# COMPARING MANUAL VS. SEMI-AUTOMATED METHODS FOR THE CODING OF CO-SPEECH GESTURES

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

University of Delaware, Harvard University
wdych@udel.edu, garvinkaree@gmail.com, kfranich@fas.harvard.edu

## ABSTRACT

While motion capture is rapidly becoming the gold-standard for research on the phonetics of co-speech gesture and its relationship to speech, traditional marker-based motion capture technology is not always feasible, meaning researchers must code video data manually. We compare two methods for coding co-speech gestures of the hands and arms in video data of spontaneous speech: manual coding and semi-automated coding using OpenPose [1], a markerless motion capture software. We provide a comparison of the temporal alignment of gesture apexes based on video recordings of interviews with speakers of Medʉmba (Grassfields Bantu). Our results show a close correlation between the computationally calculated apexes and our hand-annotated apexes, suggesting that both methods are equally valid for coding video data. The use of markerless motion capture technology for gesture coding will enable more rapid coding of manual gestures, while still allowing for direct comparison with manually-coded data.

**Keywords:** co-speech gestures, gesture coding methods, speech timing, speech-gesture alignment

## 1. INTRODUCTION

Conducting research on speech and co-speech gestures can be challenging due to the time-consuming process of manually annotating gestures in video data when marker-based motion capture methods are not available. This can be especially cumbersome when studying gestures in languages spoken in regions where this technology is not readily accessible. With the advent of markerless motion capture technology, semi-automated coding of gestures from video data is now possible [2], and the potential for rapid and accurate annotation of gestures is greatly expanded. An open question concerns how well traditional manual coding methods align with results of semi-automated methods. This paper investigates manual and semi-automated annotation methods to assess the validity and comparability of the methods.

## 2. METHODS

### 2.1. Data

Gesture analysis in this study is based on a corpus of video/audio data collected with four Medʉmba speakers (2 male and 2 female) through interviews in Banganté, in the West Region of Cameroon. The participants were recorded in pairs engaged in conversation with each other and the interviewer, providing a naturalistic dataset for analysis.

### 2.2. Manual method

Manual gestures were coded using ELAN [3] and the MIT Gesture Studies Coding Manual as a guide. This manual outlines several phases of gestures based on [4], including preparations, strokes, and holds. To this, we added a phase labeled the ’apex’ of the gesture, which occurs within the stroke phase. Each annotated phase is exemplified in Figure 1. The apex has been identified in prior research as an important landmark in speech-gesture timing.

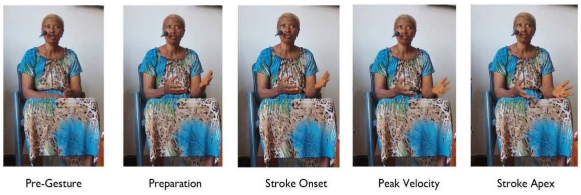

**Figure 1:** Gesture Landmarks

Though the apex is thought of traditionally as the point of maximum extension of the articulators (e.g. the fingers, in the example in Figure 1) [4], this landmark proved difficult to reliably identify in video data due to the limitations of the video frame rate. Namely, the point of maximum extension of a given articulator, e.g., the fingertips, may occur between frames. Thus, we instead used the point at which the hands displayed peak velocity of movement, which corresponded to the largest visualized change in position of the articulator between video frames, often observed by coders as an increase in blurriness between two frames. Furthermore, ELAN only allows for the annotation of intervals, rather than single points in time, and as a result, each apex was annotated as an interval, and the first time-point (T1) of the manually-coded apex was taken as the true timing of the apex for the purpose of calculations in this study.

For each participant in our study, two trained researchers annotated the data independently and then compared annotations to resolve discrepancies. Next, a consensus round was then conducted with an expanded set of coders to resolve any remaining discrepancies. Researchers coded the data and resolved discrepancies in coding with the audio muted so as to avoid any auditory bias in their coding decisions. This approach maximized consistency in coding, though the method still introduces points of potential error as the coding of specific gesture landmarks is subject to each coder's perception and interpretation.

### 2.3. Semi-automated method

Semi-automated coding was performed using OpenPose, a software developed by the Perceptual Computing Lab at Carnegie Mellon University [1]. Pre-processing and analysis were conducted in ELAN and R [5], incorporating elements from the workflow developed by Pouw and Trujillo [2]. This multi-faceted approach allowed us to effectively analyze and interpret the data obtained from the computational coding process. First, each video was run through the Openpose software which identifies 25 articulators, or keypoints, and tracks the X and Y positions of the keypoint relative to the resolution of the video. For this study, we isolated the movement of the right index finger and the resolution of the video was 1280x720 pixels. Openpose likewise assigns a confidence value that indicates how certain the software was that the articulator being tracked corresponds with its associated coordinates. The confidence value was important for identifying the reliability of OpenPose tracking and low confidence values, less than .1 on a scale of 0-1, led us to exclude data for a fourth participant, where poor lighting in the video led to errors in tracking. Figure 2 illustrates the keypoint identification in OpenPose for one participant.

Next, we created a time series of all keypoints produced by Openpose and aligned the time series with the video data based on the video frame rate (FPS). This time series was then used to align the coordinates with the sound file and calculate the speed of the articulator. The speed of the articulator was then smoothed using a Butterworth low-pass filter (frequency = 30 Hz). Finally, the time series was aligned with gesture annotations in ELAN and the manually coded strokes were used to identify the apex, where the apex was the point of max speed within a stroke and maximum speed is determined by the OpenPose coordinate data.

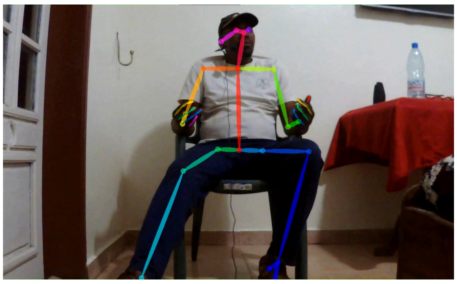

**Figure 2:** OpenPose Motion-Tracking

### 2.4. Method of comparison

In order to assess the similarities between our hand-annotated gestures and OpenPose annotations, we analyzed the consistency in apex coding between the two methods. We used "peak speed timing" as the apex landmark for the semi-automated method and the start time (T1) of the hand-annotated apex as the comparative landmark. Using these apex landmarks, we calculated two measures of apex timing. First, we calculated the relative time of each apex within the stroke, i.e., the T1 of the stroke minus either the OpenPose apex or the T1 of the manually coded apex. Second, we calculated the difference between the two apex measures, i.e., the OpenPose apex minus the T1 of the manual apex. These two measures were used to analyze the similarities of apex timing across the two coding methods, as is discussed in Section 3.1.

We also compared the alignment of apex annotation in the two methods and vowels. Apex alignment with phones was calculated as the T1 of the manually coded apex minus the T1 of the nearest vowel, for manual gesture coding, or the point of maximum speed minus the T1 of the nearest vowel, for semi-automated gesture coding. We then compared the relative apex and vowel alignment of the two methods, as discussed in section 3.2.

## 3. RESULTS

### 3.1. Agreement in apex alignment

Overall, our results show a close alignment between hand-annotated and semi-automatically annotated gesture apexes. Figure 3 shows a Bland-Altman plot characterizing agreement between manual and OpenPose-based measures of apex timing. The y-axis in the plot indexes the difference in timing measurements for apexes within each manually-coded stroke, i.e., the time between the manually-coded apex minus the timing of the corresponding OpenPose estimated apex. The x-axis plots the average of the time points between the two apex measures. The two red dotted lines represent +/- 2 standard deviations from the mean. Results indicate the average difference ('bias') in measurements is around 40 ms—just over 1 video frame—between the two measures, with manually coded apexes marked slightly later than OpenPose apexes.

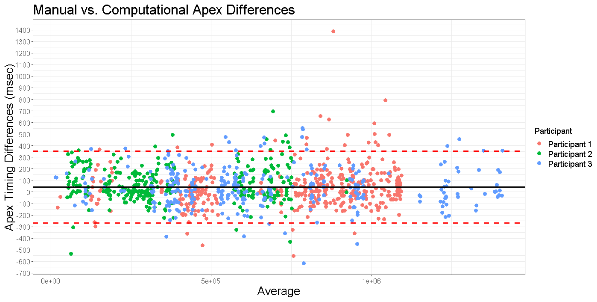

**Figure 3:** Bland-Altman of timing differences between apex measures

For the comparison between apexes across the two methods, 87.5% of the apexes fell within a two-frame (60ms) distance between each other, and 47% of the apexes fell within one frame (30ms) of each other. Larger differences in timing are due to either a tracking issue within Openpose, differences in manual coder interpretation, or a combination of these two factors. For example, coders sometimes annotated longer apexes than was standard in our process when the apex is difficult to discern, resulting in greater timing differences between annotation methods.

### 3.2. Agreement in apex and phone alignment

Since phonetic studies of gesture timing focus on the relative timing between apexes and segments in the speech signal, we also examined the alignment of manual and computational apexes with vowel, the results of which are illustrated in Figure 4.

We analyzed the time between apexes and vowels for both methods and found that the manual method showed an average time-to-vowel of 287 ms while the computational method showed an average time-to-vowel of 300 ms, an average difference of 13 ms. This measure matches our observation stated in Section 3.1, where the manual apexes tended to occur earlier than the semi-automated apexes. Figure 4 illustrates this trend across all three participants, confirming the regularity of the pattern between manually and semi-automatically coded apexes. Thus, while differences between the two methods are minimal and relate predominantly to framerate, the two methods differ in a consistent and therefore predictable way. Overall, our results demonstrate high comparability between apexes coded using the two methods.

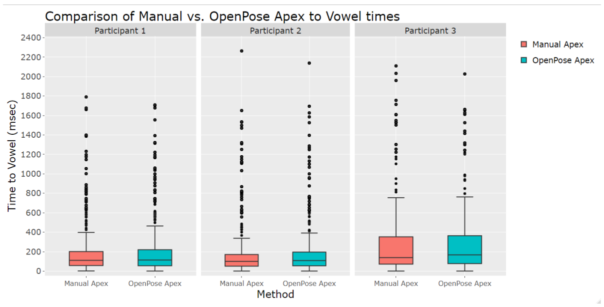

**Figure 4:** Comparison of apex annotation method times to vowel

## 4. FURTHER AUTOMATION OF GESTURAL PHASE ANNOTATION

In our semi-automated analysis, OpenPose estimated apex timing is calculated using manually-coded stroke intervals. Nevertheless, stroke annotation would ideally rely on semi-automated coding to further reduce manual labor. Thus, a future direction in our work will be to estimate stroke location based on gesture kinematics alone, similar to analyses of electromagnetic articulography (EMA) data [6]. A challenge of relying solely on kinematic landmarks is the complexity of movement in co-speech gestures and the corresponding degree of variability across gestures and their kinematic profiles. For example, Figures 5 and 6 provide speed profiles for two different beat gestures; Figure 5 shows a gradual decline in speed, a sharp spike in speed at the apex, and into another gradual decline. However, Figure 6 shows a beat gesture where there is a gradual rise and fall over the course of the gesture.

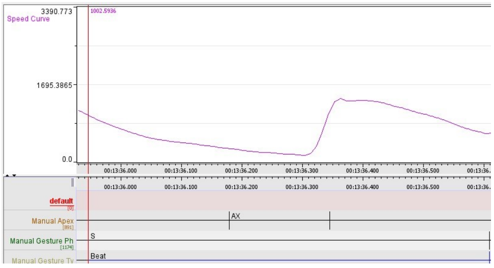

**Figure 5:** Speed profile of beat gesture A

In comparison, the speed envelope for Figure 7 shows an iconic cyclic gesture which is very similar to the speed profile of the second beat gesture shown in Figure 6. The gestural speed profile shown in

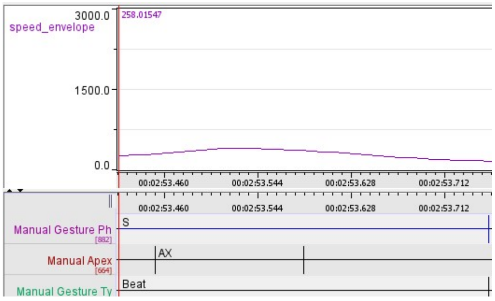

**Figure 6:** Speed profile of beat gesture B

these three figures is only a sample of the degree of variation across gestures and demonstrates just how difficult it is to characterize the stroke of a co-speech gesture from the speed profile alone. The corresponding x and y coordinates are similarly diverse and complex as the gestures themselves are complex and varied. While further investigation of gesture types and forms and their corresponding speed profiles may yield a typical speed profile for a given gesture type and form, additional research on this topic is necessary in order to further automate markerless tracking methodology.

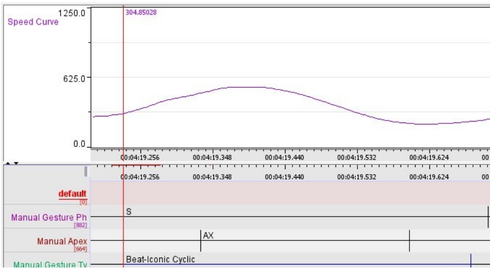

**Figure 7:** Speed profile of an iconic cyclic gesture

Ultimately, the analysis of stroke speed profiles uncovers the necessity of manual annotations, at least for phase identification in gestures. While automated processes can effectively identify gestural landmarks where they are quantitatively defined, manual annotators are able to identify more qualitative aspects of the video. For instance, in training researchers to annotate gestural phases, the importance of intentionally in a movement is often discussed as an important aspect of identifying gestures from fidgets. However, what quantifies intentionally is an open question. Thus, while this study illustrates the similarities between manually coded and computationally coded apexes, there is still a great deal of progress to be made in order to fully automate gestural annotation.

## 5. DISCUSSION

In the absence of marker-based motion tracking technology, two main methods are used to annotate co-speech gestures in video data: manual annotation and markerless tracking, such as OpenPose. While manual annotation has traditionally been used in co-speech gesture research, markerless tracking technology offers a potential means to automate aspects of the annotation process and improve efficiency. This study compares manually coded gesture apexes with those coded semi-automatically using OpenPose, using a corpus of conversations between four Medʉmba speakers.

Our results demonstrate a close alignment between apexes coded using the two methodologies, validating the landmark identification of both methods and the comparability of data across annotation methods. Our analysis demonstrates that the vast majority of apexes (87.5%) coded using the two methods fall within two frames of one another. Furthermore, a comparison of the alignment between phones and apexes using two methods likewise shows a high degree of similarity, with apexes coded using the manual coding method occurring slightly earlier than those using the semi-automated coding process. Thus, while differences between the two methods are minimal, those that do arise are consistent and predictable.

While markerless tracking software offers an efficient approach to annotating gesture apexes, it still relies on manual coding for some tasks. For example, both semi-automated and manual coding methods rely on manually defined stroke boundaries to identify apexes. Further research on the kinematic characteristics of stroke phases is needed to enable automation of this process. Yet, it may be possible to automate differentiation between gestural phases and types through machine learning, using a large corpus for training. While the implementation of this approach would require significant resources, current technology makes it feasible.

Overall, semi-automated methods of gesture annotation offer a reliable and efficient means of streamlining gesture research that is comparable to traditional manual annotation, with similar timing of apexes and alignment between vowels and apexes between the two methods. Therefore, semi-automated annotation is a valid method in co-speech gesture timing research. This study contributes to the field by demonstrating a way to expand co-speech gesture corpora without using motion capture technology, enabling further research on co-speech gesture and the automation of gestural annotation using markerless tracking technology.

## 6. REFERENCES

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," 2018. [Online]. Available: https://arxiv.org/abs/1812.08008

[2] W. Pouw and J. Trujillo, "Materials tutorial gespin2019 - using video-based motion tracking to quantify speech-gesture synchrony,." [Online]. Available: osf.io/rxb8j

[3] Max Planck Institute for Psycholinguistics, "Elan." [Online]. Available: https://archive.mpi.nl/tla/elan/download

[4] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in The Relationship of Verbal and Nonverbal Communication. DE GRUYTER MOUTON, Dec. 1980, pp. 207–228.

[5] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[6] M. Tiede, "Mview: Multi-channel visualization application for displaying dynamic sensor movements," development, 2010.