Creating Multimodal Corpora for Co-Speech Gesture Research

Walter Dych[1], Karee Garvin[2], and Kathryn Franich[2]

[1]Binghamton University, [2]Harvard University

Studies of the relationship between speech and co-speech gesture have the potential to inform many areas of linguistic inquiry, including phonetics/phonology, prosody, information structure, and semantics/pragmatics [1]. Multimodal corpora of conversational speech provide a rich context in which to explore the nuances of how gesture behaves across languages. Until recently, the gold standard in multimodal speech research has involved manual coding of co-speech gestures from video data. Coding of gestures by hand is time-consuming, and, following best practices, usually requires at least two coders for the establishment of inter-rater reliability [2]. While marker-based motion-capture technologies can be useful for avoiding pitfalls of manual annotation, such systems are often not available for the study of under-documented languages spoken in areas of the world where linguistics labs are not common. Here, we present a set of tools adapted for the automatic coding of co-speech gestures in simple video data. These tools have been developed based on a variety of data types from several different languages, including multiple varieties of English (US, Cameroonian, and Nigerian), as well as several under-documented Niger-Congo languages (Medʉmba, Kejom, and Igbo). Our data processing pipeline uses MediaPipe [3] markerless motion capture technology to track 2D or 3D movement of the articulators from video inputs to extract keypoints, as shown in Fig 1. Our toolkit allows for automatic annotation of gesture movement onset and offset, an interval that comprises the preparation, stroke, hold, and recovery of a gesture (Fig. 2), in addition to automatic apex annotation based on movement speed peaks or zeros (Fig. 3). We demonstrate how this method can generate apex annotations that closely correspond with manually coded apexes for multiple gesture types, including pointing gestures and bimanual beat gestures. We show that differences in manual vs. automated results typically reflect erroneous inclusion of non-gesture events (e.g. fidgets), and propose a set of next steps for fine-tuning the algorithm so as to exclude such non-gesture events. We then discuss challenges and potential solutions for tackling the range of more complex gestures that tend to occur in more naturalistic conversation, with the goal of developing methods which will be useful for a broader range of data types.

[1] Maricchiolo, Fridanna, De Dominicis, Stefano, Ganucci Cancellieri, Uberta, Di Conza, Angiola, Gnisci, Augusto and Bonaiuto, Marino. "109. Co-speech gestures: Structures and functions". Volume 2, edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Jana Bressem, Berlin, München, Boston: De Gruyter Mouton, 2014, pp. 1461-1473. https://doi.org/10.1515/9783110302028.1461

[2] MIT Speech Communication Group, Co-speech gesture coding manual. [Online]. Available: http: //scg.mit.edu/gesture/coding-manual.html.

[3] C. Lugaresi, J. Tang, H. Nash, et al. 2019. Mediapipe: A framework for building perception pipelines. DOI: 10.48550/ARXIV.1906.08172.