Multimodal Applications from Mobile to Kiosk

Michael Johnston

AT&T Labs – Research

W3C Sophia-Antipolis July 2004

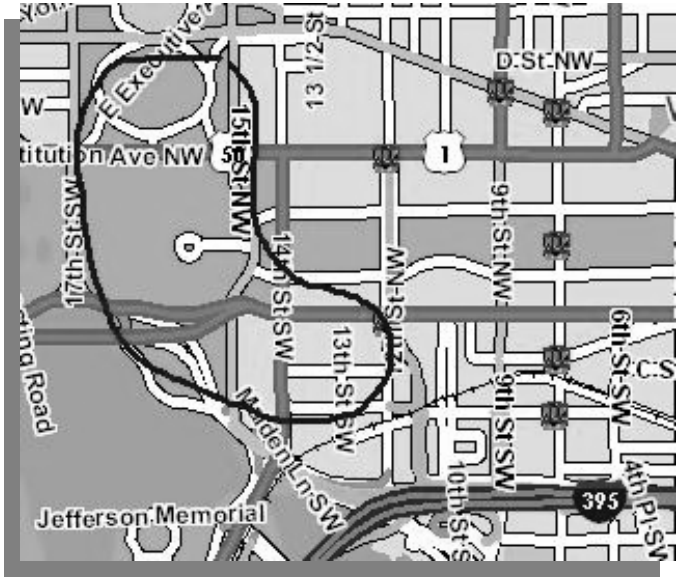
RETURN ON COMMUNICATIONS

Overview

- Composite multimodality
- MATCH: Multimodal Access To City Help
- MATCHKiosk
- Design Issues: Mobile vs. Kiosk
- Multimodal architecture
- Multimodal language understanding
- Conclusion

Composite Multimodality

- Composite Input
  - Enabling users to provide a single contribution (single turn) which is optimally distributed over the available input modes
  - e.g. Speech + Ink “zoom in here”

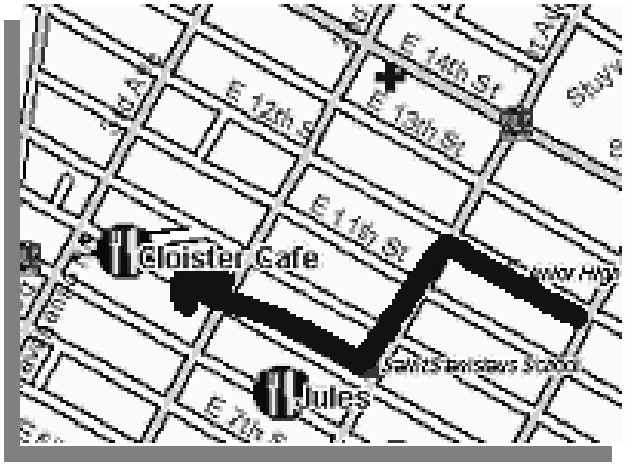


Composite Multimodality (cont.)

- Motivation
  - Naturalness: human communication is multimodal
  - Certain kinds of content within a single communicative act are best suited to particular modes, e.g.
    - Speech for stating complex queries or constraints, reference to objects not currently visible or intangible
    - Ink/Gesture for selection, indicating complex graphical features
  - Empirical studies
    - Task performance and user preference advantages
      - Oviatt et al 1997
    - Compensation for errors
      - Oviatt 1999
      - Bangalore and Johnston 2000

Composite Multimodality (cont.)

- Composite output
  - Similar motivations apply to output
  - System output to be optimally distributed across the available modes
  - For example:
    - High level summary in speech, details in graphics
    - “Take this route across town to the Cloister Café”



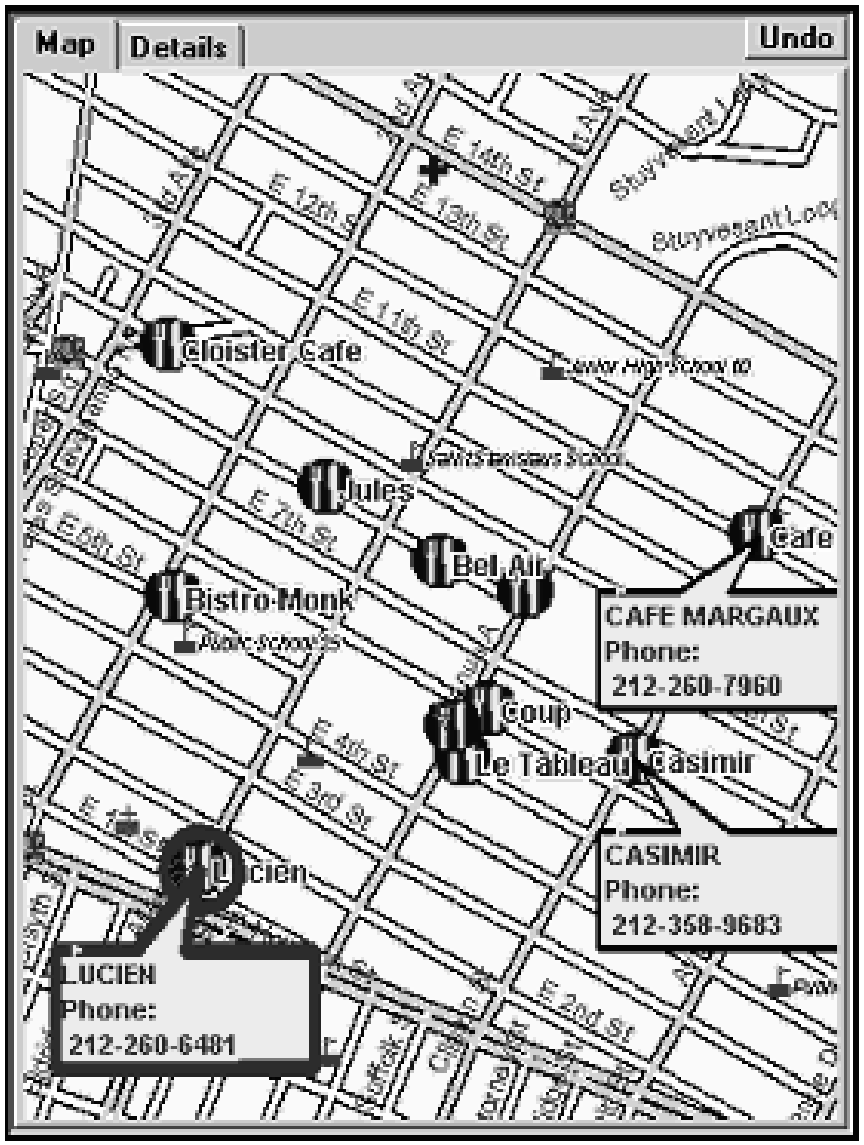
Composite Multimodality (cont.)

- Composite output
  - Another sample use
  - Multimodal help providing examples for the user
    - Hastie, Johnston, Ehlen 2002 (ICMI)
    - “To get the phone number for a restaurant, circle one like *this*, and say or write *phone*”

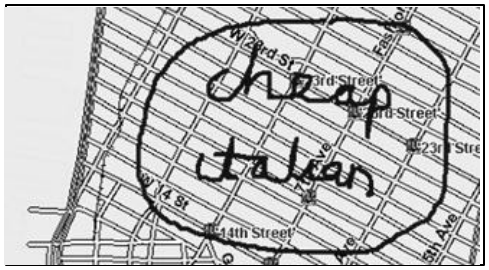
MATCH: Multimodal Access To City Help

- Interactive city guide and navigation for information-rich urban environments
- Provides information about restaurants, points of interest, and subway routes for New York and Washington, DC
- Mobile: Runs standalone on tablet or distributed over wireless network
- See Johnston et al 2001 (ASRU), Johnston et al 2002 (ACL)
- **Composite input**: Speech + Ink
- **Composite output**: Speech + Ink + Graphics

MATCH: Multimodal Access To City Help



- Finding restaurants
  - **Speech**: “show inexpensive italian places in chelsea”
  - **Multimodal**: “cheap italian places in *this* area”
  - **Pen**:



- Get information
  - “numbers for these three”
- Subway routes
  - “how do I get here from Broadway and 95th street”
- Pan/Zoom Map
  - “zoom in here”

Multimodal Interfaces for Public Kiosks

- Since introduction of ATMs in 70s, public kiosks have been deployed to provide users with a broad range of information and services
- Majority have rigid system-initiative graphical interfaces with user input by touch or keypad
  - Can only support simple tasks for able-bodied users
- To support more complex tasks for a broader range of users, kiosks will need to provide a more flexible and natural user interface
  - Multimodal interfaces provide naturalness and flexiblilty
  - e.g Gustafson et al 1999 (August), Narayanan et al 2000 (MVPQ), Raisamo 1998, Lamel et al 2002 (MASK), Wahlster 2003 (SmartKom Public), Cassell et 2002 (MACK)

MATCHKiosk

- Interactive multimodal kiosk providing city guide for Washington, DC. and NYC
- Supports both composite input and output
  - Speech, Ink, Graphics
- Deployed in AT&T visitor center in DC

MATCHKiosk Hardware

MATCHKiosk User Interface

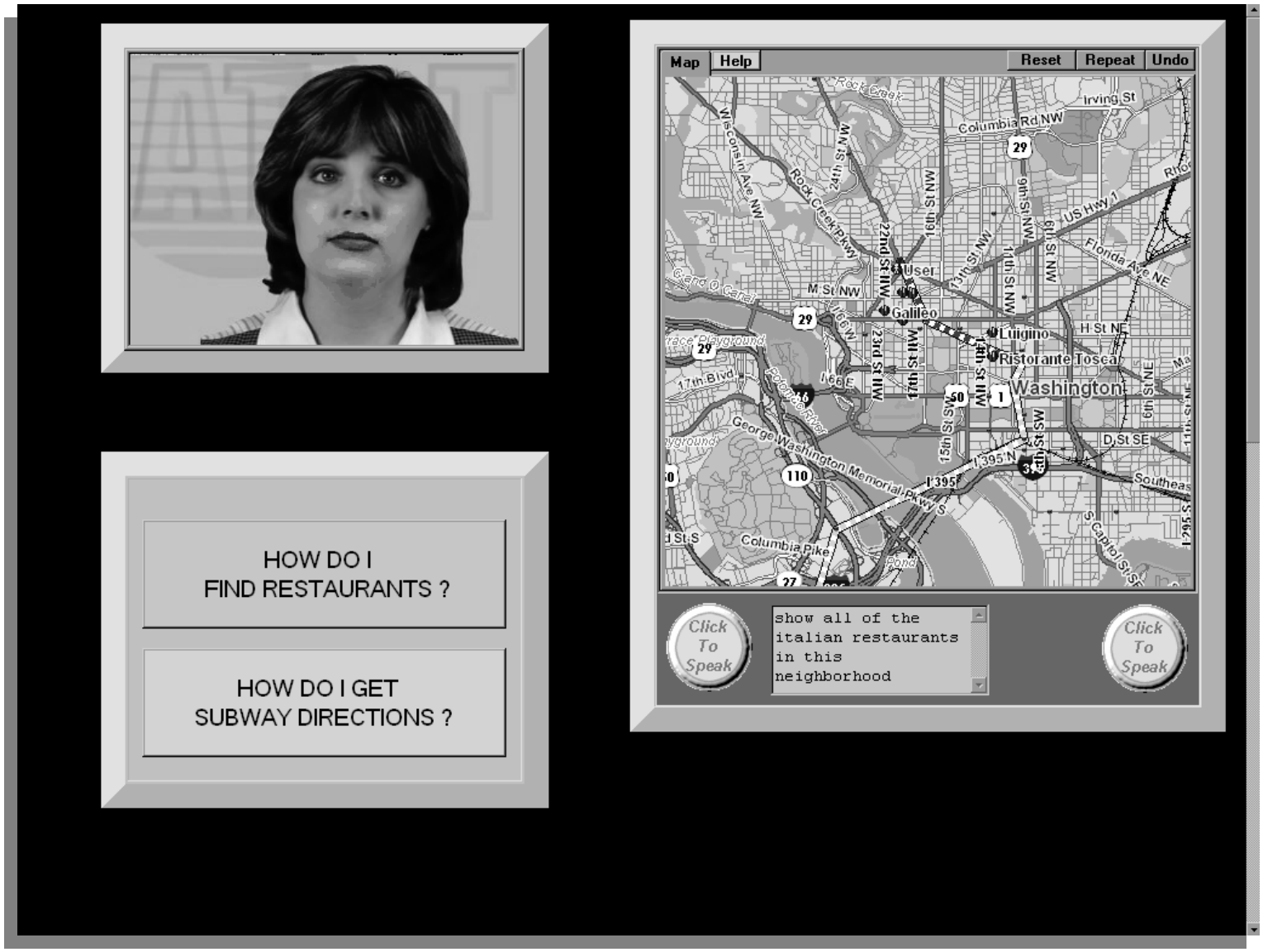
Life-like virtual human - Cosatto and Graf 2000

Dynamic Map Display

Context Dependent GUI Buttons

Click-to-Speak Buttons

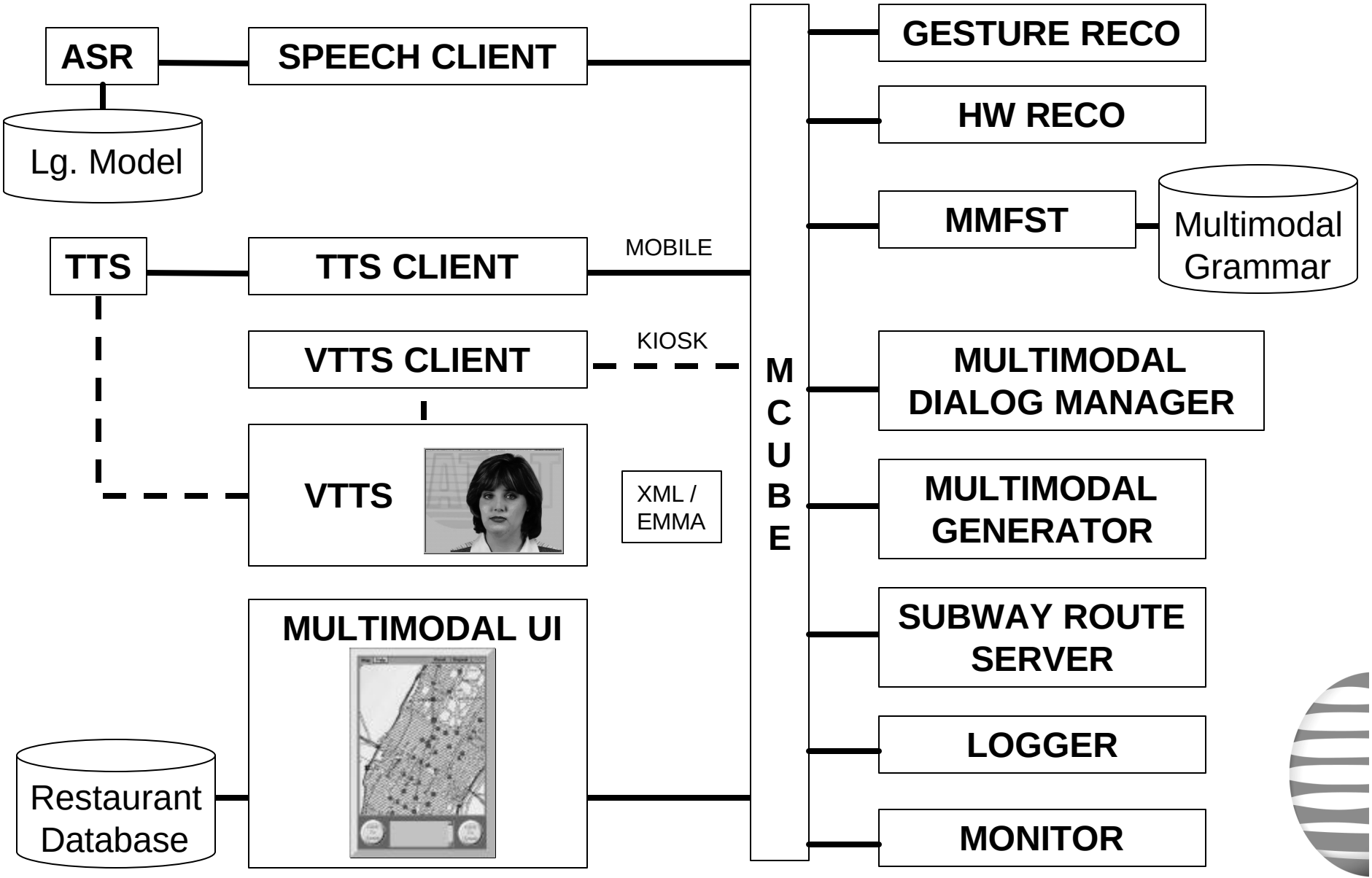
MATCHKiosk DEMO

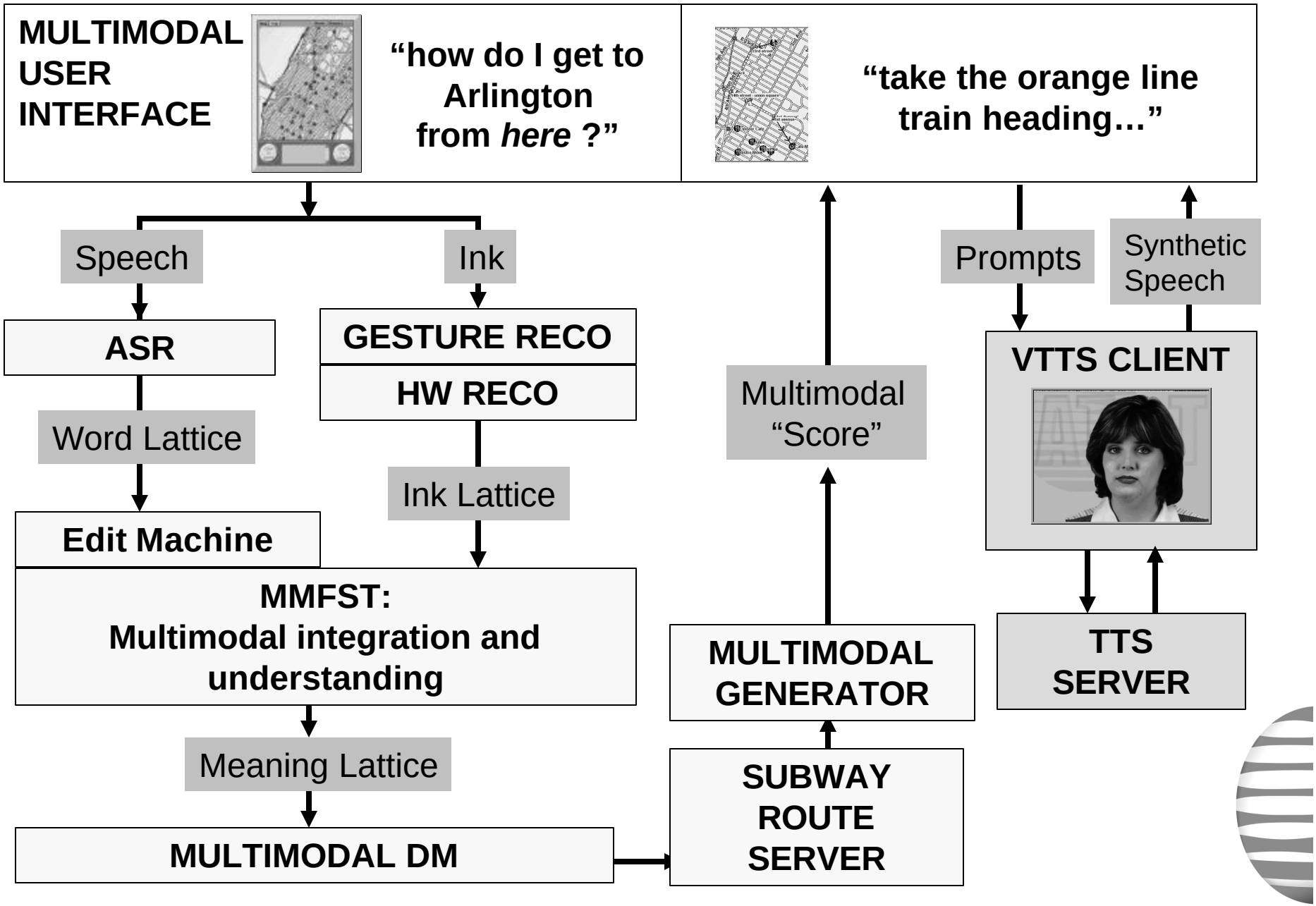


Design Issues for Mobile vs. Kiosk

- Array Microphone
- Robust Recognition and Understanding
  - Stochastic language model + Edit-machine
    - Bangalore and Johnston 2004, HLT-NAACL
- Social Interaction
- Context-sensitive GUI Buttons
- Printed output as a modality

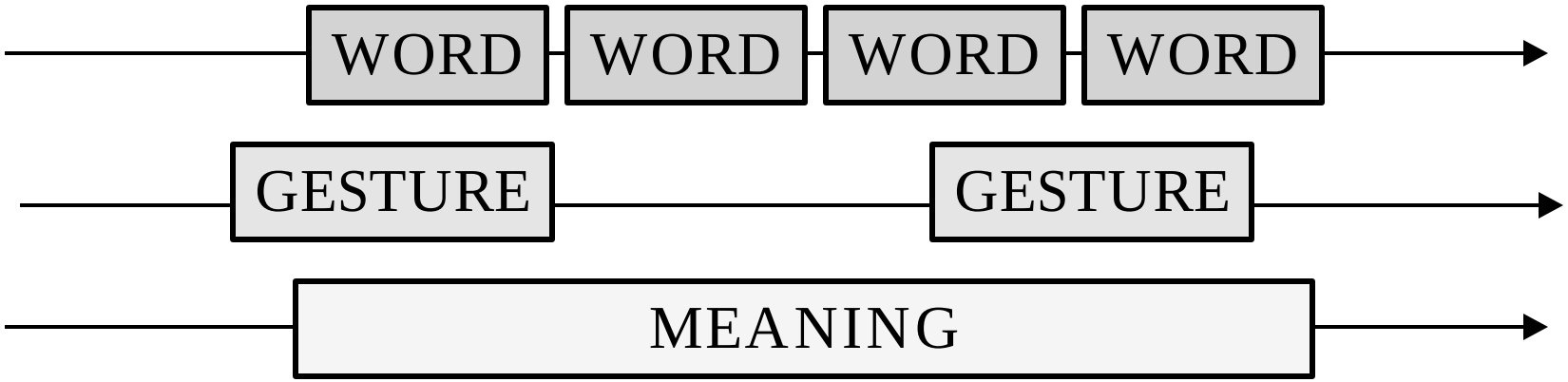
MATCH Multimodal Architecture




MULTIMODAL USER INTERFACE
"how do I get to Arlington from *here* ?"
"take the orange line train heading…"
Speech
Ink
ASR
GESTURE RECO
HW RECO
Word Lattice
Ink Lattice
Edit Machine
MMFST: Multimodal integration and understanding
Meaning Lattice
MULTIMODAL DM
SUBWAY ROUTE SERVER
MULTIMODAL GENERATOR
Multimodal "Score"
Prompts
Synthetic Speech
VTTS CLIENT
TTS SERVER

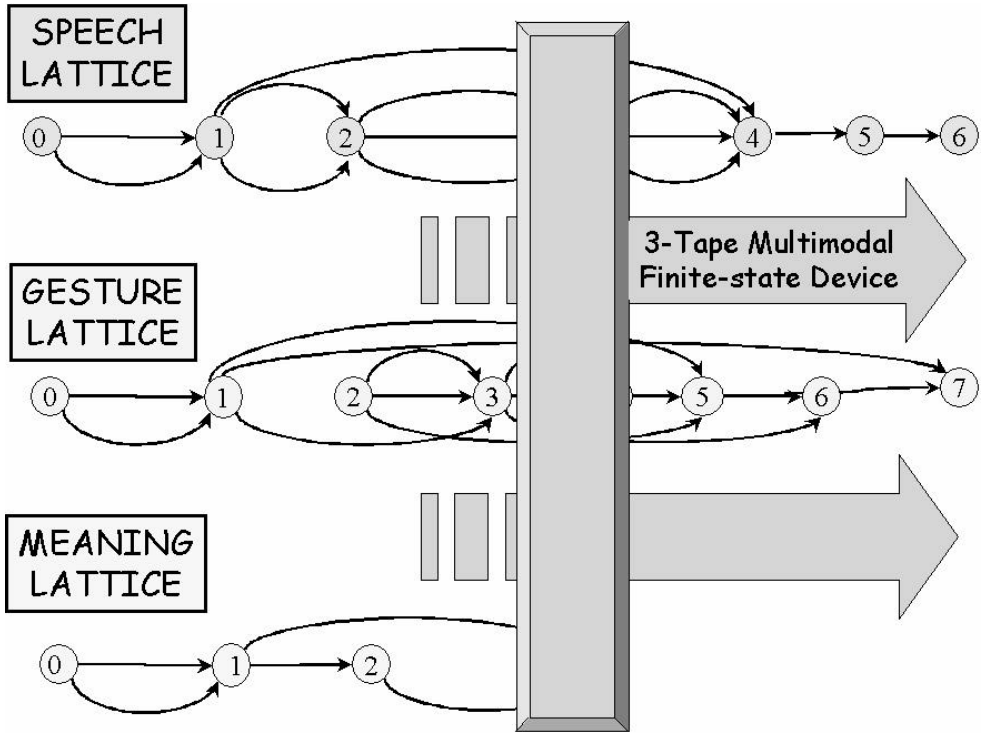
Multimodal Language Understanding

- Speech/text understanding
  - Associate word sequence with meaning
- Multimodal understanding
  - Associate word sequence + gesture sequence with meaning
  - Also associate gesture sequence to meaning



Finite-state Multimodal Language Understanding

- Speech and gesture parsing, multimodal integration, and understanding all captured within a single multimodal grammar model
  - Johnston and Bangalore 2000, 2004
  - Model can be compiled to efficient finite-state device
  - Interprets speech, pen, and multimodal inputs
  - Robust, Efficient, scalable framework for multimodal language processing
  - Enables compensation for errors in individual modes



Conclusion

- Multimodal applications supporting composite input and output
  - MATCH: Multimodal Access To City Help
  - MATCHKiosk
  - Multimodal grammars and finite-state multimodal understanding
- Positions
  1. In order to be effective, standards/frameworks for multimodal interaction should provide support for both composite input and composite output
  2. Composite input should be achieved by extending existing NLP techniques, parsing, understanding to operate over terminals in multiple input streams