# Multimodal Dialogue Interfaces

a MUSCLE e-team presentation

Manolis Perakakis
Alexandros Potamianos
Tech. Univ. of Crete

WP5

# Research Goals

- Build **state-of-the-art** MDS (GUI + Speech)
  - Demonstration and evaluation test-bed
- Demonstrate/exploit the **synergies** between modalities, e.g. :
  - Input : consistent (GUI), inconsistent (speech)
  - Output : fast (GUI), slow (speech)
- Investigate the "**optimal modality input mix**"
  - How/why do users select input modality?
  - Is unimodal efficiency the only criterion?

## Speech: an interaction modality and more …

- Speech is a strong correlate for
  - Gender
  - Emotion
  - Personality
  - Speaker's face
- In human-human communication people expect
  - Reciprocity
  - Symmetry
  - Collaboration
- Speech communication is a social act that implies presence

# Idiosyncrasies of the speech modality

- Speech modality does not “respect” fundamental human-computer interface design principles(!)
  - Control
  - Efficiency
  - Consistency
  - Familiarity and Transparency
  - Forgiveness and Recovery

# Design principles for multimodal dialogue systems

- HCI design principles for multimodal systems
  - Consistency between interaction modalities
    - Symmetric multimodality
    - No representation without presentation
  - Efficiency and synergy
  - Robustness
  - Compositionality

# Multimodal Dialogue Systems and Synergies

To build efficient MM systems we need to **exploit the synergies** between the modalities :

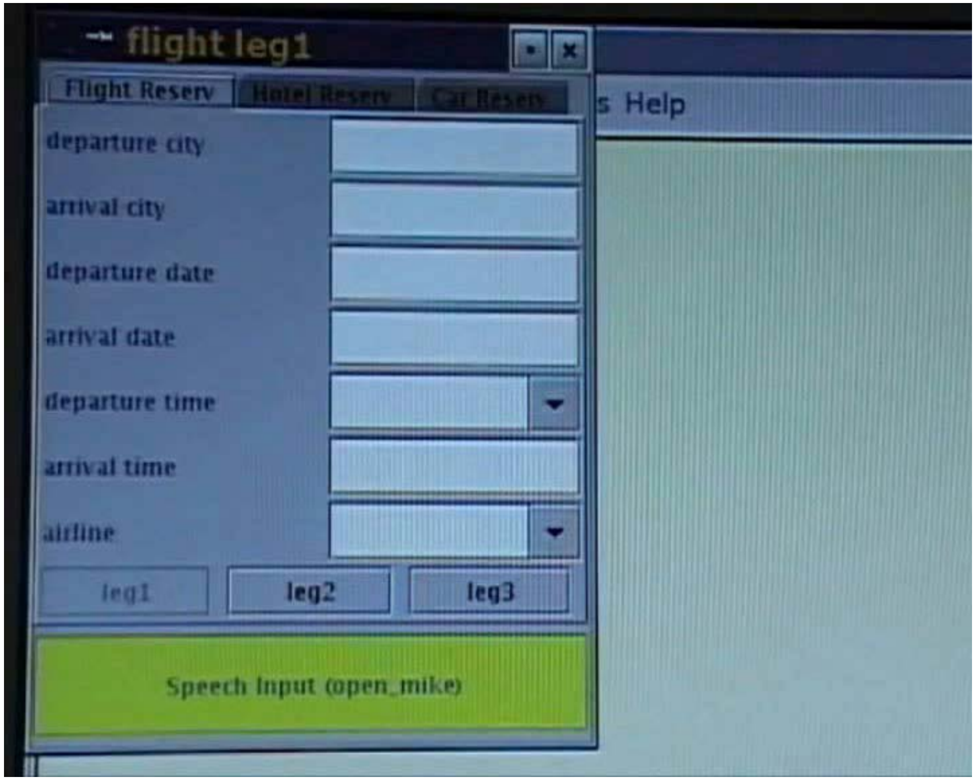
- **Output : Attributes values** are displayed at the GUI and focus (context) of speech is **highlighted**
- **Output : Speech prompts** are significantly **shorter!**

  (mostly used to emphasize information displayed visually)
- **Input :** Freedom of **input choice** : Speech or GUI
- **Error correction : Erroneous** values/ambiguity can be easily **corrected** via the GUI

# Interaction Modes Evaluated

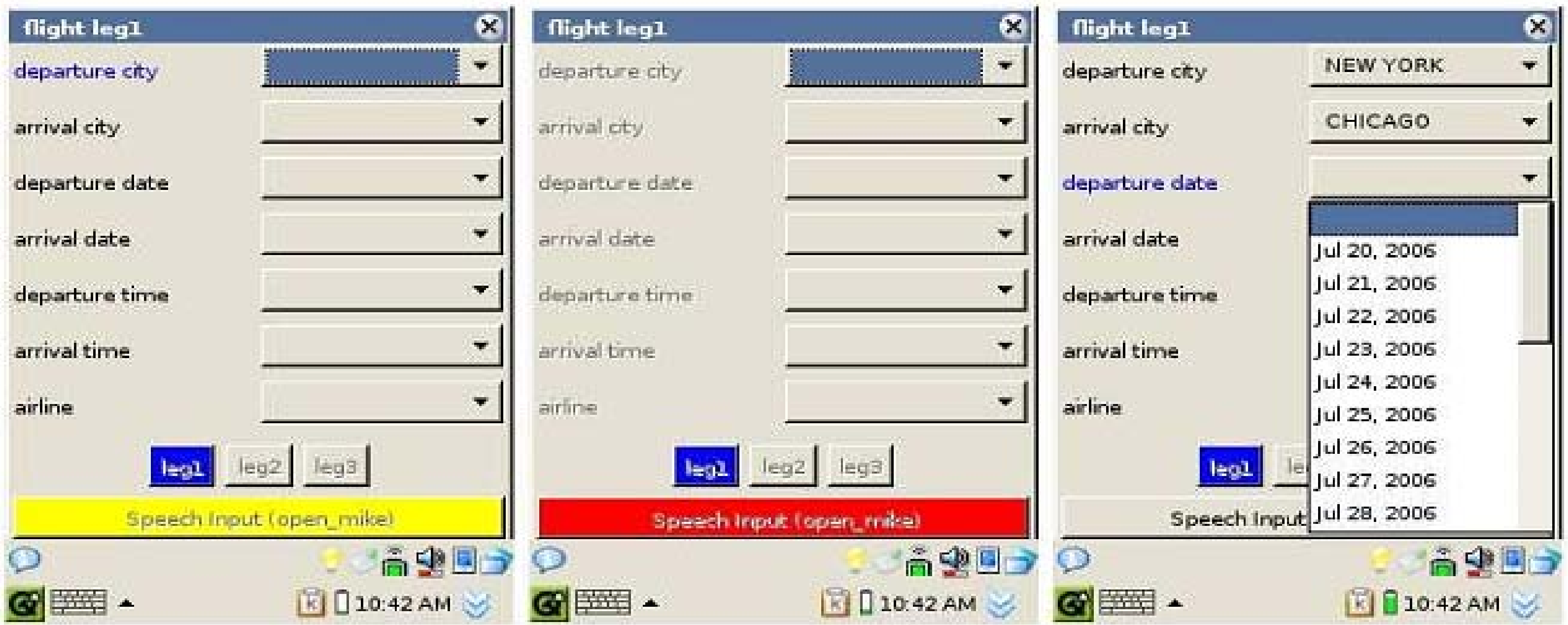
- **Unimodal interaction**
  - “Speech-Only” [SO]
  - “GUI-Only” [GO]
- **3 multimodal (MM) systems :**
  - “**Click-to-Talk**” [CTT] : GUI is the default input mode
  - “**Open-Mike**” [OM] : speech is the default input mode
  - “**Modality-Selection**” [MS] : selects default input based on unimodal efficiency considerations – current attribute size
  - **NOTE** : users can **override** proposed input modality
- **Open-Mike with Speech input [OMSI]**
  - Investigate **visual feedback** effect

# System Demo (Desktop version)



# PDA environment : “Modality Selection” example

## Input : From New York to Chicago




## Default input based on current attribute size :

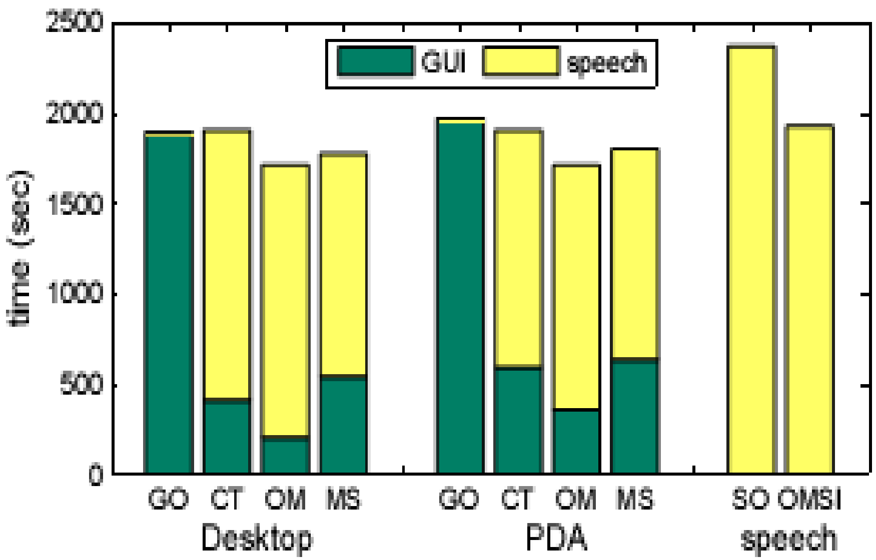
**a. System in Open-Mike mode (departure city is a long attribute)**

**b. Voice activity detected**

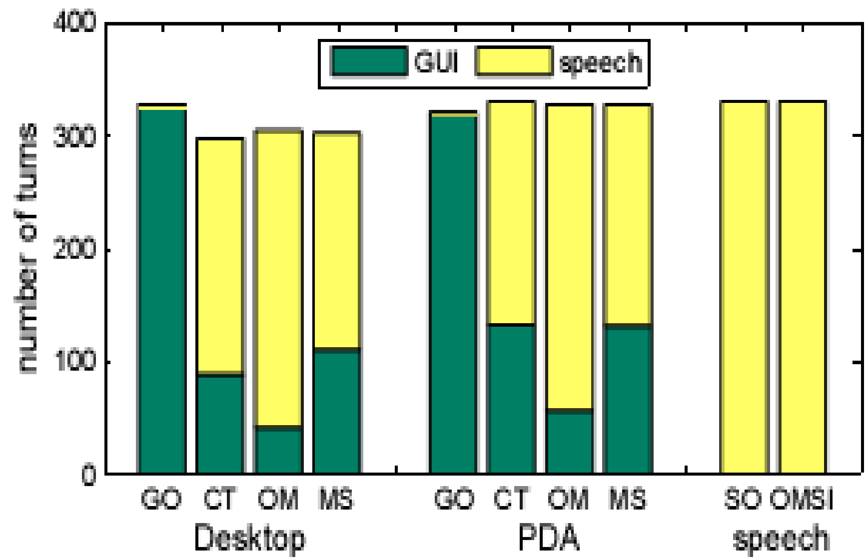
**c. System transitions to Click-to-Talk mode (date is a short attribute)**

# Evaluation and Mode Statistics

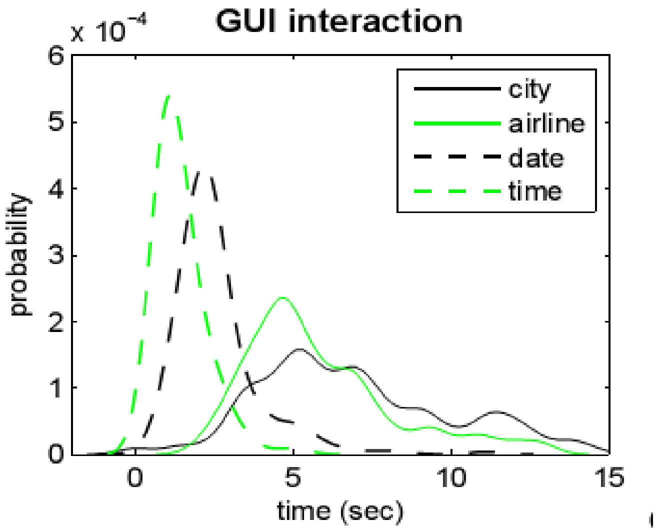
- Application : form-filling travel reservation
- 5 scenarios: 1/2/3 leg flight, round-trip with car/hotel reservation
- 2 speech systems (SO/OMSI) and 4 (GO + 3 MM) for each platform
- **Mode statistics :**

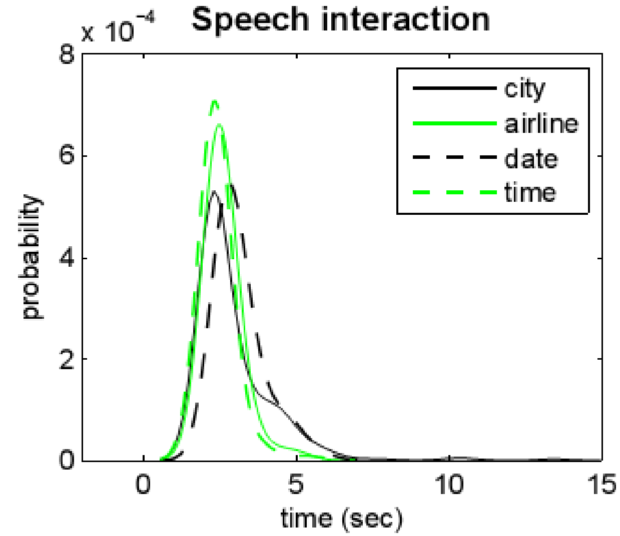


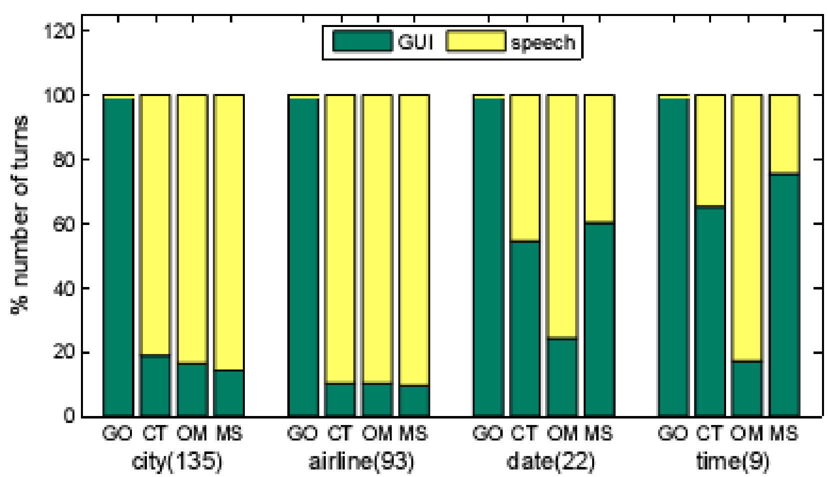

(a)



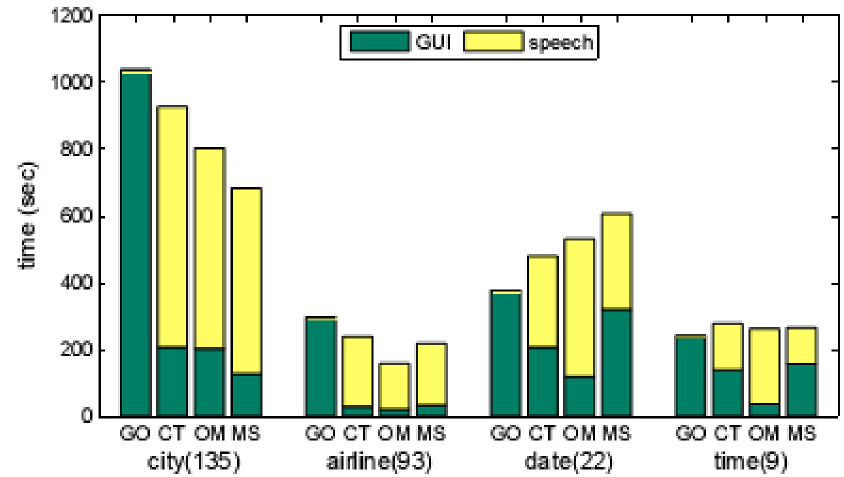
# Modality selection and unimodal efficiency (context statistics)




(a)



# Evaluation of multimodal form filling systems

- Traditional evaluation metrics fail to provide valuable information and identify usability problems
- We propose two new metrics :
- Relative modality efficiency can identify suboptimal use of modalities
- Multimodal synergy measures the added value from combining multiple input modalities and can be used as a single measure of the quality of modality fusion & fission in multimodal systems

WP5

# Relative modality efficiency

- Relative modality efficiency :

$N_s, N_g$ : number of fields filled correctly using speech/GUI
$T_s, T_g$ : overall time spent using speech/GUI

- Relative modality usage :

$$U_s = \frac{T_s}{T_s + T_g}.$$

- Relative modality usage efficiency :

$$E_s = \frac{N_s T_g}{N_s T_g + T_s N_g}$$

# Multimodal synergy

- Multimodal synergy :

$D_s, D_g$ : Completion time for "GUI - only" & "Speech - only" unimodal systems

$D_r$ : completion time for the random multimodal system :

$$D_r = U_s D_s + U_g D_g$$

$D_m$ : time to completion for the actual multimodal system

$$S_m = \frac{D_r - D_m}{D_r}$$

# Random Multimodal synergy

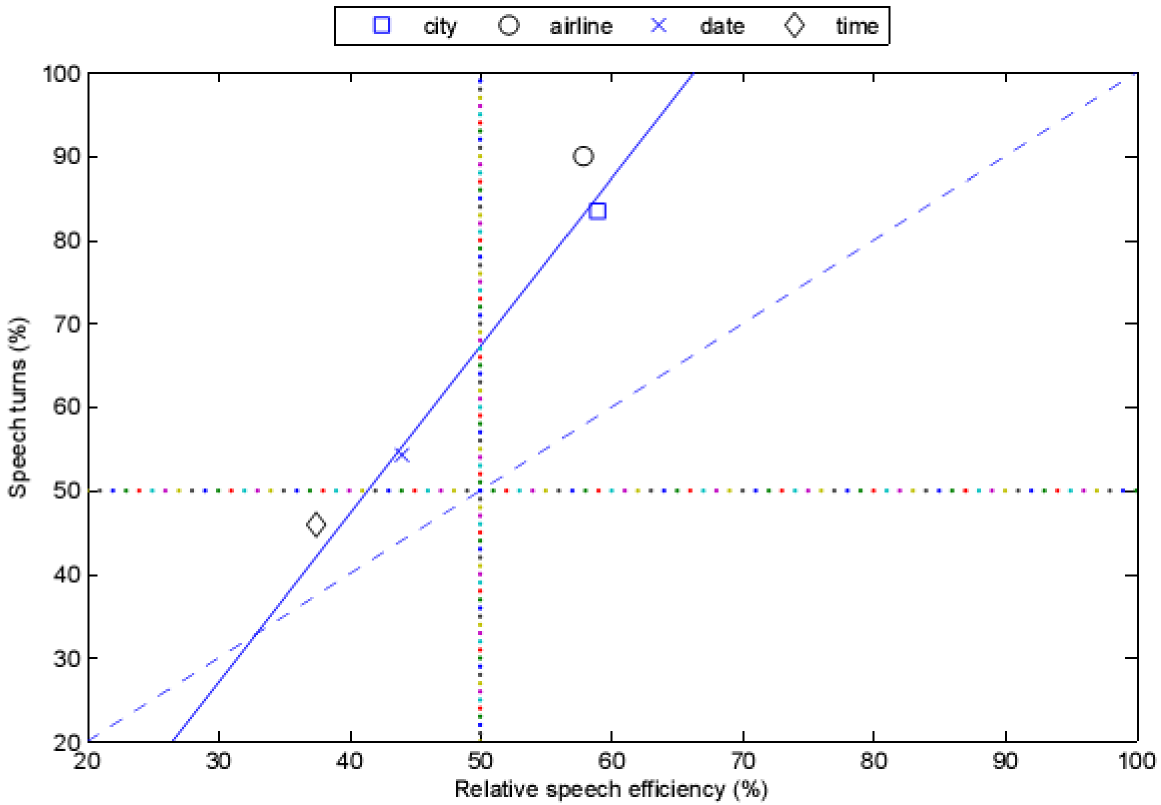
- Multimodal synergy based on random modality choice:
- Completion time for the "true random" multimodal system :

$$D_r^R = (1/\bar{N})\sum_{i=1}^{N} D_i$$

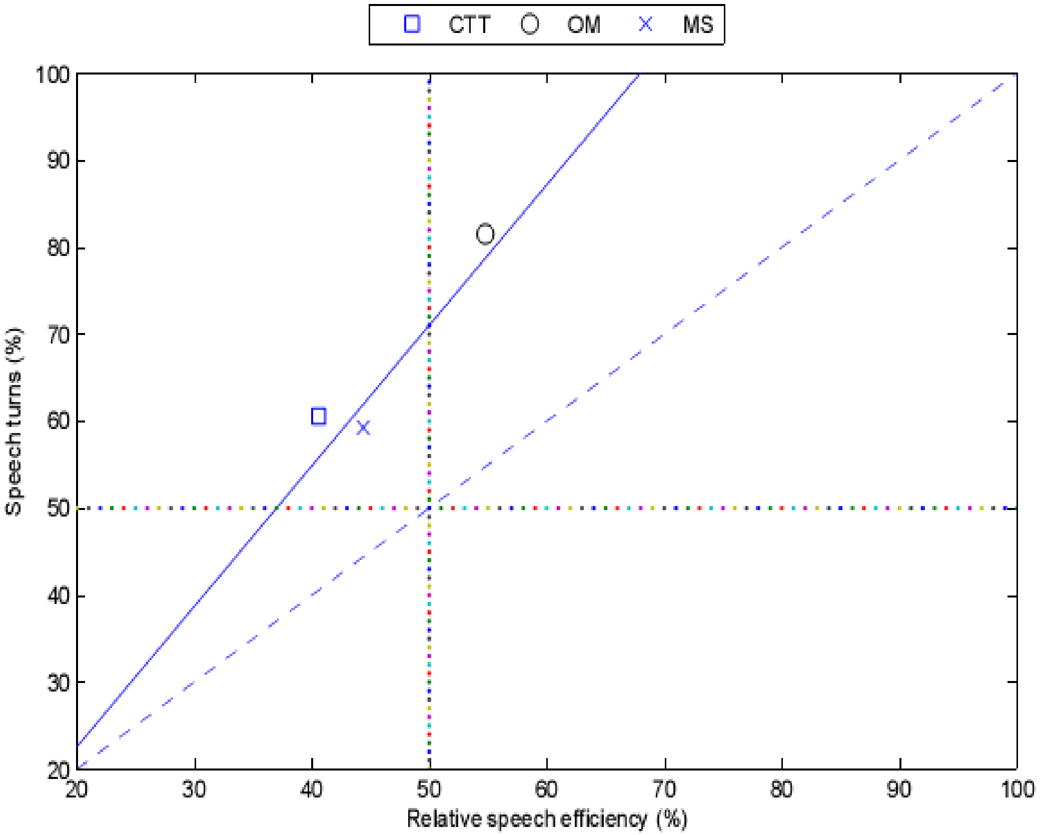
- "Random multimodal synergy" :

$$S_m^R = \frac{D_r^R - D_m}{D_r^R} = 1 - \frac{N\ D_m}{\sum_{i=1}^{N} D_i}$$

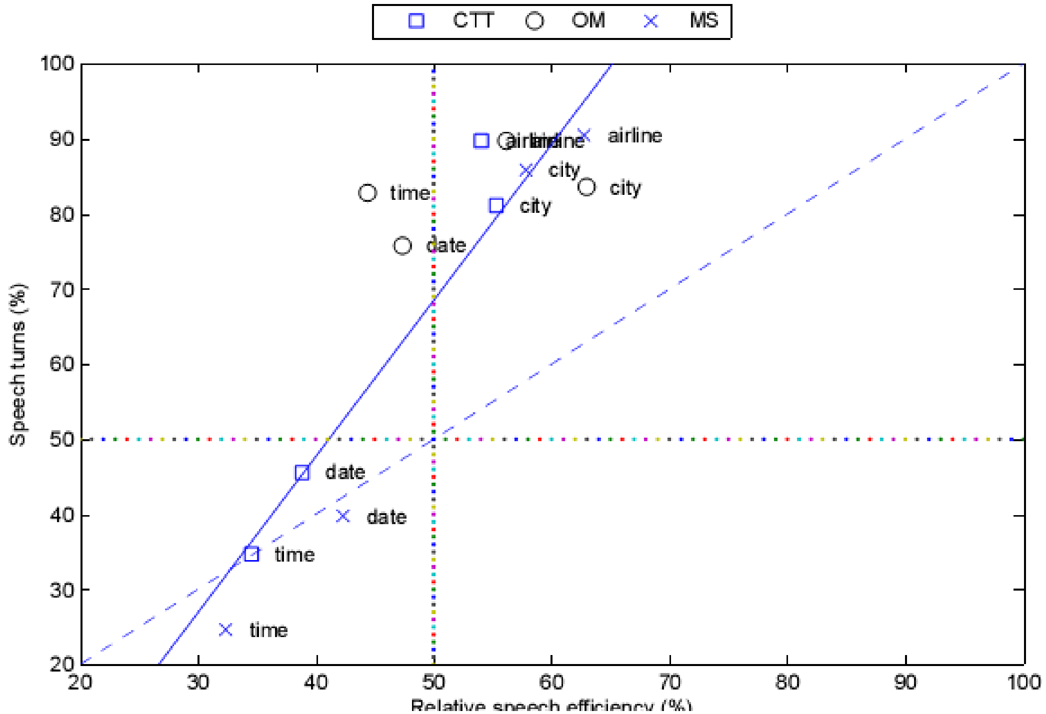
# Relative speech efficiency for the four contexts



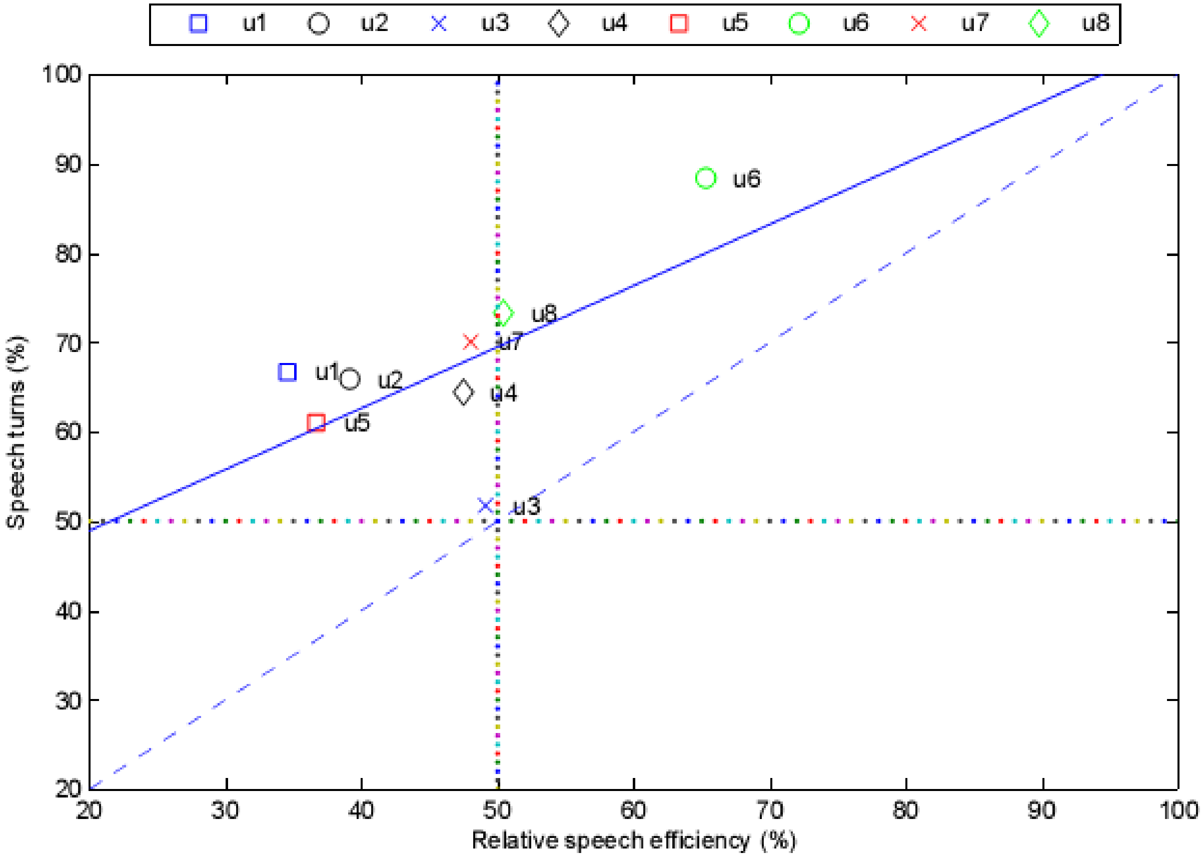
# Relative speech efficiency for the three modes



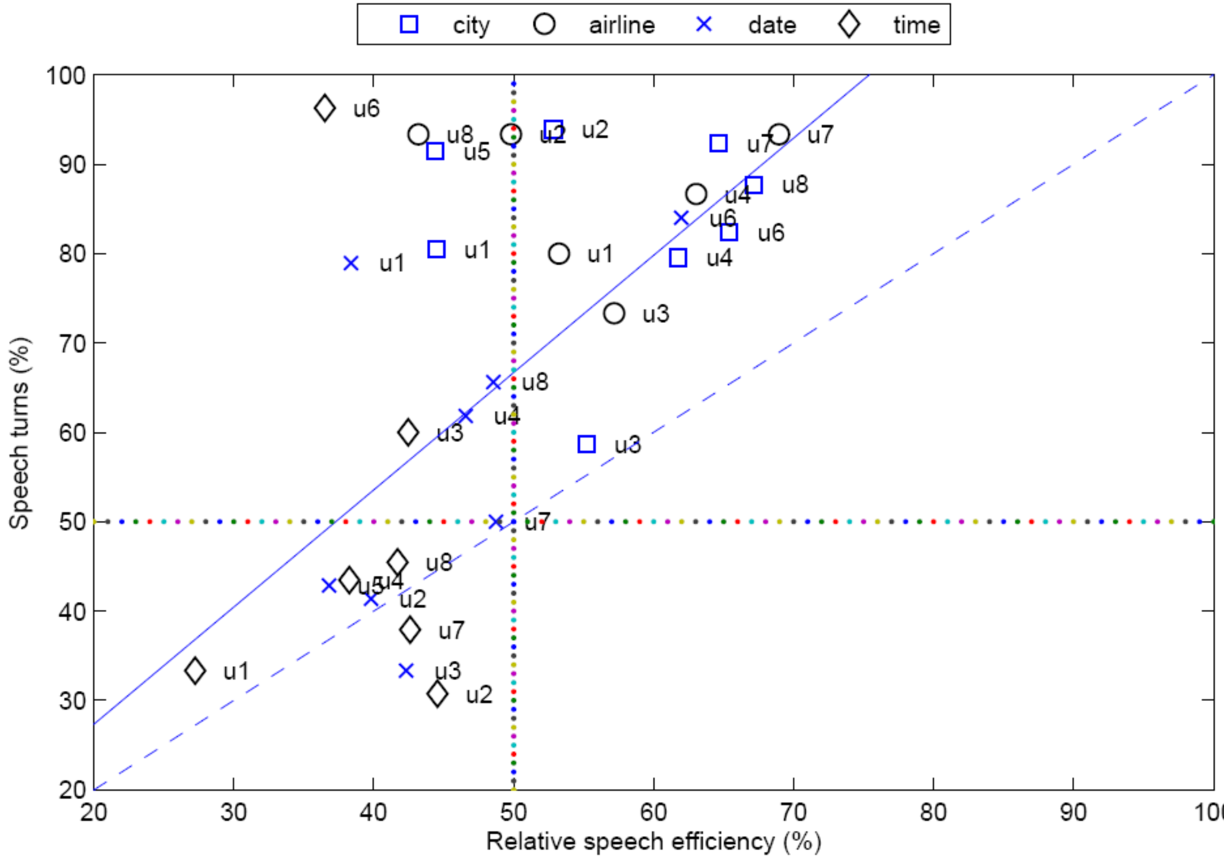
# Relative speech efficiency for mode/context



# Relative speech efficiency for the eight users



# Relative speech efficiency for users & contexts



# Results : synergy and multimodal modes

- Synergy results for the three multimodal interaction modes

| Mode | click-to-talk | open-mike | modality-selection |
|---|---|---|---|
| inactivity | -2.6 | 25.5 | 0.0 |
| interaction | 24.0 | 17.8 | 31.0 |
| overall | 12.7 | 21.1 | 17.8 |

- Results show modality-selection has the highest synergy for interaction times; they used input modality based on efficiency considerations more times compared to other systems

# Results : synergy and contexts

- For interaction times there is a clear separation for long and short attributes
- Synergy > 30% for city/airline due to input modality choice (use speech input since it is much more efficient compared to pen input )
- Synergy is much lower for short attributes. The difference in unimodal efficiency between the two modalities is smaller.

| context | city (135) | airline (93) | date (22) | time (9) |
|---|---|---|---|---|
| inactivity | -8.1 | 21.6 | 4.9 | 24.9 |
| interaction | 33.1 | 31.5 | 6.6 | 10.3 |
| overall | 18.7 | 27.6 | 5.8 | 18.4 |

# Results : synergy and users

- For inactivity times there is high variability. Some users even show negative synergy (u4, u5), demonstrating high cognitive load
- For interaction times there is high variability. User u7 has an impressive 39% over combined unimodal efficiency.
- Users helped by system design, can improve considerably their performance compared to unimodal systems.

| User | u1 | u2 | u3 | u4 | u5 | u6 | u7 | u8 | mean | std |
|---|---|---|---|---|---|---|---|---|---|---|
| inactivity | 16.4 | 21.4 | 8.4 | -21.1 | -2.7 | 9.6 | 24.8 | 2.5 | 7.4 | 14.7 |
| interaction | 26.5 | 33.2 | 15.5 | 30.5 | 17.2 | 14.4 | 39.0 | 13.4 | 23.7 | 9.85 |
| overall | 22.8 | 28.2 | 12.5 | 11.0 | 10.0 | 12.0 | 32.5 | 8.2 | 17.2 | 9.33 |

# Results : synergy and users II

- Synergy across the eight users and the three multimodal modes is shown. The mean and standard deviation is also shown in the right part.
- Again note the disparities among users.

| Time | Mode/User | u1 | u2 | u3 | u4 | u5 | u6 | u7 | u8 | mean | std |
|---|---|---|---|---|---|---|---|---|---|---|---|
| inactivity | CT | 22.6 | 22.5 | -13.1 | -19.8 | -29.6 | -0.2 | 3.5 | -8.2 | -2.8 | 18.8 |
| | OM | 29.3 | 25.0 | 29.1 | -16.0 | 23.5 | 30.2 | 48.6 | 27.5 | 24.7 | 18.2 |
| | MS | -5.2 | 16.8 | 6.5 | -27.8 | -0.8 | -0.0 | 21.7 | -12.1 | -0.1 | 15.8 |
| interaction | CT | 22.8 | 38.5 | 16.1 | 32.9 | 21.3 | 2.3 | 38.8 | 13.1 | 23.2 | 12.9 |
| | OM | 24.5 | 21.7 | 10.8 | 24.1 | 6.5 | 9.5 | 34.6 | 5.9 | 17.2 | 10.5 |
| | MS | 32.9 | 38.9 | 19.9 | 35.1 | 23.8 | 30.4 | 43.5 | 21.7 | 30.8 | 8.5 |
| overall | CT | 22.7 | 31.8 | 3.6 | 12.9 | 2.8 | 1.1 | 22.7 | 2.9 | 12.6 | 11.8 |
| | OM | 26.2 | 23.1 | 18.6 | 9.0 | 12.7 | 19.9 | 41.0 | 16.3 | 20.2 | 9.8 |
| | MS | 19.1 | 29.6 | 14.2 | 11.3 | 14.9 | 15.1 | 33.6 | 5.5 | 17.9 | 9.4 |

# Summary

- Unimodal efficiency affects input modality choice but it is not the only factor (speech bias)
- The interface design of a multimodal system can affect user behavior e.g., speech usage in open-mike mode
- Multimodal interaction will not work for all users from the start
- User adaptation can potentially yield significantly higher synergy and interface efficiency