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

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

[1]University of Delaware, [2]Harvard University
wdych@udel.edu, kareegarvin@fas.harvard.edu, kfranich@fas.harvard.edu

ABSTRACT

While motion capture is rapidly becoming the gold-standard for research on the intricacies 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 Medumba (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 marker-based technology is not readily accessible, increasing the necessity of manually annotating data. 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 Medumba 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 about cultural practices specific to the region. The speech examined in the study is therefore relatively 'naturalistic,' meaning that participants were free to speak and gesture as they pleased, and the topic of conversation was relatively unconstrained.

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. We also coded the 'apex' of the gesture, which occurs within the stroke phase. The apex has been identified in prior research as an important landmark in speech-gesture timing [5, 6, 7]. Each annotated phase is exemplified in Figure 1.

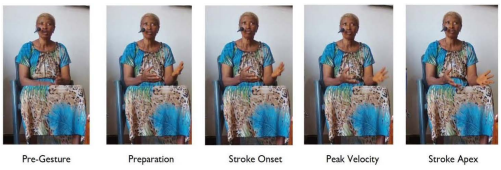

Figure 1: Gesture Landmarks

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. 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. Researchers coded the data and resolved discrepancies in coding with the audio muted to avoid any auditory bias. Next, a consensus round was conducted with an expanded set of coders to resolve any remaining discrepancies. This approach maximized consistency in coding, though the manual approach is still prone to subjective bias.

2.3. Semi-automated method

Semi-automated coding was performed using OpenPose, a software tool developed by the Perceptual Computing Lab at Carnegie Mellon University [1]. Pre-processing and analysis were conducted in ELAN and R [8], 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 to identify 25 articulators, or keypoints, and track 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 (all subjects were right-handed) with a video resolution of 1280x720 pixels. OpenPose also assigns a confidence value (between 0 and 1) that indicates the probability of the model's tracking accuracy. Due to low confidence values (<.1), we excluded data for the fourth participant, where poor lighting likely 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

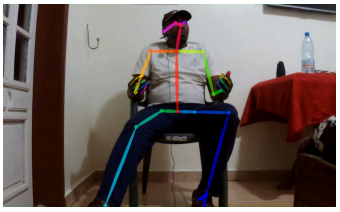

Figure 2: OpenPose Motion-Tracking

was then smoothed using a Butterworth low-pass filter (frequency = 30 Hz). Finally, the time series was aligned with gesture annotations in ELAN. The manually-coded strokes were used to identify the apex, where the apex was the point of maximum speed within a stroke based on coordinate data.

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 gesture apexes with vowels in the accompanying speech across the two methods in order to determine the comparability of the gesture annotation methods in relation to speech timing. 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.

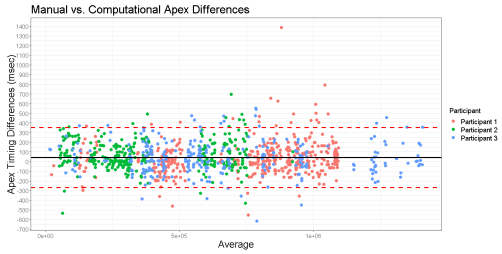

Figure 3: Bland-Altman of timing differences between apex measures

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 was 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 [6, 9, 10, 11, 12], we also examined the alignment of manual and computational apexes with vowels; the results 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

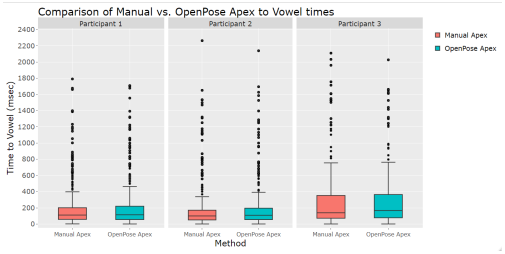

Figure 4: Comparison of apex annotation method times to vowel

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.

4. FURTHER AUTOMATION OF GESTURAL PHASE ANNOTATION

Currently, this method allows us to automatically code gesture apexes based on *manually*-coded gesture strokes. Thus, further automating gestural annotation to further reduce manual labor would require a method for automatically identifying stroke intervals. Articulatory kinematics have been used to estimate gesture onsets in speech articluation data (EMA) [13], and a similar approach could be applied in the case of co-speech gestures. However, the level of variability in articulatory parameters such as hand shape, orientation, and gesture location pose challenges to this approach. Related to this, the velocity profiles for different gestures can look quite different. For example, Figures 5 and 6 provide speed profiles for two different gestures, the first a bimanual combined depictive-beat gesture (where the hands are moving downward and outlining the shape of an item being described), and the second, a single-handed beat gesture with the hand extending outward and downward. Figure 5 shows a gradual decline in speed, a sharp spike in speed before the apex, into gradual decline. In contrast, Figure 6 shows a gradual rise and fall over the course of the gesture. In comparison, the speed profile for Figure 7 shows a cyclic gesture —where the speaker moves his two hands in a fluid, circular motion—which is very similar to the speed profile of the second beat gesture shown in Figure 6. The gestural speed profile shown in 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.

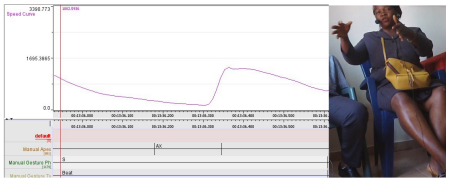

Figure 5: Speed profile of beat-depictive gesture

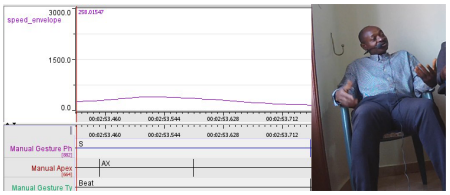

Figure 6: Speed profile of beat gesture

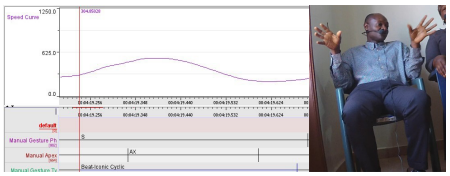

Figure 7: Speed profile of depictive cyclic gesture

Another challenge of relying exclusively on kinematic data to code stroke intervals is that not all hand movements are gestures; i.e., speakers move their hands to fidget, adjust clothing, etc. Human annotators can easily distinguish gestures from fidgets due to the difference in form and intention behind gestures; however, computationally defining what constitutes a fidget is more challenging.

Ultimately, the analysis of stroke speed profiles uncovers the necessity of manual annotations, for identifying stroke boundaries. Although automated processes can effectively identify gestural landmarks where they are quantitatively defined, manual annotators are better able to identify more qualitative aspects of the video. 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, and may even distinguish gestures from fidgets, additional research on this topic is necessary in order to further automate markerless tracking methodology. Thus, this study illustrates the similarities between manually-coded and computationally coded apexes and presents areas for further progress 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, e.g., through the use of OpenPose software. While manual annotation has traditionally been used, markerless tracking technology offers a potential means to automate aspects of the annotation process, improving efficiency. This study compares manually-coded gesture apexes with those coded semi-automatically using OpenPose, using a corpus of conversations between four Medumba speakers.

Our results demonstrate close alignment between apexes coded using the two methods, 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%) fall within two frames of one another. A comparison of the alignment between vowels and apexes likewise shows a high degree of similarity, with manually-coded apexes occurring slightly earlier (13ms) 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, our approach still relies on manual coding for some tasks. For example, both semi-automated and manual coding methods rely on manually defined stroke boundaries, based on which apex timing is determined. Further research on the kinematic characteristics of stroke phases is needed to enable automation of this process. Machine learning techniques provide a promising avenue for classifying different gesture types in terms of their kinematic profiles.

Overall, semi-automated methods of gesture annotation offer a reliable and efficient means of streamlining co-speech gesture research. Automatic apex detection yields similar results to manual apex coding with human coders (with some small, predictable differences), indicating that data annotated with the two methods is comparable. 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 marker-based motion capture technology.

6. ACKNOWLEDGMENTS

The authors would like to thank the Medumba speakers who participated in the study and to Dr. Ange Bergson Lendja for his assistance in coordinating the study. The authors would also like to express their sincere gratitude to the University of Delaware PhonLab team for their invaluable contributions and unwavering commitment to this study. The collective efforts of Crystal Akalu, Christine Belden, Kylie Boggs, Nicole Curnan, Luc DiNardi, Walter Dych, Bryant Graybeal, Lindsay Hawtof, Jacquelyn Janocha, Brett Lindsey, Zoe Lipkin, Madison Miller, Meghan O'Brien, Victoria Ochlan, Anna Schumeyer, Nicole Taylor, Ta-tran Tran, and Juliette Winnard in processing and analyzing data have been instrumental in achieving the results presented in this paper. Their expertise, dedication, and camaraderie have greatly enriched our research and fostered an environment conducive to innovation and discovery.

This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are our own.

7. 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] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements," Journal of Speech, Language, and Hearing Research, vol. 56, no. 3, pp. 850–864, 2013. [Online]. Available: https://doi.org/10.1044/1092-4388(2012/12-0049)

[6] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," Laboratory Phonology, vol. 3, no. 1, 2012. [Online]. Available: https://doi.org/10.1515/lp-2012-0006

[7] A. B. Hostetter and M. W. Alibali, "Visible embodiment: Gestures as simulated action," Psychonomic Bulletin & Review, vol. 15, no. 3, pp. 495–514, 2008.

[8] 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/

[9] K. Franich and H. Keupdjio, "The Influence of Tone on the Alignment of Speech and Co-Speech Gesture," in Proc. Speech Prosody 2022, 2022, pp. 307–311.

[10] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," Speech Communication, vol. 57, pp. 209–232, 2014.

[11] I. Biazzo, C. Canestrari, and E. Demuru, "Gesture-prosody synchrony: Evidence from italian," Gesture, vol. 14, no. 2, pp. 171–193, 2014.

[12] A. Watanabe and Y. Hirose, "Correlates between prosodic features and manual gesture in japanese spontaneous speech," Language and Speech, vol. 58, no. 2, pp. 225–243, 2015.

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