Multiple-Level Models for Multi-Modal Interaction

Martin Russell¹, Antje S. Meyer², Stephen Cox³, Alan Wing²
¹School of Engineering, University of Birmingham
²School of Psychology, University of Birmingham
³School 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
  - …
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
  – …
Example: speech and gaze
Results from ‘map task’ experiment

Distance of Giver's Eye from Object 0

Giver

Follower
Results from map task

Distance of Giver's Eye from Object 1

- **giver**
- **follower**
Results from map task

Distance of Giver's Eye from Object 2

giver
follower
Results from map task

Distance of Giver's Eye from Object 2

Spoken Language and HCI Grand Challenge: slide 15
Object naming

From ESRC Meyer, Wheeldon
Object naming
Lessons from psychology

- Gaze-to-speech lags

![Graph showing speech-to-gaze lags for monosyllabic and disyllabic words over repetition.](image)
More lessons…

- Gaze duration

![Graph showing viewing times (ms) across repetitions](image-url)
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
CETaDL meeting room