza** (in relazione alle abilità, alle competenze fisico-cognitive, al livello di esperienza)
  - **Incrementare la "robustezza"** dell'interazione (maggiore "fault tolerance" attraverso ridondanze e complementarietà)
  - Indirizzare questioni di **privatezza** e **sicurezza** (combinando più forme sensoriali d'interazione)

Alcuni contesti d'impiego

- **Applicazioni desktop avanzate** (es. sistemi di dettatura, interfacce multisensoriali audio-visuali-tattili)
- **Pervasive and ubiquitous mobile computing**
- **Sistemi di sicurezza** (es. sistemi M3 – Multibiometric, Multimodal, Multisensor)
- **Sistemi "assistivi"** di supporto **all'accessibilità** ed **all'autonomia** di utenti disabili e di categorie deboli o svantaggiate ([ETSI], [WAI])

Tre definizioni

- **Modo interattivo (Modo):** canale hw/sw associato ad un senso umano per interagire con sistemi automatici e/o applicazioni software (es. dialogo audio-vocale, interfaccia visuale, dispositivo di puntamento e tattile, gesto manuale, sguardo, movimento del corpo)
- **Sistema multicanale (SMC):** sistema hw/sw che consente la ricezione, l'interpretazione, il processamento (in input) e la generazione (in output) di due o più modi interattivi in maniera indipendente
- **Sistema multimodale (SMM):** sistema hw/sw che consente la ricezione, l'interpretazione, il processamento (in input) e la generazione (in output) di due o più modi interattivi in maniera integrata e coordinata

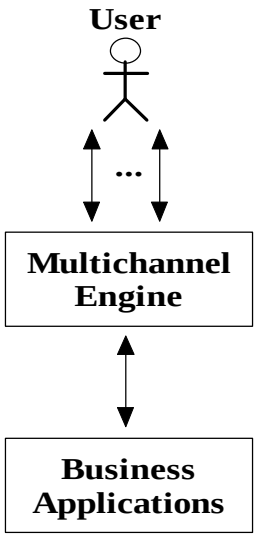
La multicanalità

Un altro obiettivo…([MWI])

"Making Web access from a mobile device as simple, easy and convenient as Web access from a desktop device."

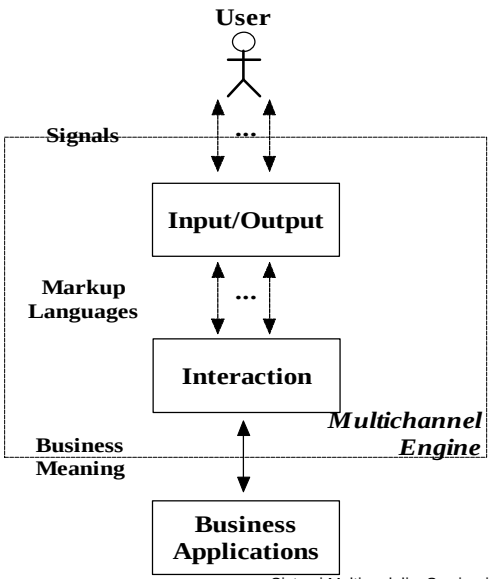
W3C – MWI (Mobile Web Initiative)

Architettura multicanale (1/3) [MAS]



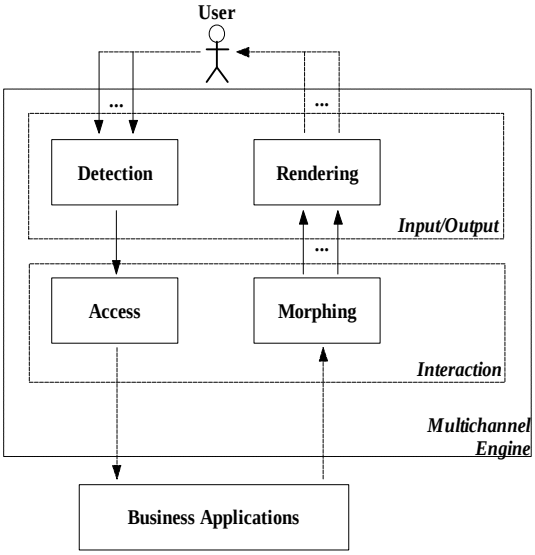
- Utente usa più modi di interazione, uno alla volta
- Motore di adattamento multicanale tra utente e logica applicativa (dati, processi, servizi)
- Logica applicativa tradizionale

Architettura multicanale (2/3)



- Segnali: in corrispondenza ai sensi associati ai modi
- Linguaggi di markup: rappresentazione Web-based dell'IN/OUT (HTML, WML, VoiceXML, cHTML, XHTML, InkML)
- Dati di business: il significato per la specifica logica applicativa

Architettura multicanale (3/3)



- Riconoscimento del modo (es. CC/PP [CCPP], UAProf [UAProf])
- Parsing ed estrazione dei dati di INPUT
- Adattamento dei dati di OUTPUT (es. XSLT [XSLT])
- Presentazione esterna

Pregi…

- Singoli linguaggi di markup noti e supportati da ambienti di sviluppo adeguati
- Sviluppo di applicazioni ormai consolidato e diffuso
- Ampia disponibilità di prodotti commerciali ad hoc ([MAS]):
  - Apache Cocoon
  - Volantis Mariner
  - MobileAware Everix
  - Oracle 9iAS Mobile
  - Sybase Mobile Application Studio
  - @Hand Mobile Application Server

…e limiti

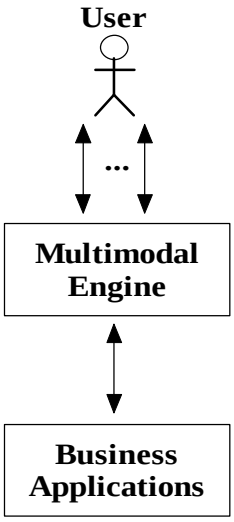
- Ogni modo ha i suoi problemi:
  - Inserimento ed accesso visuale ai dati difficoltoso su dispositivi mobili (schermi e tastiere piccoli) e progressivamente miniaturizzati
  - Riconoscimento vocale ancora potenzialmente incerto e, quindi, bloccante
  - Output vocale potenzialmente ingestibile se lungo
- Inadeguatezza di un “canale alla volta” per tutte le circostanze di vita umana
- Impossibilità di un’interazione continuata (“seamless”) con applicazioni e servizi

La multimodalità

Principali necessità

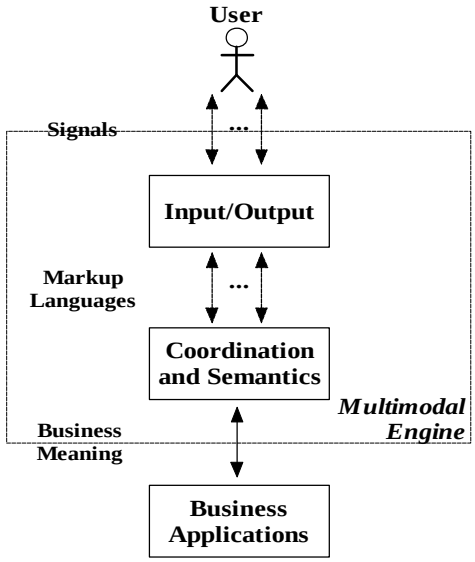
- Integrazione dei modi (in input)
  - A livello di segnale (es. sistemi vocali con riconoscimento del labiale)
  - A livello semantico (es. sistemi vocali con riconoscimento di gesti manuali)
    - Frame-based (pattern-matching)
    - Unification-based (programmazione logica)
- Time-stamping o temporal-cascading (in input)
- “Mutual disambiguation” dei modi (in input)
  - Es. effetto di un input modale su altri ed in relazione al contesto
- Scheduling temporale (in output)
  - Coordinamento degli output modali

Architettura multimodale (1/3)



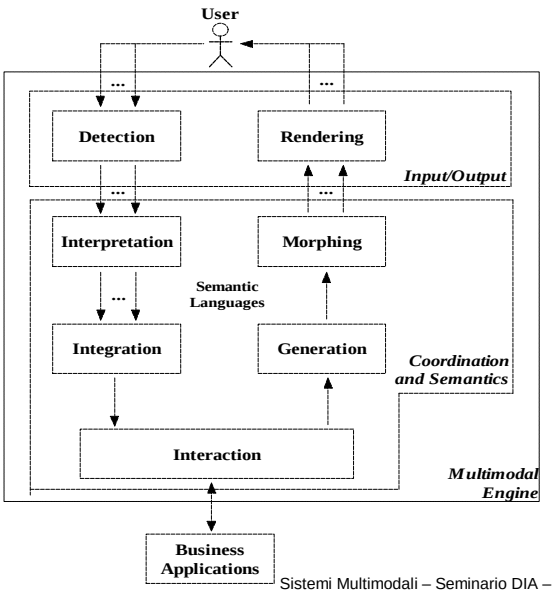
- Utente usa più modi di interazione, in maniera integrata e coordinata
- Motore di adattamento multimodale tra utente e logica applicativa (dati, processi, servizi)
- Logica applicativa tradizionale

Architettura multimodale (2/3)



- Segnali: in corrispondenza ai sensi associati ai modi
- Linguaggi di markup: rappresentazione Web-based dell'IN/OUT
- Dati di business: il significato per la specifica logica applicativa

Architettura multimodale (3/3)



- Quale linguaggio per la rappresentazione e l'interpretazione dei modi ?
- Come dare significato ad una combinazione interattiva ?
- Come "sincronizzare" il funzionamento del motore?

Sincronizzazione

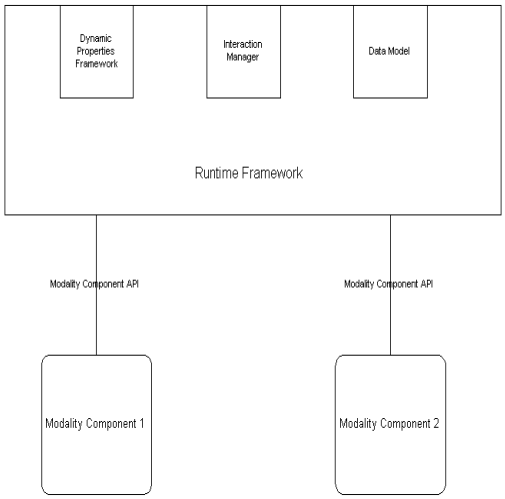
- **[Oviatt], [MMI]**

- **SMM in Input**
  - **Sequenziale:**
    - interpretazione del passo interattivo dipendente da un solo modo (può esserci alternanza)
  - **Sincronizzato tempo-indipendente (raro):**
    - interpretazione del passo interattivo dipendente da due o più modi (può esserci simultaneità)
  - **Sincronizzato tempo-dipendente:**
    - interpretazione del passo interattivo dipendente da due o più modi;
    - dipendenza semantica temporale stretta dei modi
- **SMM in Output**
  - **Sequenziale**:
    - un solo modo di presentazione all'utente per volta
  - **Sincronizzato**:
    - più modi di presentazione alla volta, simultanei o meno

Livello della sincronizzazione

- **Event-level**: Input in un modo sono recepiti a livello di eventi DOM ed immediatamente propagati ad un altro modo
- **Field-level**: Input in un modo sono propagati ad un altro modo dopo che l'utente cambia "fuoco" (es. movimento da un campo d'input ad un altro) or completa l'interazione con un campo (es. completa la selezione in un menu)
- **Form-level**: Input in un modo sono propagati ad un altro modo solo dopo che un particolare punto dell'interazione è stato raggiunto (es. dopo che un certo numero di campi di una form è stato completato)
- **Session-level**: Input in un modo sono propagati ad un altro modo solo dopo un passaggio esplicito di modo

Multimodal Engine – Una possibile implementazione (1/2)



- Adozione del paradigma MVC (Model-View-Controller)
- Componenti software:
  - Modality Components (views)
  - Service Components (controller)
  - Data Components (model)
- Interazione tra componenti:
  - via eventi asincroni
  - secondo modello publish/subscribe

Multimodal Engine – Una possibile implementazione (2/2)

- Esempio:
  - Runtime Framework ospitato a bordo di un browser multimodale integrato
  - EMMA usato come linguaggio di markup semantico dell'Interaction Manager del browser multimodale
  - XHTML può essere usato come linguaggio di markup per un Modality Component
  - VoiceXML può essere usato come linguaggio di markup per un Modality Component
  - SVG (Scalable Vectorial Graphics) può essere usato come linguaggio di markup per un Modality Component

Visuale vs. Multimodale

- Sequenzialità stretta degli eventi visuali
- Certezza simbolica dell'input e dell'output
- Permanenza del gestore di presentazione visuale sul client (basso footprint)
- Parallelismo degli eventi multimodali
- Ambiguità ed incertezza in input ed output (es. comprensione audio-vocale)
- Distribuzione dei componenti architetturali
- Architetture "time-sensitive"

Scenari architetturali

- Server-based o Thin Client
  - Distribuzione dei componenti del Multimodal Engine "across the network"
  - Riuso dei linguaggi standard di markup "tradizionali" (HTML, XHTML, WML, cHTML, VoiceXML) e dei rispettivi browser già esistenti
  - Necessità di sviluppo separato dei componenti del Multimodal Engine
- Device-based o Fat Client
  - Possibilità di pensare ad un unico browser multimodale in cui "collassare" le componenti del Multimodal Engine
  - Sviluppo "naturale" di nuovi linguaggi di markup per la multimodalità (utilizzabili peraltro anche nel caso sopra): XHTML+Voice (X+V), SALT, EMMA

Linguaggi per la multimodalità

- XHTML+Voice (W3C Submission by IBM, Motorola and Opera Software)
  - [X+V]
- SALT (CISCO, Intel, Microsoft et al.)
  - [SALT]
- EMMA (W3C Extensible MultiModal Annotation markup language)
  - [EMMA]

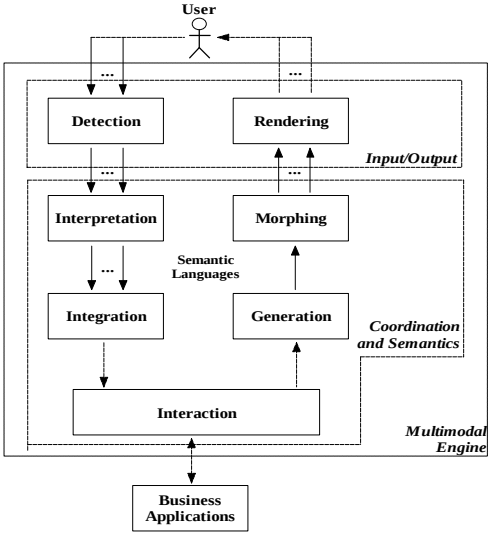
XHTML+Voice

```
<?xml version="1.0"?>
   <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML +Voice//EN"
   "xhtml+voice10.dtd">
   <html xmlns=http://www.w3.org/1999/xhtml
        xmlns:vxml="http://www.w3.org/2001/voicexml20"
        xmlns:ev="http://www.w3.org/2001/xml-events">
        <head>
               <title>Skeleton XHTML+Voice Document</title>
               <!-- voice handlers -->
               <vxml:form id="sayHello">
                      <vxml:block>Hello World</vxml:block>
               </vxml:form>
        </head>
        <body>
               <h1>Skeleton XHTML+Voice Document</h1>
               <p ev:event="onclick" ev:handler="#sayHello">
               Clicking anywhere on this paragraph results in a
               welcome message being spoken on account of
               attaching <code>vxml:form</code> handler to this
               paragraph.
               </p>
        </body>
   </html>
```

SALT

```
<?xml version="1.0"?>
<!-- HTML -->
<html xmlns:salt="urn:saltforum.org/schemas/020124">
    <body onload="RunAsk()"> 1
     <form id="travelForm">
    <input name="txtBoxDestCity" type="text" />
     </form>
<!-- Speech Application Language Tags -->
     <salt:prompt id="askDestCity"> Where would you like to go to? </salt:prompt> 3
     <salt:prompt id="sayDidntUnderstand" onComplete="RunAsk()">Sorry, I didn't
    understand.</salt:prompt>
     <salt:listen id="recoDestCity"
          onReco="procDestCity()"onNoReco="sayDidntUnderstand.Start()"><salt:grammar
          src="city.xml" /></salt:listen> 5
<!-- Script -->
   <script>
    function RunAsk() {
          if (travelForm.txtBoxDestCity.value=="") {
                    askDestCity.Start(); 2
                    recoDestCity.Start(); 4
          }
      }
    function procDestCity() {
          travelForm.txtBoxDestCity.value=recoDestCity.text; 6
          travelForm.submit();
      }
     </script>
  </body>
</html>
```


EMMA – Il linguaggio



- Linguaggio di markup XML-based usato per descrivere ed rappresentare *semantica e significato* di dati multimodali in input
- Componenti che generano EMMA:
  - Interpretation, Integration
- Componenti che usano EMMA:
  - Integration,Interaction
- Un documento EMMA può contenere tre tipi di dato:
  - Instance data
  - Data model
  - Metadata (Annotations)

Instance data

- Informazioni in input da interpretare a partire da vari dispositivi (telefono, “thin clients”, “rich clients”):
  - Modo (in ogni linguaggio umano)
    - Testo
    - Parlato (speech)
    - Scrittura a mano (handwriting)
    - Puntamento (pointing)
    - Video
  - Combinazione di modi
    - Singola
    - Sequenziale
    - Simultanea
    - Composta

Data model

- Vincoli su struttura e contenuto degli instance data (attività di validation):
  - Prestabiliti da un applicazione in formati arbitrari
    - XML
    - XMLSchema
    - XForms
    - Relax-NG
  - Impliciti (non specificati)

Metadata (1/2)

- Annotazioni (compatibili con il framework concettuale RDF) relative agli instance data:
  - Generali
    - Mancanza di input
    - Input non interpretabile
    - Timestamps
    - Posizione relativa di eventi di input
    - Raggruppamento temporale di eventi di input
    - Modi di input previsti
    - Lingue di input ammesse
  - Strutturali
    - Riferimento ad un data model
    - Rappresentazione di un input dalla composizione di più modi

Metadata (2/2)

- Di riconoscimento
  - Riferimento al metodo usato
  - Ambiguità
  - Confidenza di riconoscimento
- Di interpretazione
  - Riferimento al metodo usato
  - Ambiguità
  - Confidenza di interpretazione
- Dipendenti dal modo
  - Parlato
  - Scritto
  - Puntato

Esempio A (Integration)

```
<emma:emma emma:version="1.0"
   xmlns:emma="http://www.w3.org/2003/04/emma#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <emma:group emma:id="grp">
        <emma:interpretation emma:id="raw“
                emma:start="2004-03-02T0:00:00.15"
                emma:end="2004-03-02T0:00:00.515">
                <answer>From Boston to here tomorrow</answer>
        </emma:interpretation>
        <emma:interpretation emma:id="better">
                <emma:derived-from resource="#raw" composite=“false"/>
                <origin>Boston</origin>
                <destination>here</destination>
                <date>tomorrow</date>
        </emma:interpretation>
        <emma:interpretation emma:id="best">
                <emma:derived-from resource="#better" composite=“false"/>
                <origin>Boston</origin>
                <destination>here</destination>
                <date>20040303</date>
        </emma:interpretation>
        <emma:interpretation>
                <x>0.866</x>
                <y>0.724</y>
        </emma:interpretation>
   </emma:group>
```


Esempio B (Integration) (1/2)

```
<emma:emma emma:version="1.0"
   xmlns="http://www.w3.org/2003/04/emma#">
   <emma:interpretation emma:id="speech1"
        emma:start="2003-03-26T0:00:00.2"
        emma:end="2003-03-26T0:00:00.4"
        emma:process="http://example.com/myasr.xml"
        emma:source="http://example.com/microphone/NC-61"
        emma:signal="http://example.com/signals/sg23.wav"
        emma:confidence="0.6"
        emma:medium="acoustic"
        emma:mode="speech"
        emma:function="dialog"
        emma:verbal="true"
        emma:lang="en-US"
        emma:tokens="destination">
                <rawinput>destination</rawinput>
   </emma:interpretation>
```

Esempio B (2/2)

```
    <emma:interpretation emma:id="pen1"
      emma:start="2003-03-26T0:00:00.1"
      emma:end="2003-03-26T0:00:00.3"
      emma:process="http://example.com/mygesturereco.xml"
      emma:source="http://example.com/pen/wacom123"
      emma:signal="http://example.com/signals/ink5.inkml"
      emma:confidence="0.5"
      emma:medium="tactile"
      emma:mode="ink"
      emma:function="dialog"
      emma:verbal="false">
        <rawinput>Boston</rawinput>
    </emma:interpretation>
    <emma:interpretation emma:id="multimodal1"
      emma:process="http://example.com/myintegrator.xml">
      <emma:derived-from resource="#speech1" composite="true"/>
      <emma:derived-from resource="#pen1" composite="true"/>
        <destination>Boston</destination>
    </emma:interpretation>
</emma:emma>
```


Backup

Caratteristiche auspicate per sistemi multimodali (1/3)

[Reeves]

- **Accessibilità:**
  - massimizzazione abilità umane fisiche e cognitive in relazione al contesto
  - evitare combinazioni di modi non necessarie aggiungendo supporto a modalità combinate solo se aumenta efficacia ed efficienza dell'utente in un certo contesto
  - interazione audio-vocale preferita per dare comandi e rilevare allarmi
  - sequenzializzazione stretta tra puntamento ed interazione vocale
  - accoppiamento visuale-vocale-gestuale per informazioni spaziali
- **Adattabilità:**
  - ripartizione dell'informazione in funzione del modo
  - descrizioni audio di immagini grafiche
  - gesti accoppiati a comandi audio-vocali per interazioni in ambienti rumorosi o da parte di utenti con problemi audio-vocali

Caratteristiche auspicate per sistemi multimodali (2/3)

- **Flessibilità**
  - possibilità di scelta del modo (o dei modi) da parte dell'utente (user-initiated) o da parte del sistema (system-initiated)
- **Consistenza**
  - correttezza e completezza delle informazioni in input ed output indipendentemente dal modo interattivo
- **Feedback all'utente**
  - consapevolezza dell'utente del modo (o dei modi) usati e disponibili
  - meccanismi di notifica modi attraverso icone descrittive per modo visuale piuttosto "speech bubbles" per modo audio-vocale
  - conferme delle interpretazioni da parte del sistema, ad esempio dopo un'interpretazione integrata di più modo (caso "point and tell")

Caratteristiche auspicate per sistemi multimodali (3/3)

- **Gestione/Prevenzione congiunta di errori**
  - integrazione modi complementari per compensare debolezze di uno con forze dell'altro
  - monitoraggio e gestione delle performance di ciascun modo in contesti d'uso e controllo di scelta da parte dell'utente (es. comprensione di situazioni che richiedono grammatiche di riconoscimento aperte e cambio ad un modo visuale)
  - consentire cambio di modo in caso di errore (es. prevedere un switch ad una GUI dopo un numero fissato di errori d'interpretazione da parte di un ASR – Automatic Speech Recognition engine)
  - gestione congiunta su più modi di informazioni ambigue (ad es. presentando a video scelte in base alle possibili alternative "capite" da un ASR)

Principali riferimenti bibliografici

- [CCPP] Composite Capabilities Preferences Profile http://www.w3.org/Mobile/CCPP/
- [DI] W3C Device Independence (http://www.w3.org/2001/di/)
- [EMMA] W3C Extensible MultiModal Annotation markup language (http://www.w3.org/TR/emma)
- [ETSI] European Telecommunications Standards Institute. Human factors (HF): Multimodal interaction, communication and navigation guidelines. (Report No. ETSI EG 202 191 v 1.1.1 (2003-08). ETSI, Sophia Antipolis, France
- [MAS] Jain, R., Puglia, S., Wullert, J., Parmeswaran, K., Bakker, J.L. The Mobile Application Server (MAS): An Infrastructure Platform for Mobile Wireless Services. *Information Systems Frontiers Journal Vol.6 Issue 1*, Kluwer Academic Publishers, 2004, 23-34
- [MMI] W3C Multimodal Interaction Activity (http://www.w3.org/2002/mmi/)
- [MWI] W3C Mobile Web Initiative (http://www.w3.org/2005/MWI/)
- [Oviatt] Oviatt, S.L. Multimodal interfaces. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Application*s. J. Jacko and A. Sears, Eds. Lawrence Erlbaum, Mahwah, NJ, 2003, 286–304
- [Reeves] Reeves, L.M. et al. Guidelines for Multimodal User Interface Design. *Commun. ACM 47, 1* (Jan 2004), 57-59
- [SALT] SALT (http://www.saltforum.org)
- [WAI] W3C Web Accessibility Initiative (http://www.w3.org/WAI/)
- [XSLT] W3C XSL Transformations (http://www.w3.org/TR/xslt)
- [X+V] XHTML+Voice (http://www.w3.org/Submission/2001/13/) - W3C Submission by IBM, Motorola and Opera Software
- [UAProf] http://www.wapforum.org/what/technical/SPEC-UAProf-19991110.pdf