THE UNIVERSITY OF BIRMINGHAM

# Multiple-Level Models for Multi-Modal Interaction

Martin Russell[1], Antje S. Meyer[2], Stephen Cox[3], Alan Wing[2]
[1]School of Engineering, University of Birmingham
[2]School of Psychology, University of Birmingham
[3]School of Computing, University of East Anglia

# Outline of talk

- Motivation for multi-modal interaction
- Multiple-level representations to explain variability
- Multiple-level representations to integrate modalities
- Issues in combining modalities
- Example: speech and gaze
- Proposed research
- Conclusions

# Motivation

- Linguistic utterances rarely unambiguous, but communication succeeds
  - Shared world knowledge
  - Common discourse model
  - Speech augmented with eye-gaze and gesture

# Psycholinguistic perspective

- In psycholinguistic theories the processes of retrieving and combining words are far better described than the processes of using world and discourse knowledge, eye gaze or gestures

# Computational perspective

- *Automatic* spoken language processing lacks knowledge and theory to explain ambiguity
  - Assumes direct relationship between word sequences and acoustic signals
  - Variability treated as noise
- No established framework to accommodate complimentary modalities

# Challenges

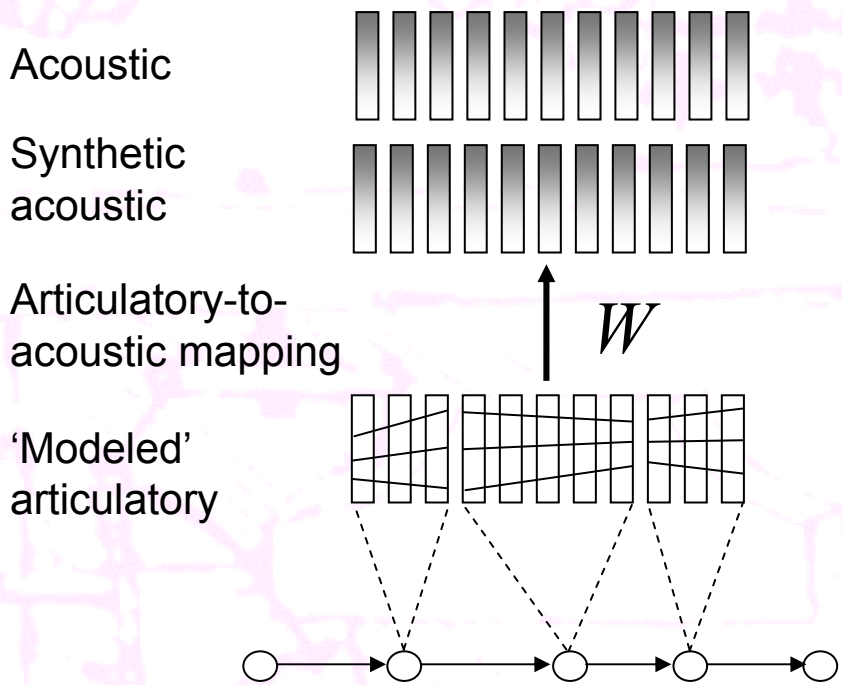
- Psycholinguistics needs:
  - Better understanding of how speakers and listeners use eye gaze and gesture to augment the speech signal
- Computational spoken language processing needs:
  - Better treatment of variability in spoken language
  - Better frameworks for augmenting speech with other modalities
- Both need fruitful interaction between psycholinguistics and computational spoken language processing

# Example: acoustic variability

- Sources of acoustic variability not naturally characterised in the acoustic domain:
  - Speech dynamics
  - Individual speaker differences
  - Speaking styles
  - …

THE UNIVERSITY OF BIRMINGHAM

# A model of acoustic variability




- Introduce intermediate, 'articulatory' layer
- Speech dynamics modelled as trajectory in this layer
- Trajectory mapped into acoustic space
- Probabilities calculated in acoustic space

# Combining modalities

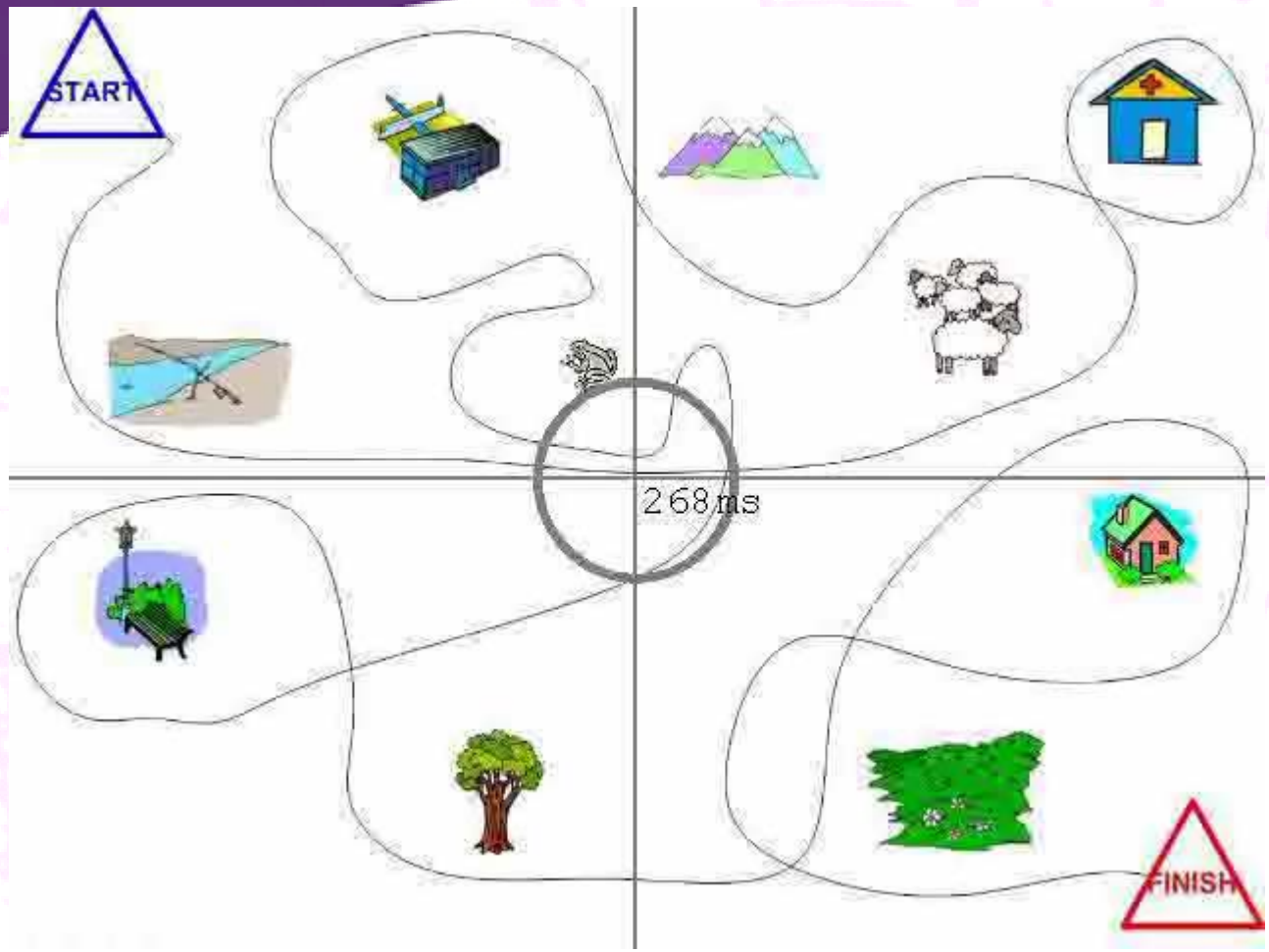
- Examples:
  - Lip-shape correlates with speech at the acoustic level…
  - … but this is not the case in general
  - Correlation between speech and eye-movement (when it exists) likely to be at conceptual level

# Multiple-level models

- Different levels of representation needed:
  - To model causes of variability in speech
  - To capture relationship between speech and other modalities
- Candidate formalisms already exist:
  - Graphical models,
  - Bayesian networks,
  - layered HMMs
  - …

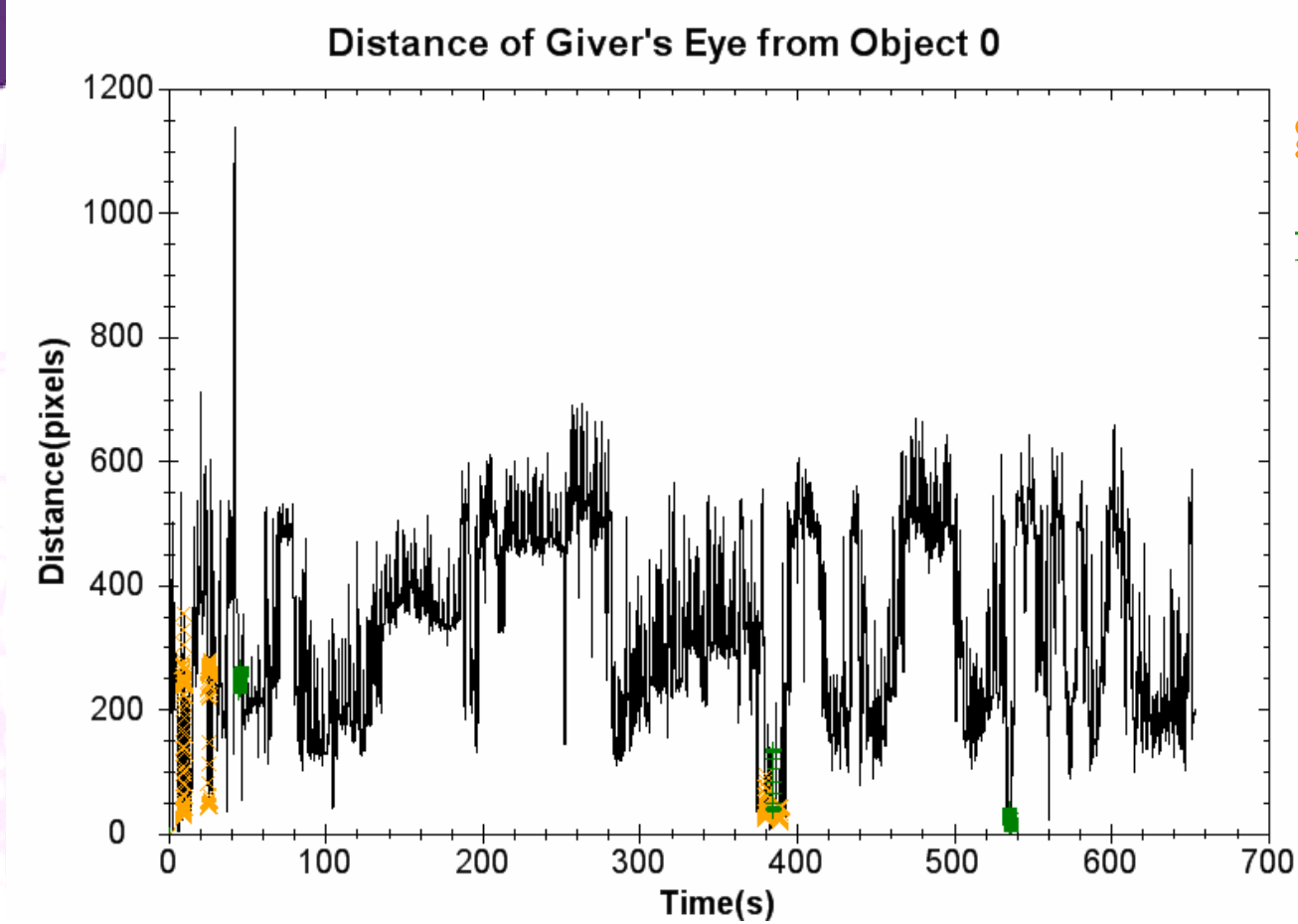
THE UNIVERSITY OF BIRMINGHAM

# Example: speech and gaze



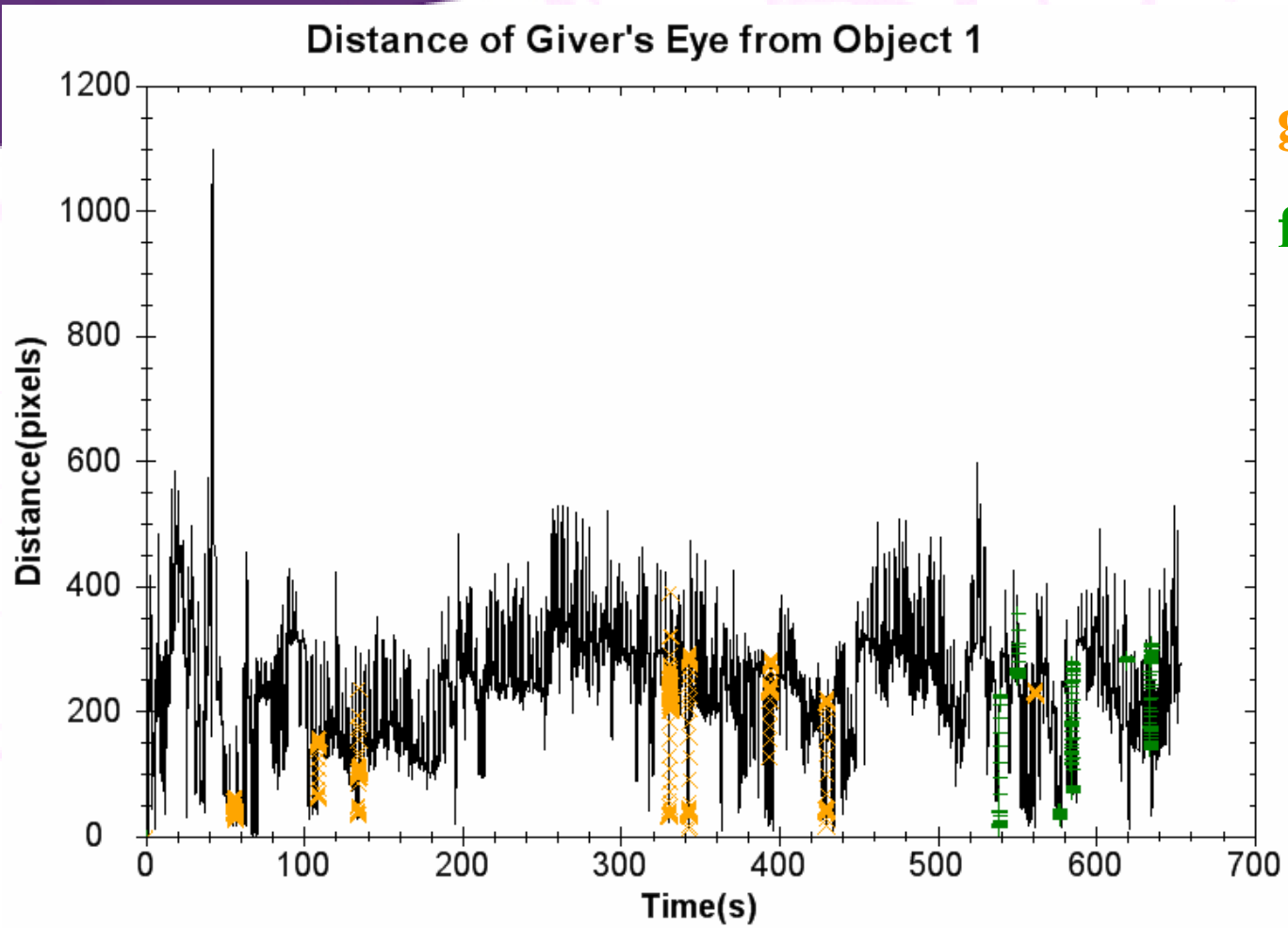
THE UNIVERSITY OF BIRMINGHAM

# Results from ‘map task’ experiment



THE UNIVERSITY OF BIRMINGHAM

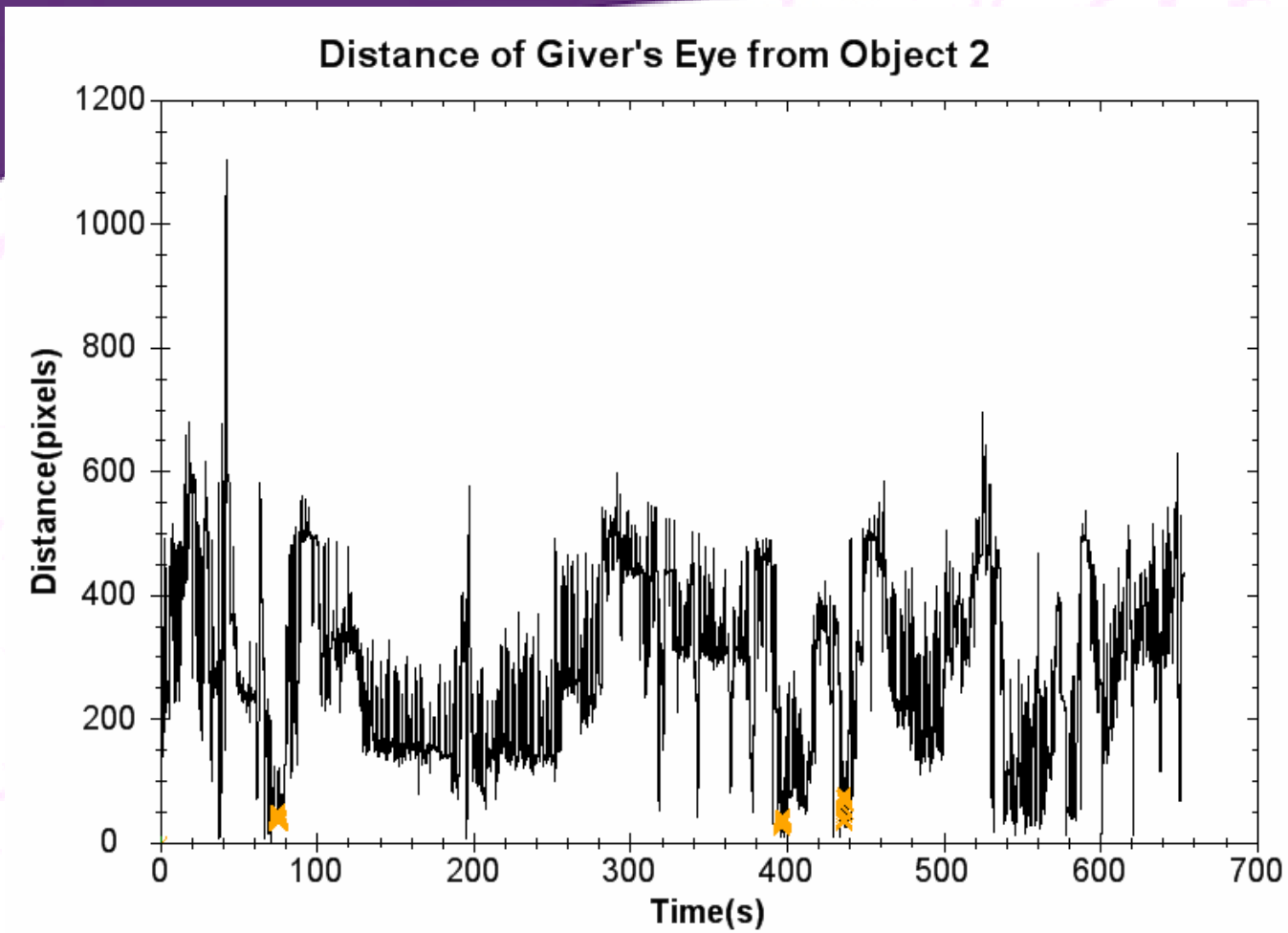
# Results from map task




**giver**

**follower**

# Results from map task

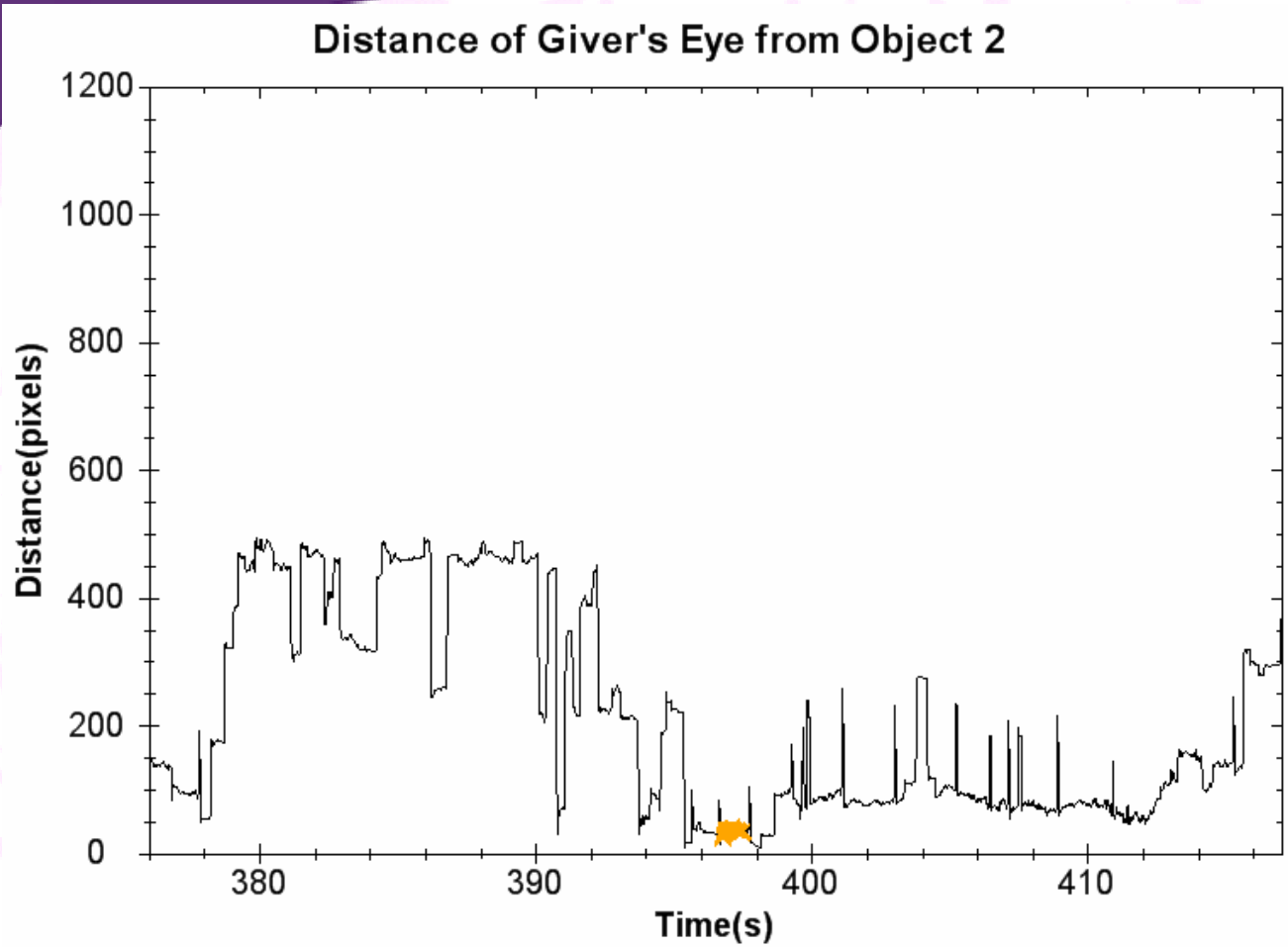



**giver**

**follower**

THE UNIVERSITY OF BIRMINGHAM

# Results from map task




**giver**

**follower**

THE UNIVERSITY OF BIRMINGHAM

# Object naming

4

gaze

lag

speech onset

time

*Planning to phonological level*

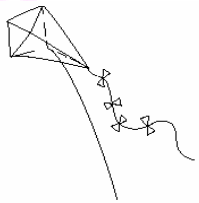
*Phonetic, articulatory planning*

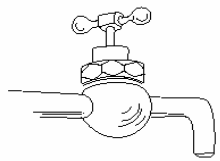
*PLUS advanced planning for next object*

*From ESRC Meyer, Wheeldon*

THE UNIVERSITY OF BIRMINGHAM

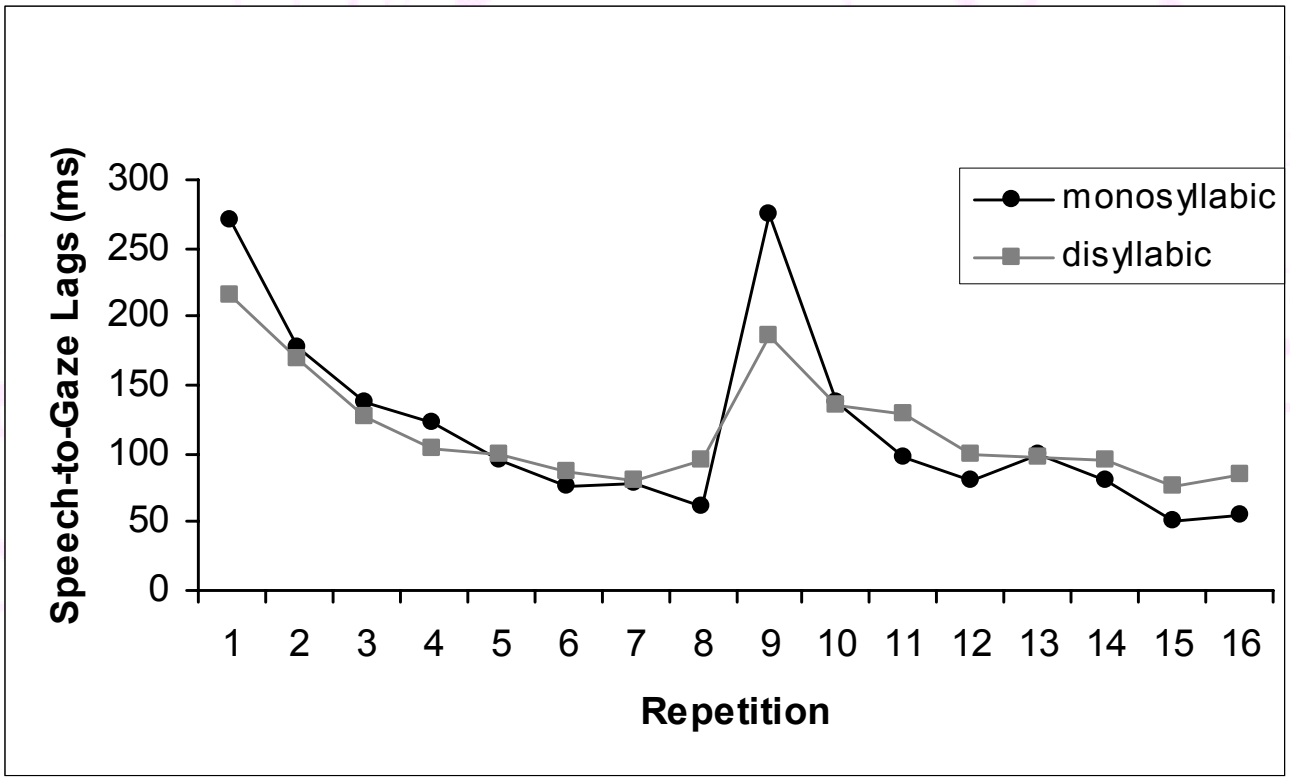
# Object naming





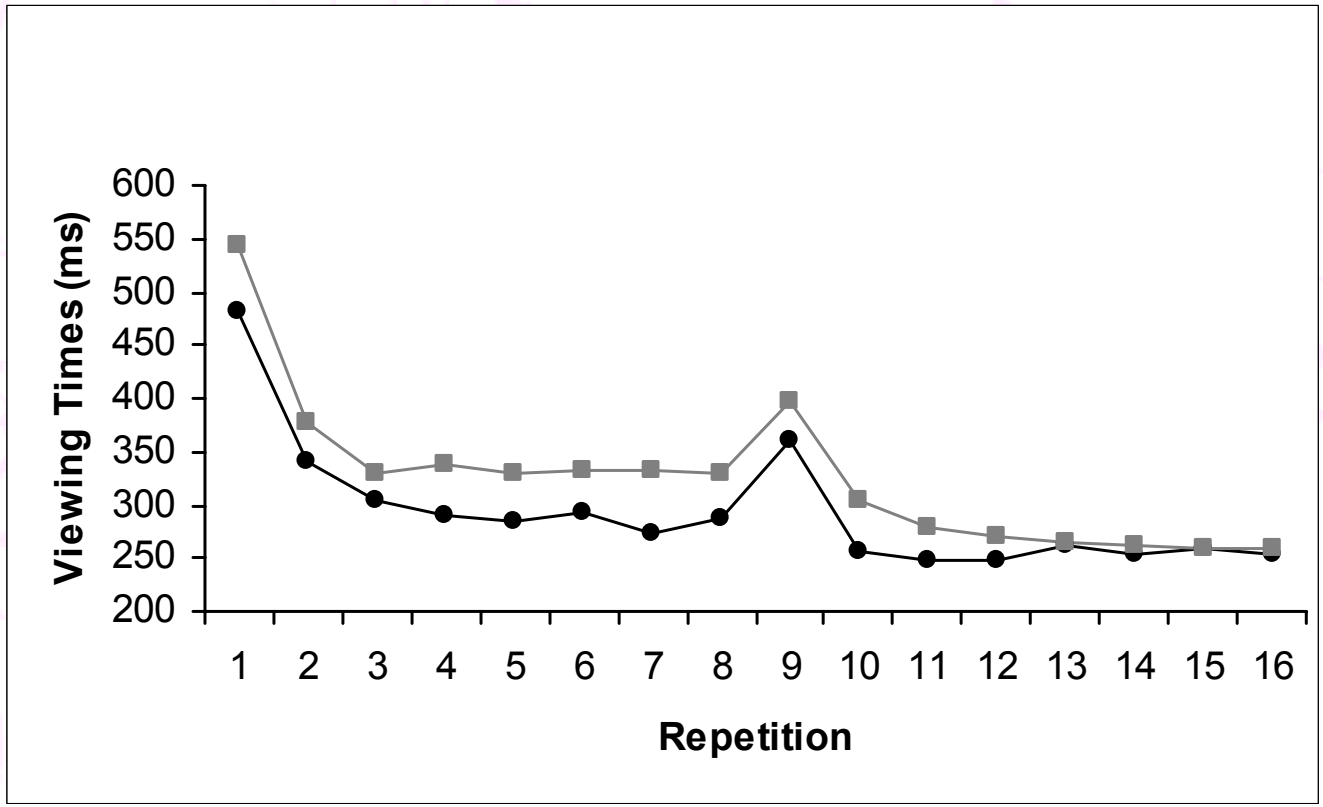
# Lessons from psychology

- Gaze-to-speech lags



# More lessons…

- Gaze duration



# Speech and gaze

- In general, a speaker who looks at an object might:
  a) Name the object,
  b) Say something about the object
  c) Say something about a different topic altogether
  d) Say nothing at all
- There will be a delay (200-300ms for object naming) between finishing looking at an object and talking about it
- The delay will be less if the object was discussed previously

# Speech and gaze (continued)

- Alternatively, gaze might provide an important cue for classifying the ‘state’ of a communication (e.g. meeting)
  - Monologue (all eyes on one subject)
  - Discussion (eyes move between subjects)

# Proposed research

- **Goal**: Improved understanding of user goals and communication states through integration of speech, gaze and gesture
- Integrated, multi-disciplinary project, involving psycholinguistics, speech and language processing, and mathematical modeling

# Proposed research (1)

- Experimental study of speech, gaze and gesture in referential communication and matching tasks, to determine:
  - How speakers' and listeners' gaze are coordinated spatially and in time
  - Functional significance of eye gaze and gesture information (by allowing or preventing mutual eye contact between the interlocutors)
  - Importance of temporal co-ordination of speaker and listener gaze

# Proposed research (2)

- Development of multiple-level computer models for integration of speech, gaze and gesture, for
  - Improved understanding of user goals
  - Improved classification of communication states (meeting actions)

# Summary

- Speech in multi-modal interfaces
- Multiple-level models for:
  - Characterising variability within a modality
  - Characterising relationships between modalities
- Proposal for collaborative research in psycholinguistics and speech technology

THE UNIVERSITY OF BIRMINGHAM

# CETaDL meeting room