Situated Multimodal Interaction for Local Search

Patrick Ehlen
AT&T Interactive

Michael Johnston
AT&T Labs

Mobile Information Access

Power of Speech

- ✓ Select from long lists
- ✓ Combine multiple constraints

to

of

”

obile

by

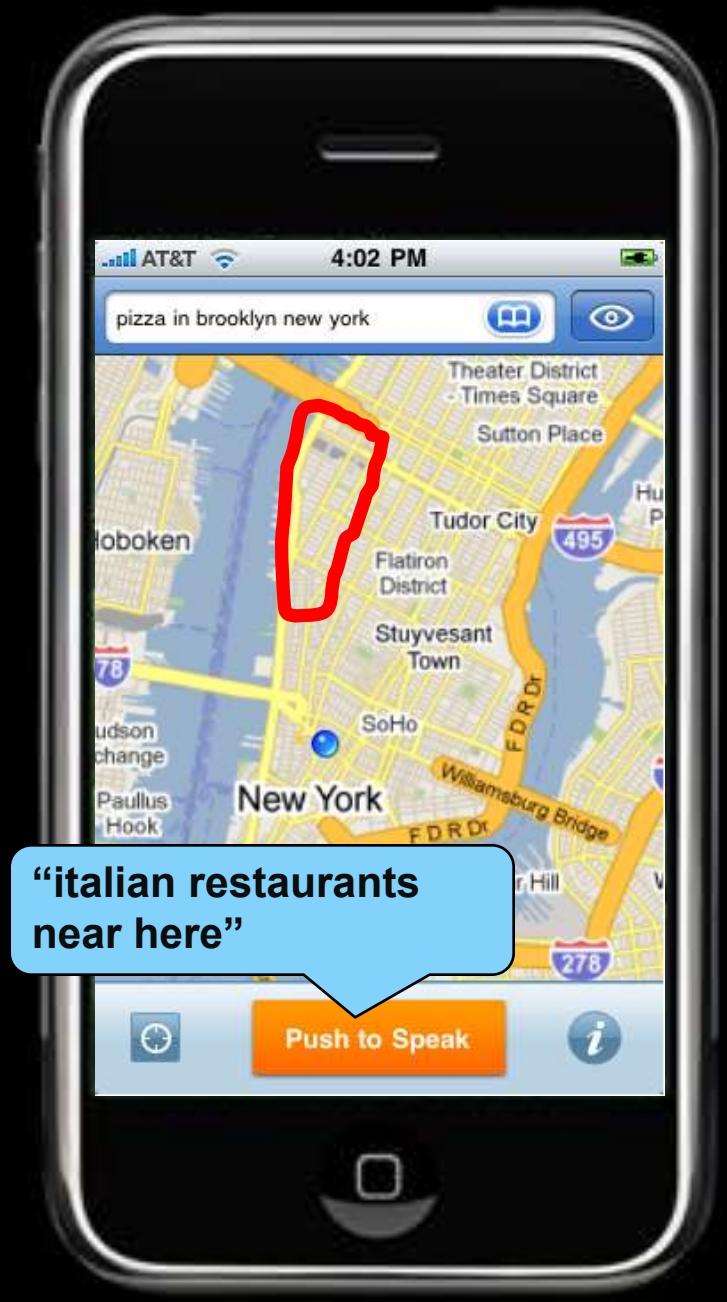
the bay”

Why Multimodal?

- ✓ Users should not have to act like their hands are tied behind their back
- ✓ Allow users to provide inputs in the most natural way possible
  - *"Certain tasks and functions cry out for particular modalities"* - Rudnicky and Hauptmann 1992
- ✓ Adaptation to the physical and social environment
  - Noise / Privacy / Eyes busy / Hands busy
- ✓ Superior error handling
  - Avoid error spirals - Oviatt and van Gent 1996
  - Mutual compensation through multimodal fusion - Oviatt 1999, Bangalore and Johnston 2000, 2005

Multimodal Local Search

- Interaction with dynamic map to search for businesses
  - MATCH (Johnston et al 2001)
  - SmartKom Mobile, CityBrowser, …
- Speak4It iPhone app
  - Empower users to refer to businesses or locations using speech, touch, or drawing
  - Don't always know the name or pronunciation of a location
  - Some locations don't have a name
  - Multimodal integration
  - Gesture recognition
  - Location salience modeling



Multimodal Architecture

Ink Capture:

"Push to Speak"
"Push to Interac

Reco

Multimodal Integration:
Fuse interpretations from

Interaction Manager:
Communicates with
Search, Geolocation, and
Location Salience services

Gesture
Reco

Interaction
Manager

Searc
h

Geo

Location
Salience
Model

Results Returned
to
Client

Situated Interaction







Situated Interaction

Situated Interaction

in ~80% queries, users speak no location

Grounding Contexts for Salient Location

"italian restaurants"

PHYSICAL

User's current location (GPS)

GESTURE

Where user touched

GUI

Location shown on map display

VERBAL

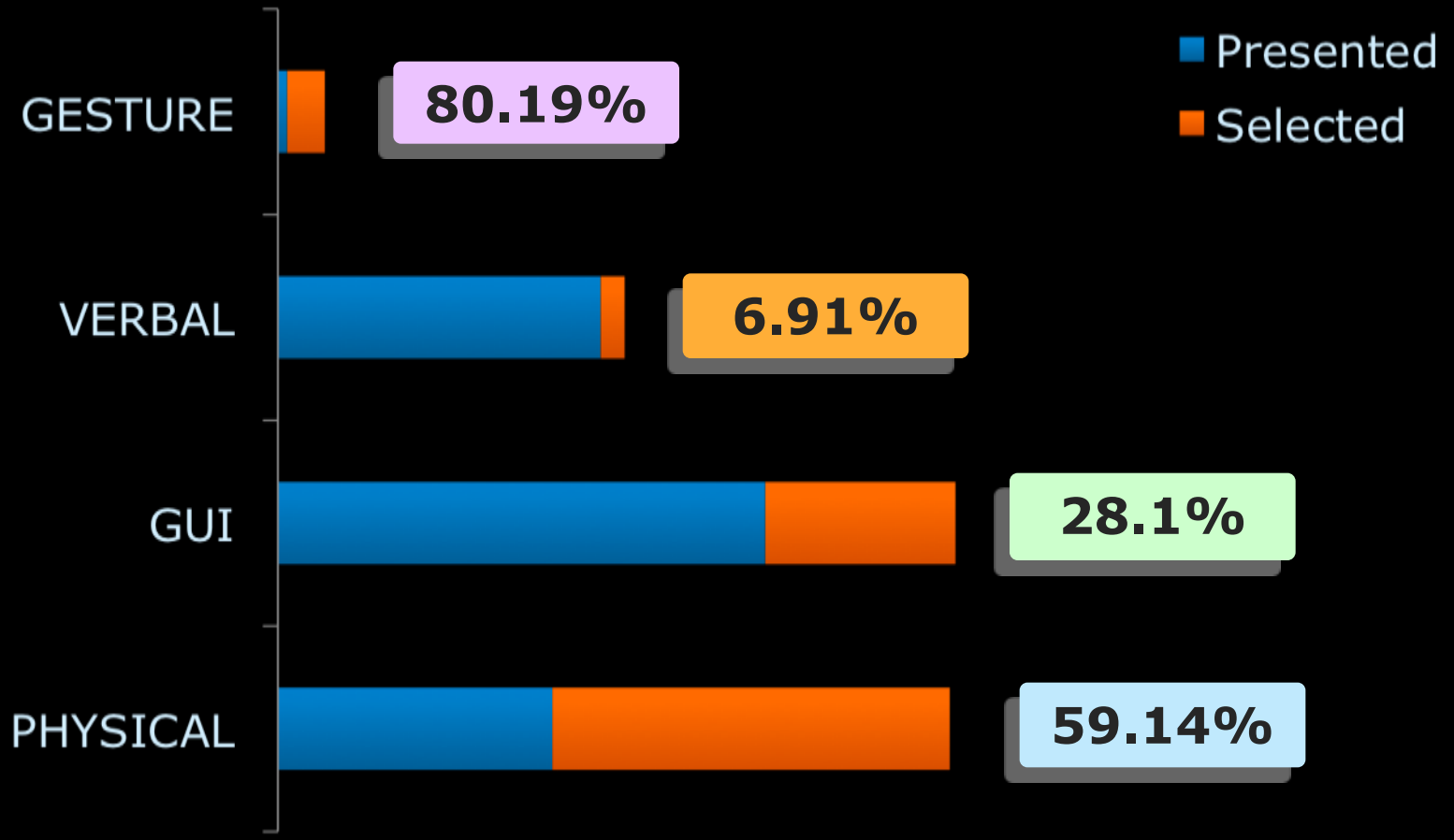
Place spoken in prior query

"Sorry I could not find french restaurants in madison"

Which Location Does The User Feel Has Been Grounded?

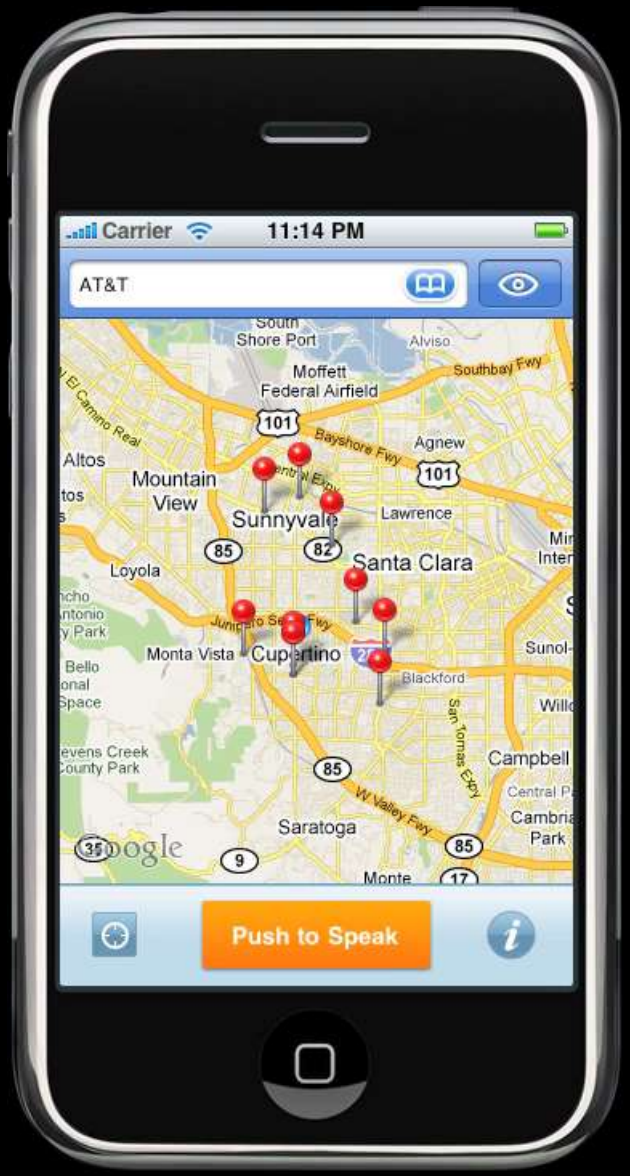
- Experiment Strategy:
  Let users say & do whatever they are inclined to say & do
- Present them with a "salient location disambiguation" screen that will gather user-generated "truth" of the intended context
- Display to 10-20% of unlocated queries for a limited time
- Train a context model using the data

Which Location Does The User Feel Has Been Grounded? (Data Evidence)



Conclusion

- ✓ Multimodality enables more natural and effective interaction for mobile voice applications
  - "users should not have to act like their hands are tied behind their back"
- ✓ Speech recognition, Understanding, and Multimodal processing need to be sensitive to the situated nature of mobile interaction
    - e.g. for Location Salience
      - Physical Context
      - GUI context
      - Dialog context


Carrier
11:14 PM
AT&T
Moffett
Federal Airfield
Southbay Fwy
101
Bayshore Fwy
Agnew
Altos
Mountain
View
Sunnyvale
Lawrence
85
82
Santa Clara
Loyola
Monta Vista
Cupertino
Blackford
Campbell
W Valley Fwy
Saratoga
9
Push to Speak