Combining Manual and Automated Gesture Annotation: a Case Study

Victoria Reshetnikova[1], *Roy Hessels*[2], *Aoju Chen*[1]

[1]Institute for Language Sciences, Utrecht University

[2]Experimental Psychology, Utrecht University

v.reshetnikova@uu.nl, r.s.hessels@uu.nl, aoju.chen@uu.nl

Studying multimodal communication implies studying both auditory and visual streams of information. One of the challenging steps in such research, for instance on gesture-speech interaction, is annotating audio-visual corpora. While there are tools that allow researchers to annotate aspects of speech prosody automatically (e.g., AASP for Dutch: Hu et al., 2020, and AuToBI for English: Rosenberg, 2010) or manually in a fast and reliable manner (e.g., RPT: Cole & Shattuck-Hufnagel, 2016), gesture annotation for describing gesture trajectory and phasing is traditionally done manually and, therefore, laborious. Thus, the following questions arise: What aspects of gesture annotation can be automated? Where is the human annotator necessary?

In the talk, we will address these questions in the light of a recent study of variation in gestures accompanying intonational phrase boundaries in infant-mother interactions (Reshetnikova et al., under revision). For the aims of this study, we recorded live interactions between nine Dutch-speaking mothers and their 5- to 9-month-old infants and focused on three aspects of eyebrow and hand gestures: type (i.e., whether a gesture was beat, deictic, iconic, metaphoric, or conventional), temporal alignment to intonational boundaries (i.e., how early before the intonational boundary a gesture peaked), and intensity (i.e., how intense the eyebrow movement was).

First, what aspects of gesture annotation can be automated? We have found that simple movement that was easy to operationalise and quantify could be automated such as eyebrow movement. Using the Facial Action Unit activation as estimated by OpenFace (Baltrušaitis, Robinson, & Morency, 2016), we were able to describe the variation in both temporal alignment between peaks in eyebrow movements and intonational boundaries and eyebrow gesture intensity within and between participants (see Figure 1).

The automation was also useful in a situation where human perception was biased. For instance, to describe eyebrow and hand gestures that were related to intonational boundaries, we needed to define whether a certain gesture was perceived as being associated with an intonational boundary. Inter-rater agreement was low (Cohen's kappa = -0.04) when this parameter was annotated manually. One potential explanation might be that prosodic features influenced the perception of gesture prominence (Krahmer & Swerts, 2007). It was not clear what the human annotator's strategy for annotating gesture association with an intonational boundary was. Therefore, we ended up operationalising gestures as being associated with a boundary when they occurred within the gesture unit.

Second, where is manual annotation necessary? In our study, aspects of hand movements, such as gesture units, peaks, and types, were fully annotated manually, following the MultiModal MultiDimensional (M3D) labeling system (Rohrer et al., 2021). Why did we opt for manual annotation over automated methods? For example, existing methods such as OpenPose (Cao et al., 2017) are able to locate important body key points (e.g. hand location) and estimating the physical movement is straightforward. However, to interpret those movements as being iconic, metaphoric, deictic, conventional, or beat gestures, manual annotation was necessary. While it would be possible to automate annotation of hand gesture units and peaks, gesture types are still to be labelled by a human annotator in the absence of substantial training sets for machine learning models. Besides, human annotators are good at adapting their strategies in problematic cases. For instance, in this study mothers were not explicitly instructed to remain at a fixed distance from the camera while interacting with their infants. Some mothers moved closer to the camera, and some hand movements were therefore out of the camera view. Such a set-up, uncontrolled for participant-camera distance, turned out to be problematic for OpenPose since it was not possible to estimate position of hands when they were "out of sight". The manual annotator was able to estimate the gesture trajectory, phasing, and type in borderline cases.

To sum up, there is a trade-off between using manual and automated gesture annotations in studies investigating gesture-speech interaction, stemming from methodological constraints inherent in both approaches, exemplified in our case study. More generally, we think the following questions ought to be discussed in the context of multimodal research: What is the current state-of-the-art in machine learning techniques for e.g. classification of gestures? How applicable are these techniques to the research practice in linguistics? How explainable and/or replicable are these techniques? Where are the largest gains in terms of progressing research on multimodal communication? And where might manual classification remain the de facto standard for the foreseeable future?

Index Terms: gesture-prosody interaction, gesture annotation, manual gesture annotation, automated gesture annotation

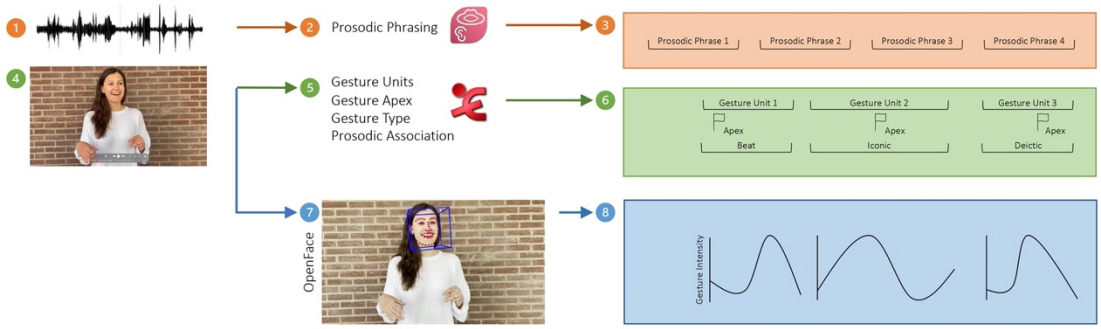

Figure 1. **From the video recording to the prosodic and gesture annotation.** 1) Audio input retrieved from the video recording. 2) Audio data are annotated for prosodic phrases in Praat. 3) Prosodic phrasing annotation output. 4) Video input. 5) Video data are annotated for hand gesture units, peak, type, and IP boundary association in ELAN. 6) Hand gesture annotation output. 7) Video data are analysed for eyebrow movement through OpenFace. 8) Annotation output of eyebrow movement.

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).

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of memory and language*, *57*(3), 396-414.

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*.