

Functional Evaluation of Multimodal Interfaces: An Experiment with Talking Heads

Emiel Kraemer (Tilburg University)

Zsófia Ruttkay (CWI)

Marc Swerts (IPO/CNTS)

Wieger Wesselink (IPO)

Motivation

Evaluation of multimodal interfaces (e.g., Embodied Conversational Agents, Talking Heads) can be difficult.

- Some standard techniques (e.g., thinking aloud) do not apply.
- Often more than one correct way to express information.

Alternative: focus on function

Case study: *Pitch, Eyebrows and the Perception of Focus*

Outline

1. Introduction
2. Materials
3. Experimental set-up
4. Procedure
5. Research questions
6. Results
7. Future research

1. Introduction

- In Germanic languages (Dutch, English, ...) pitch accents can indicate which information is in *focus* (contrastive or new).
- Rapid eyebrow movements can perform a similar function (e.g., Birdwhistell 1970, Condon 1976).
- Try uttering: **blue** square.
- Morgan (1953), Bolinger (1985): *metaphor of up and down*.
- On the other hand: speakers do more with pitch than with their eyebrows (e.g., Cavé et al. 1996).

No apparant consensus on animation of eyebrows in ECAs.

Pelachaud et al. (1996) Affect dependent

(julia prefers)_{theme} (popcorn)_{rheme}

Cassell et al. (2001) “flash” on object reference in rheme.

(julia prefers)_{theme} (popcorn)_{rheme}

Remarks:

- No evaluation.
- No insight in relative function of eyebrow or pitch.

2. Materials

- 2D Talking Head (CharToon, CWI): “blauw vierkant” (blue square).
- Six different voices (4 human, 2 synthetic).
- Four humans collected in an earlier dialogue game experiment (Krahmer & Swerts 2001).
- During game subject had to describe different coloured geometrical figures (including a blue square) on a stack of cards.
- Order of stack was systematically varied: target descriptions in different contexts.

- Define: A property is
 - **contrastive (c)** if previous object had a different value.
 - **given (g)** if previous object had the same value.
 - **in focus** if it is contrastive.
- Distribution analysis: **c** is accented, **g** deaccented.
- Human speakers used different intonation contours: two high-ending (H%) and two low ending (L%).
- Synthetic voices copy the 2 contours (“prosody transplantation”).
- Two rapid eyebrow versions: either on 1st word (“blauw”; notation $\hat{o}o$) or 2nd word (“vierkant”; $o\hat{o}$).
- Duration 300ms (compare humans: on average 375ms).

3. Experimental set-up

“Reconstructing Dialogue History” (Swerts & Krahmer 2001):
subjects watch and listen to Talking Head uttering “blauw vierkant”
(blue square).

Task: decide whether the *previously described* object:

1. a red square [cg/focus on adjective]
2. a blue triangle [gc/focus on noun]
3. a red triangle [cc/all focus]



4. Procedure

- 25 "prosodically naive" subjects.
- Individual, self-paced, forced choice.
- Training session (3 stimuli): no feedback.
- 36 stimuli (3 contexts × 2 eyebrow version × 6 voices).
- Two different random orders.

5. Research questions

1. Which features contribute to the perception of focus?
2. What happens when pitch accent and eyebrow movement do not match?
3. What, if any, is the influence of the intonation contour?
4. What, if any, is the influence of the voice (synthetic vs. human)?

6. Results

		classified as			
		cc	gc	cg	<i>total</i>
context	$\hat{c}c$	64	41	45	150
	$c\hat{c}$	59	70	21	150
	$\hat{g}c$	34	91	25	150
	$g\hat{c}$	33	90	27	150
	$\hat{c}g$	16	22	112	150
	$c\hat{g}$	16	30	104	150

In sum

- Speech is dominant for the perception of focus.
- Subjects are best in reconstructing the dialogue context in the single focus cases.
- Eyebrows have a significant effect, but only for the prosodic all focus cases **CC**.
- There is an effect of intonation contour:
 - For high-ending speakers (H%) there is a stronger overall preference for **gc** (focus on noun).
 - For low-ending speakers (L%) the effect of eyebrows is stronger.
- There is no effect of voice.

Since auditory cues are dominant:

- Results basically confirm with speech-only results of Swerts & Kraemer (2001), but here a bit more confusion. May be due to:
 - differences in subject population, or
 - visual cues distract subjects.
- Mismatches often go unnoticed . . .
- . . . but various subjects found them confusing.
- It might be that associating eyebrow movements with unfocussed information “gives listeners a harder time” (cf. Terken & Noolboom 1987).

7. Future Research

- Redo the experiment with a Romance language (Italian).
- Speech has already been collected (“triangolo nero”).
- Always a double (downstepped) accent structure (cf. Dutch **cc**).
- Speech-only dialogue reconstruction is impossible.
- Hypothesis: eyebrows contribute more than for Dutch.
- (Nota Bene: there is some evidence that Italian listeners profit more from gestures than English ones, Graham & Argyle 1975)