Use of Visual Information in Experimental End-of-Speech Detection

R.J.J.H. van Son, Wiencke Wesseling, and Louis C.W. Pols

ACLC/IFA, University of Amsterdam, The Netherlands
R.J.J.H.vanSon@uva.nl

OAP, 20 June 2008
Introduction: Visual cues in conversations

Quantify the use of visual cues in (experimental) conversations

- Humans use visual cues, eg, lips, gaze, brows, head, hands
- Interesting for understanding language and for technical use
- Studying the processing of "conversations" in the lab
- When do human subjects use visual cues and how does it help them
- Method: Shadowing pre-recorded dialogs with minimal responses
Stimuli

- 20 informal dialogs of 900 s each, 5 hours total
- Informal and unrestricted dialogs (lively speech)
- 69 kWords, 13669 utterances, 5752 Turn Switches (simplified)
- Transcribed, utterance aligned
- Annotated for dialog “function” and Gaze direction
Stimuli: Examples

Example frame of stimuli

Videos were synchronized frame-by-frame
Stimuli: Presentation

Concatenate 12 fragments of 5 minutes (310 s) each
- 2 minute practise fragment (Audio + Video)
- 2 minute pause
- Repeat four times:
  - AV: Audio + Video
  - A: Audio only
  - V: Video only
- 2 minute pauses after every four fragments (20 minutes)
- Fragments were pseudo-randomized
- *Never* present a subject two fragments from the same dialog
30 Subjects were asked to “shadow” the dialog

- Subjects should utter minimal responses whenever they “felt like it”
- Record responses with laryngograph, segment automatically
- RT defined with respect to nearest utterance end
- Works surprisingly well, over 10,000 responses
- Response delays comparable to turn-switch delays
- RT experiment with Real, Correct, and Spurious, Error, responses
- Model Spurious responses with Randomized responses
Reaction-Times

Random Walk to threshold for Central (decision) component

- Perceptual and Motor: deterministic response-time $t_0 = t_p + t_m$
- Central decision making component: integration-time $\tau = \frac{1}{\alpha}$
- $RT \sim \tau + t_0$
- $RT$ Variance $\sim \tau^3$

Results: Distribution of Turn and Response delays

A mixture of *Spurious* and *Correct* responses

- **Turns**: Original speaker turns switch delays
- **AV**: audio-visual, **A**: audio-only, and **V**: visual-only
- **Randomized** delays model *Spurious* responses
Results: Separating *Correct* and *Spurious* responses

![Graph showing distributions of responses and delays.](image)

**Expectation Maximization with one PDF fixed**

- Filled symbols: Estimated Correct delays (% correct)
- Open symbols: Randomized delays
Results: Experimental values

Three interesting values

- Fraction of Correct responses $\Rightarrow$ Precision
- Estimated Average Response delay $\Rightarrow$ RT
- Estimated variance $\Rightarrow \tau^3$ or “cognitive load”
- *But*: Statistics are problematic

Visual cues: Gaze

Annotated Gaze direction of the speaker is a correlate of visual attention

- Gaze: Speaker starts to look towards the listener
- Nogaze: Other utterances
Results: Gaze versus Nogaze utterances

High numbers of spurious responses, overlapping delay distributions

- Filled symbols: Estimated Correct delays
- Open symbols: Randomized delays
Results: Estimated Response Counts

Correct versus Spurious responses, as Signal-to-Noise ratio (dB)

- Left: High numbers of spurious responses
- Right: Gaze has better SNR, except for Audio-only (A)
Results: Estimated Reaction Times

Estimated mean RT

- **Gaze** longer than **Nogaze** even without visual cues (A)
- Compounding factors? (eg, utterance length)
Results: Estimated Standard Deviation and $\tau_{\text{Gaze}} / \tau_{\text{Nogaze}}$

SD: Gaze versus Nogaze

Cognitive “load” increases when visual Gaze cues are used

- AV, V have higher Sd and integration constant $\tau$ with Gaze ($\sim 20\%$)
- Original Turns have high processing load, not affected by Gaze
Visual cues are sufficient and helpful for *experimental* TRP detection

- Visual cues almost sufficient for TRP detection ($V \sim 25\%$ Correct)
- Visual information without Gaze has little effect ($AV \sim A$)
- Visual correlates of attention, eg, Gaze:
  - improve precision ($\sim 2dB$ SNR, $V \sim A$)
  - increase cognitive load ($\tau \sim +20\%$)
- Original dialog turn switches (Turns)
  - induce a higher cognitive load
  - unaffected by Gaze
  - visual attention only one factor among many
Thank You