Coordination as a simplifying factor in human-machine communicative interaction

Eric Vatikiotis-Bateson
(COGS and Linguistics)

20 September 2016

Abstract for Speech Robotics session, Acoustical Society of America (28 November, 2016, Honolulu, HI)

Spoken communication has traditionally been treated as a problem of sending and receiving signals containing elements whose sequential organization specifies meaningful content that is the same for both sender and receiver. Unfortunately, successful communication depends on a host of contextualizing factors that are either not present or impossible to identify in short signal streams (by either humans or machines). This presentation focuses on the crucial role of one such factor, the necessary coordination between perceiver and producer, to suggest that human communicative interaction can be conceived more simply than it has been and thereby improve the likelihood of successful human-machine interaction. For example, humans depend on physiological and cognitive coordination for successful interaction, particularly in signal parsing, alignment, and error correction. However, coordination is loosely constrained under most conditions and can be achieved in human-machine interaction without imbuing machines with human attributes; for example, shared attention. In many other cases, machines outperform humans; for example, in processing multimodal speech signals under adverse acoustic conditions to disambiguate the labial viseme, /p,b,m/. In sum, these and other factors provide texture and coherence, rather than daunting complexity, to our efforts to understand spoken communication.

Coordination between perceiver and producer

- Humans depend on coordination for successful communicative interaction, particularly in signal parsing, alignment, and error correction.
- Coordination is
  - physiological and cognitive
  - loosely constrained under most conditions
- Human-machine interaction (HMI/HCI) does not require machines to have "human" attributes; …
  - Shared attention will do, as it provides feedback about common orientation without regard for what's going on internally.
  - Thus, human and/or machine interactants are free to process the world in ways best suited to their internal structure.
    - Machines easily outperform humans in processing multimodal speech signals under certain adverse acoustic conditions (disambiguating the labial viseme, /p,b,m/), but not necessarily others (distinguishing voices at a cocktail party).

Is this argument reasonable? What are some alternatives?

About the readings

- Ibbotson and Tomasello's rebuttal of Chomsky's universal grammar
  - Humans share many physiological and cognitive traits; what plays out in language and communication depends on these traits **and** on learning.
  - Predictability of communicative intent and meaning is context dependent.
- Brian Scassellati's PhD thesis (MIT) on robot *theory of mind*
  - An elegant attempt to imbue COG (the AI Lab's humanoid robot) with the supposedly human communicative goal of "correctly"attributing beliefs, goals, and percepts to other people.
  - COG achieves what appears to be *shared attention*, but what does it actually know or need to know about communicative intent?
- Cynthia Breazeal's PhD thesis (MIT) on *sociable machines*
  - *Kismet* is a small, cuddly robot, wired to generate caricatures of *infant-toddler* social behavior.
  - It's small size, cuteness, and ability to adapt its output to match ("mimic") the behavior of adult interactants all enhance its acceptability as a communicative partner.

COG and Kismet samples

provided by Rod Brooks, former Director of the MIT AI-Lab

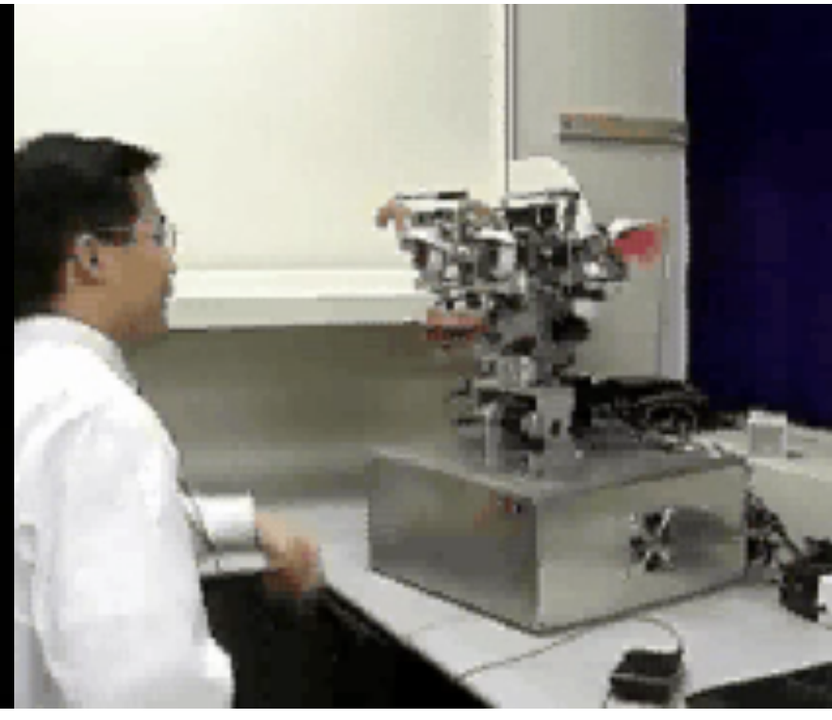
COG

Kismet

Affect

Social amplification

Turn-taking



My take

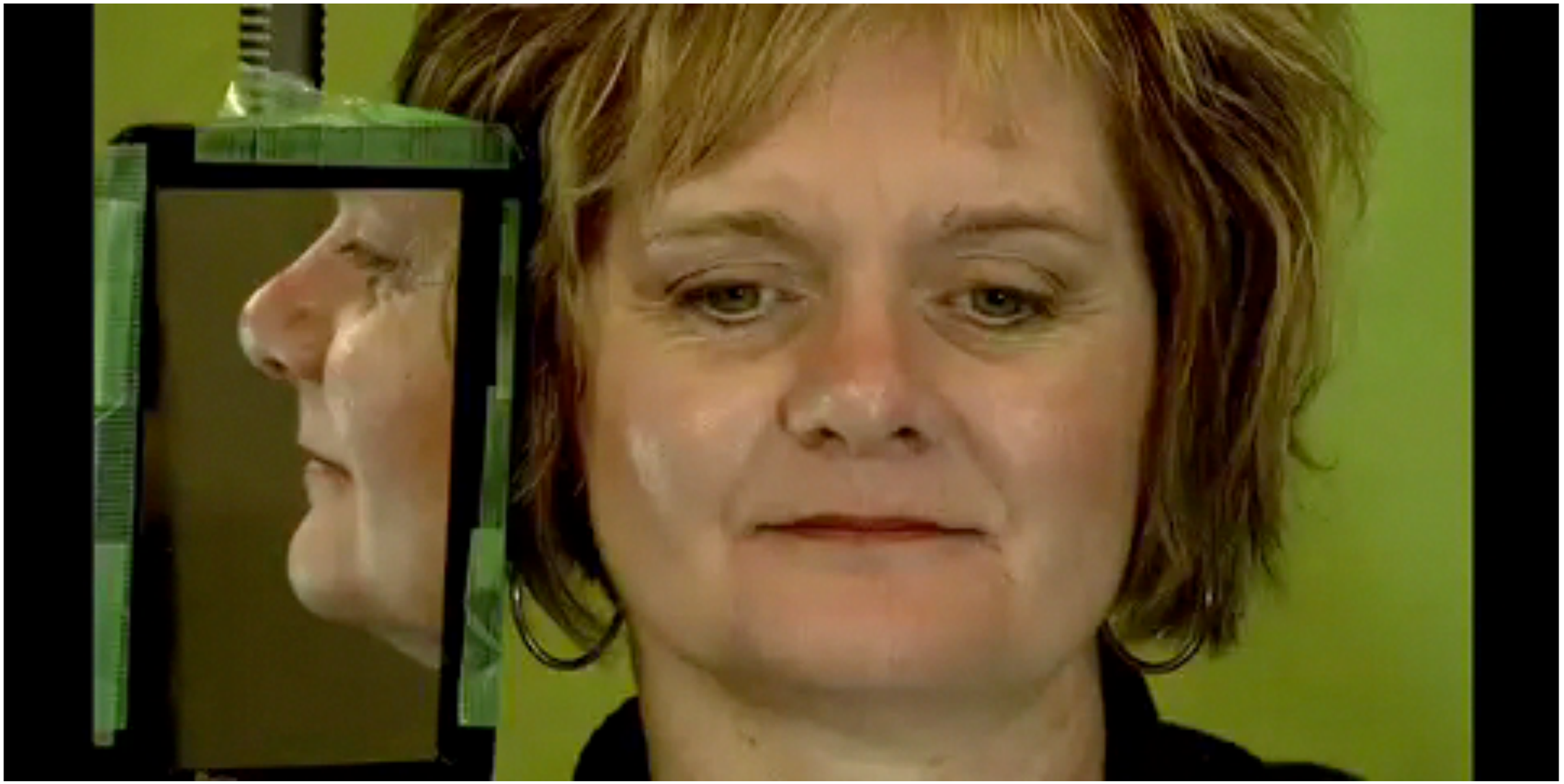
- How human communicative interaction is packaged is more import than its content.
- Adaptive coordination of behavioral patterns is sufficient for successful social interaction and does not require shared intentional goals.
- Machines can achieve this in human-machine interaction, but …

Machines can outperform humans

from *Abel et al. 2011*

Sample videos

talker 1



Sample videos

talker 2

Stimulus pair for discrimination

Results

- **Discrimination** (Chi-square) of labials in front and profile
- huge individual perceiver bias
- Problems: small *N* (77 pairs), modulated babble tract, and VCVs

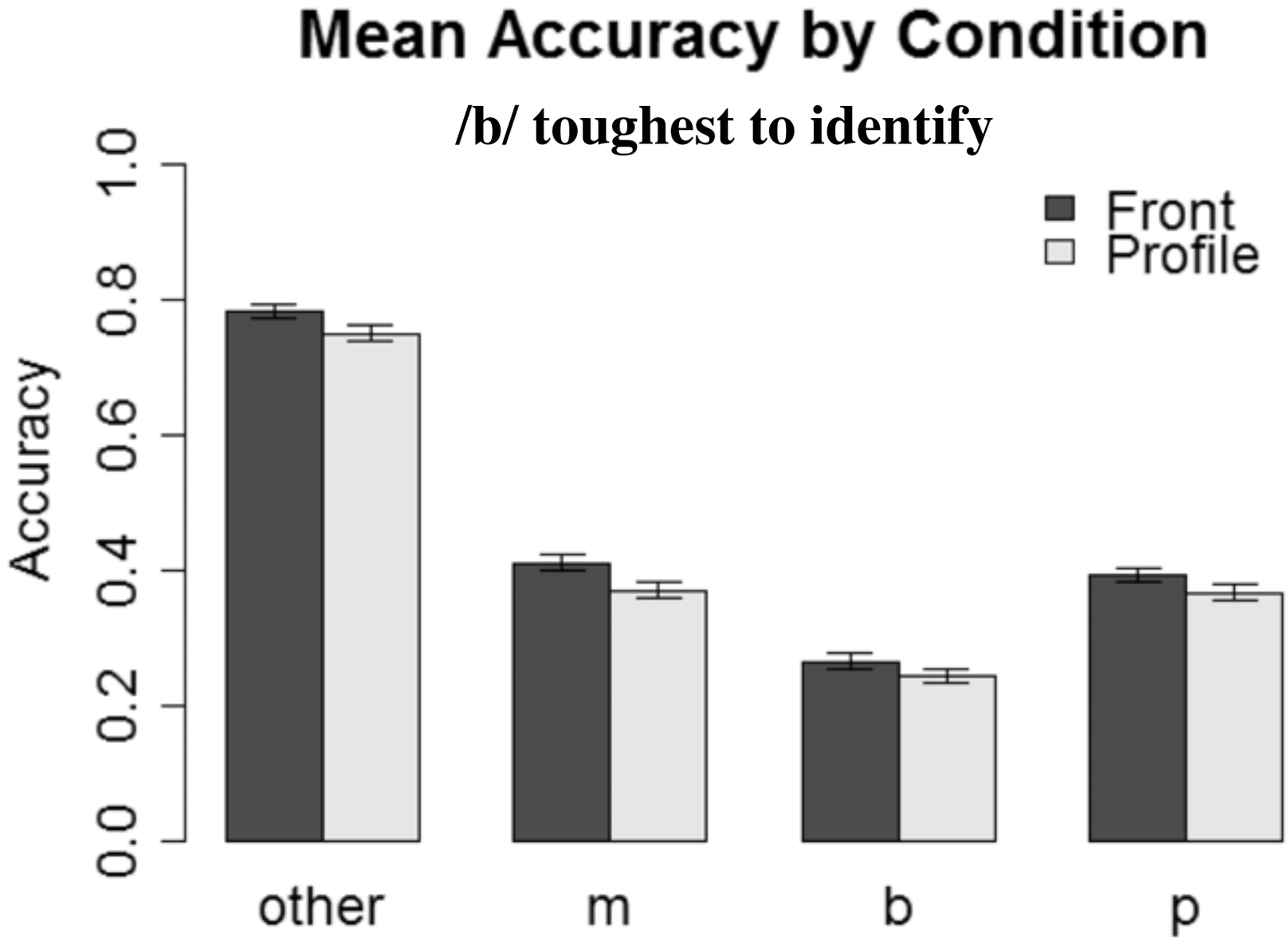
Stimulus for identification: front

N = 396

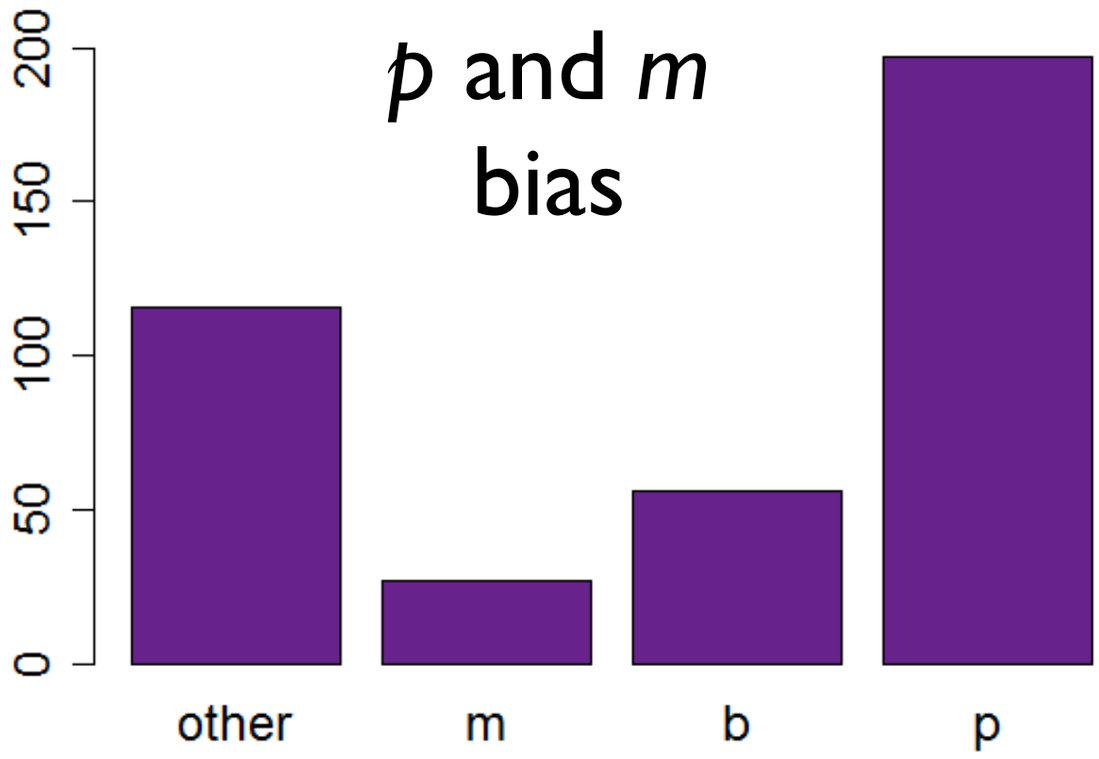
Stimulus for identification: profile

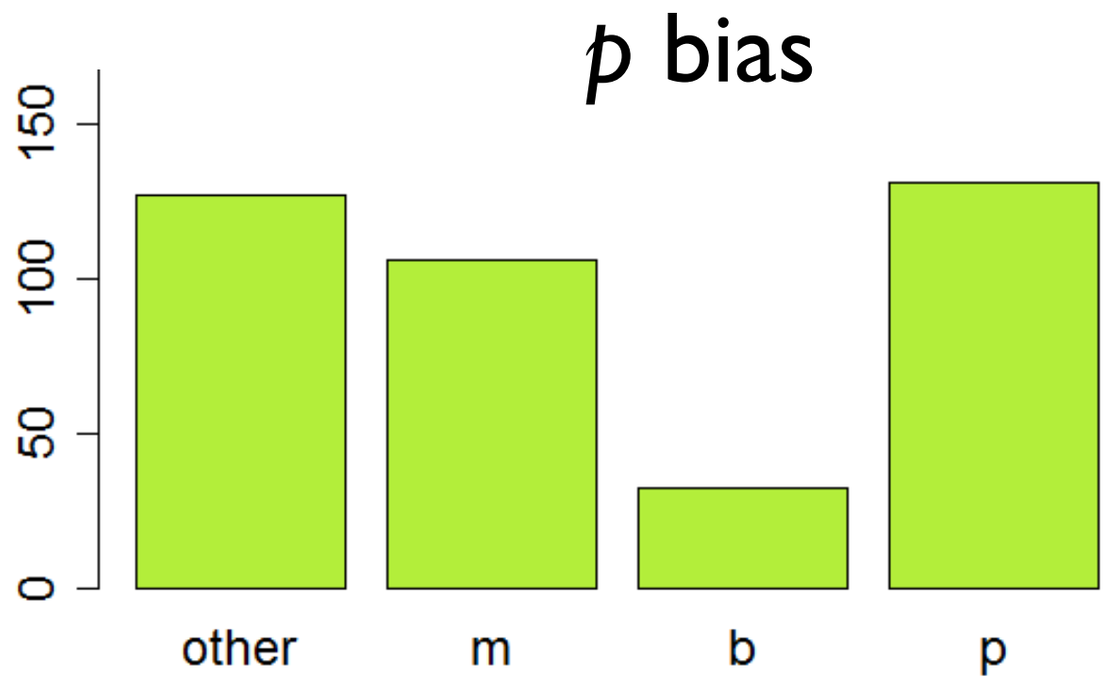
N = 380

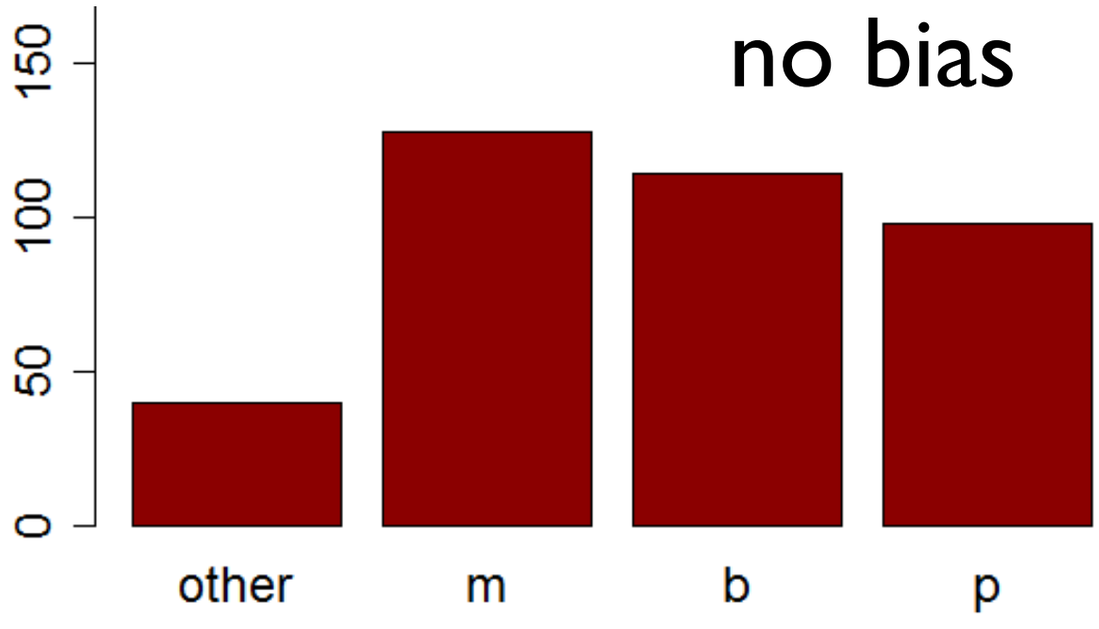
Identification study



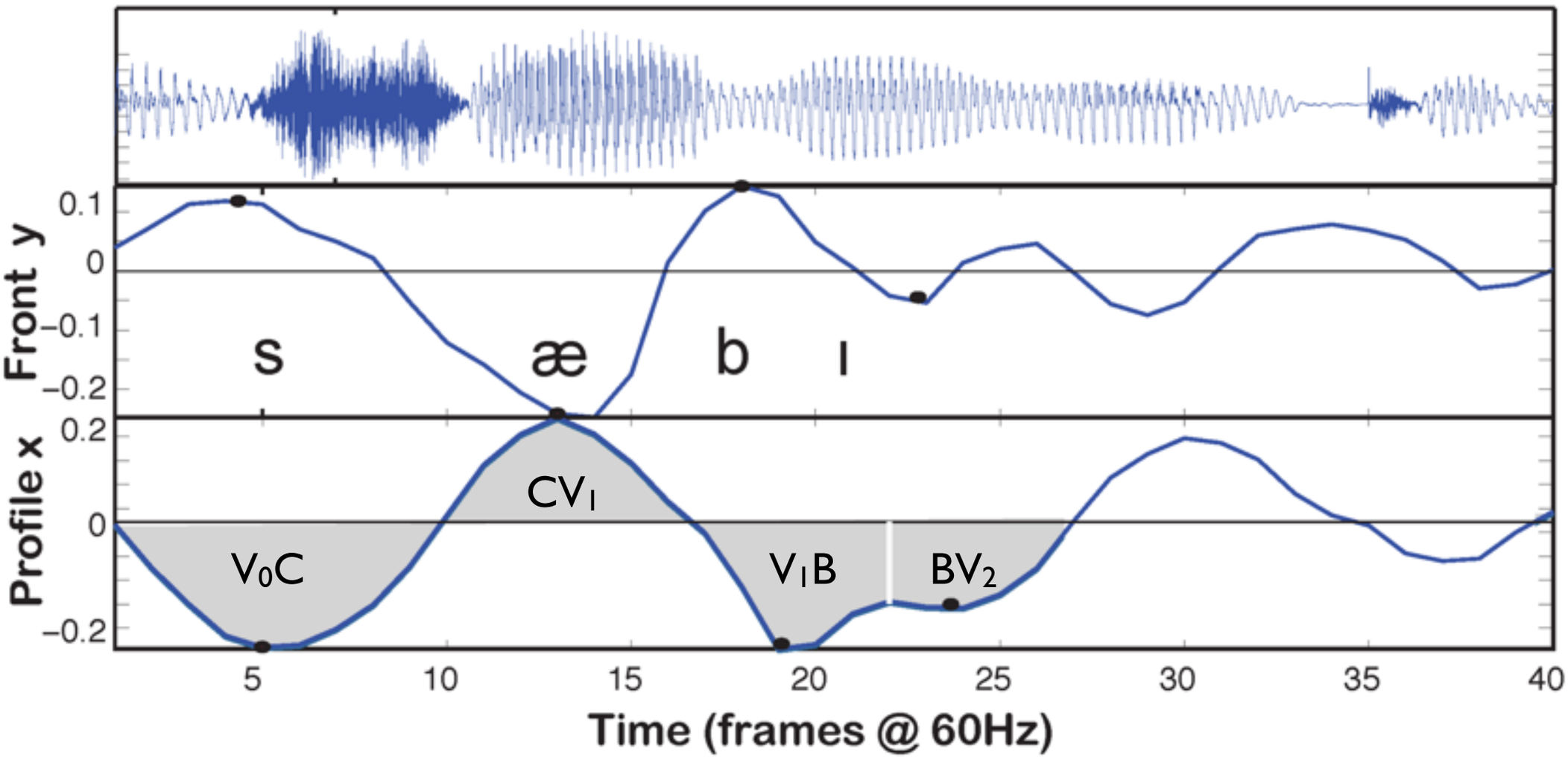
Individual response strategies







Motion peaks and AUC



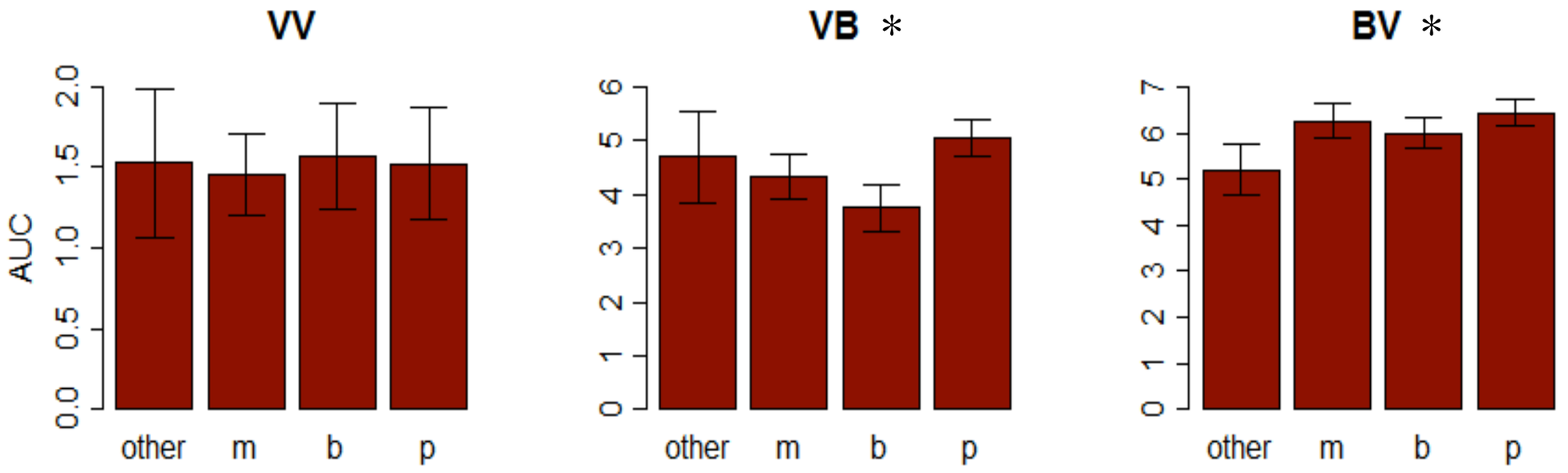
Peak motion results:

- Sparse reliable differences between *p, b, m* (ANOVA):
- Few BV_2, fewer V_1B, vertical motion, front view, always lower-face ROI
- No V_0C or CV_1

Area Under the Curve (AUC)

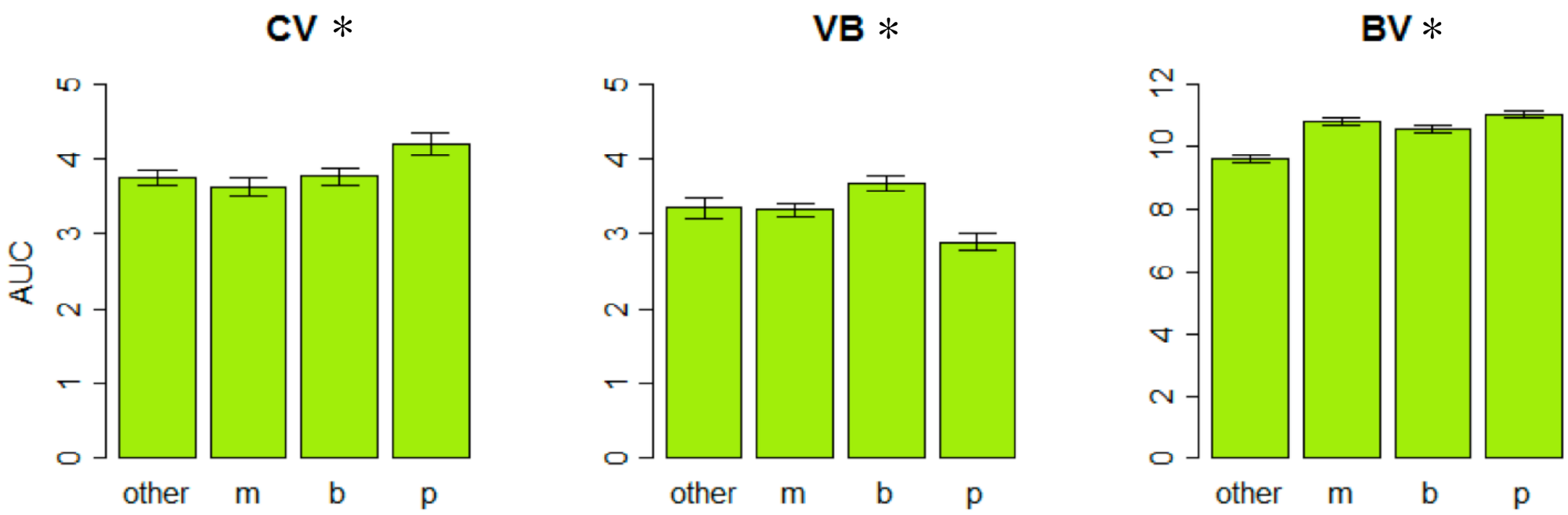
S1 Vertical Profile

(Small N)



S2 Vertical Front

(Larger N)



Summary of labials study

- Perception does not fully exploit production.
  - ➡ Machine analysis should not take its cue from perception.

- Production and perception do not have the same task
  - Production acts collectively.
  - Perception achieves decomposition and selection via attention.

- Problems:
  - Running speech reduces both articulatory and acoustic attributes of segment identity – no running speech in this study.
  - Simple measure of total face motion – surprisingly effective – but not necessarily the most sensitive measure of orofacial motion.