Combining manual and automated gesture annotation: a case study

Victoria Reshetnikova[1], Roy S. Hessels[2] & Aoju Chen[1]
[1] Institute for Language Sciences, Utrecht University, Utrecht, the Netherlands
[2] Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, the Netherlands

Presenting author:
Victoria Reshetnikova
v.reshetnikova@uu.nl

Problem statement

Studying multimodal interaction requires analysing both auditory and visual information. There are tools available to annotate speech prosody both automatically (e.g., AASP for Dutch: Hu et al., 2020 and AuToBI for English: Rosenberg, 2010) and manually (e.g., RPT: Cole & Shattuck-Hufnagel, 2016) in a swift manner. Fewer instruments exist for gesture annotation.

What aspects of gesture annotation can be automated?

Where is the human annotator necessary?

Examples from our recent study on variation in gestures accompanying intonational phrase (IP) boundaries in infant-mother interaction

A. Manual annotation: gesture type in context

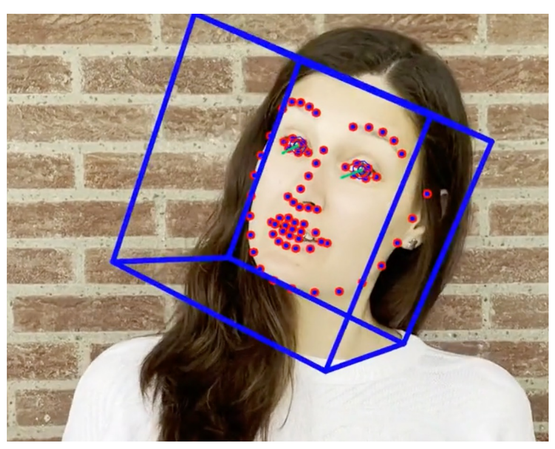
Beat gesture

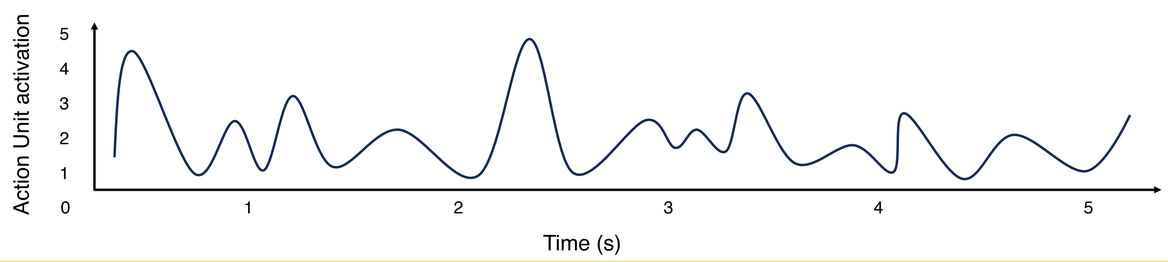
"… and **today**, I'm gonna tell you the Little Red Riding Hood story…"

Conventional gesture

"… and **hurray**, Grandma and Little Red Riding Hood are free…"

B. Automated annotation: intensity of eyebrow movements





C. Problematic situation 1: from movement to gesture?

Beat gesture

Not a gesture

"… and **today**, I'm gonna tell you the Little Red Riding Hood story…"

D. Problematic situation 2: articulators (partially) invisible

Articulators visible

Articulators (partially) visible

"… and **hurray**, Grandma and Little Red Riding Hood are free…"

Conclusion & points for discussion

Trade-off between manual and automated gesture annotation stems from methodological constraints.

- Machine-learning techniques may be useful, but what about explainability?
- Where does manual annotation remain the de facto standard?

Video examples

References
•Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-10). IEEE.
•Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).
•Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. Laboratory Phonology, 7(1).
•Hu, N., Janssen, B., Hansen, J., Gussenhoven, C., & Chen, A. J. (2020). Automatic analysis of speech prosody in Dutch. In Proceedings of Interspeech 2020 (pp. 155-159).
•Reshetnikova,V., Hessels, R., & Chen, A. (under revision). Variation in gestural input related to prosodic phrasing in infant-directed interaction. Language and Cognition.
•Rohrer, P. L., Vilà-Giménez, I., Florit-Pons, J., Gurrado, G., Gibert, N. E., Ren, P., Shattuck-Hufnagel, S., Prieto, P. (2021, February 24). The MultiModal MultiDimensional (M3D) labeling system. https://doi.org/10.17605/OSF.IO/ANKDX
•Rosenberg, A. (2010). Autobi-a tool for automatic tobi annotation. In Eleventh Annual Conference of the International Speech Communication Association.