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

XXX

XXX
XXX

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 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 co-speech gestures and their relationship with speech 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-automatic 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 from this type of automated method. This paper aims to understand what differences arise between manual and semi-automatic 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, 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 Kendon [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 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 between annotations. 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-Automatic Method

Semi-automatic coding was performed using OpenPose, a software developed by the Perceptual Computing Lab at Carnegie Mellon University [1]. Pre-processing and analysis was conducted in ELAN and R [5], incorporating elements from the workflow developed by Wim Pouw and James Trujillo [2] from the Donders Institute. 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 which 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 articular. The speed of the articulator was then smoothed using a Butterworth low-pass

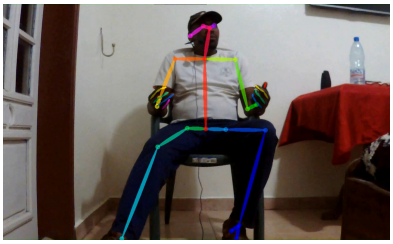

Figure 2: OpenPose Motion-Tracking

filter (frequency = 30 Hz). Finally, the time series was likewise aligned with gesture annotations in ELAN and the manually coded strokes were used to identify the apex, where the apex was the point of maximum speed within a stroke and maximum speed is determined by the OpenPose 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. To this end, we used "peak speed timing" as the apex landmark for the semi-automatic 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.g., 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 relative similarities of apex timing across the two coding methods, as is discussed in Section 3.1.

3. RESULTS

3.1. Analysis of Agreement on Apex Alignment

Overall, our results show a close correlation 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 in the data. The x-axis on the plot represents the average of the time points between the two apex measures. The two red dotted lines represent +/- 2 standard deviations from the mean, respectively. 94.5% of our data fall within 2 standard deviations of the mean.

For the comparison between apexes across the

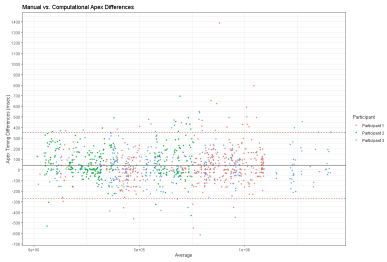

Figure 3: Bland-Altman of timing differences between the 3 participants

two methods, 87.5 percent of the apexes fell within a two-frame (60ms) distance between each other, and 47 percent of the apexes fell within one frame (30ms) of each other. The results of this analysis show a high degree of similarity between apexes across the two methods, suggesting that landmark identification for the apex is similar across the two methods. Larger differences in timing are due to either a tracking issue within Openpose, errors in manual coding, or a combination of these two factors.

We also examined the alignment of manual and computational apexes with phonetic information by analyzing the time between the apex of a gesture and its associated phone in speech. Figures 4 and 5 illustrate the alignment between phones and manually and computationally coded apexes, respectively.

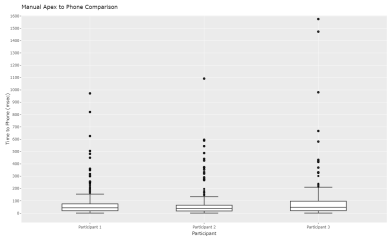

Figure 4: Time from manual apex to phone

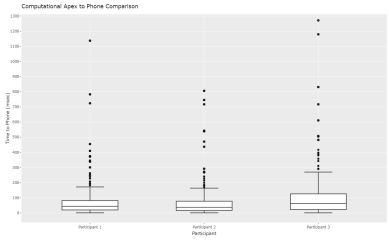

Figure 5: Time from computational apex to phone

As illustrated by the figures, the alignment between phones and gesture apexes is similar across the apex coding methods; however, the computational coding method was slightly more accurate, with only 7% of alignments producing outliers, compared to 10% for the manual method. This alignment between apexes and phones is important because the time-to-phone from the apex of a gesture is a commonly relied upon method in gesture-speech analysis. Thus, these results demonstrate the validity of both methods in understanding the timing relationship between speech and co-speech gesture.

4. FURTHER AUTOMATION OF GESTURAL PHASE ANNOTATION

At present, both manually coded apexes and computationally coded apexes rely on manually coded strokes as even the computational method identifies the apex as the point of maximum speed within the manually coded stroke. Thus, a future direction in streamlining the semi-automatic annotation process is to computationally identify the stroke of a gesture. One avenue for computational annotation of the stroke is to identify the gestural landmarks of a stroke from the speed profile, similar to methods used for analyzing point tracking data like electromagnetic articulograph (EMA). However, a challenge in this method of identification is the complexity of movement in co-speech gestures and the corresponding degree of variability across gestures and their respective speed profiles. For example, Figure 6 and 7 provide the speed profile for a typical beat and iconic cyclic gesture, respectively. As is illustrated by these figures, the speed profile differs substantially between the two gestures, where the beat gesture is characterized by a gradual decrease in speed, followed by a sharp increase, followed by another gradual decrease in speed. Conversely, the iconic cyclic gesture is characterized by a gradual increase and decrease in speed.

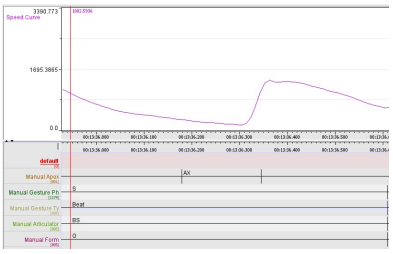

Figure 6: Example of beat gesture with speed curve

Figure 7 is a representation of the speed cure of a cyclic gesture, where the speed curve is more gradual and similar to the shape of a sine wave.

The gestural speed profile shown in these two 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

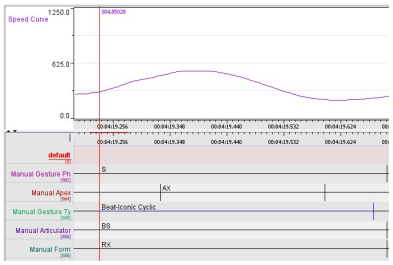

Figure 7: Example of cyclic gesture with speed curve

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 for further automation of the markerless-tracking methodology.

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 quantitative 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

Where marked-based motion tracking technology is not available, there are two main methods of annotating gestural phases from video data: manual annotation and markerless tracking software, e.g. OpenPose. While manual annotation has long been relied on for co-speech gesture research, markerless tracking technology provides a promising means to automate parts of the gesture annotation process in order to streamline and expedite the coding process. This study compares manually coded gesture apexes with semi-automatically coded apexes using the markerless tracking software, OpenPose. We analyze a corpus of conversations between four speakers of Medumba.

The results of our analysis demonstrate a high degree of similarity across the two coding measures validating landmark identification of both methods and the comparability of data annotated across the two methodologies. Our analysis demonstrated that the vast majority of apexes (87.5%) coded using the two methods fell within two frames of one another. Likewise, in comparing the alignment between phones and the two apex identification methods similarly showed a high degree of similarity, with slightly higher accuracy of the computational method.

While markerless tracking software provides a streamlined approach to apex annotation that serves to ease the labor-intensive process of manual gesture coding, markerless tracking technology still relies on elements of the manual coding process. Namely, while OpenPose tracking can be used to obtain the speed profile of the gesture, apex identification relies on the boundaries of the manually coded stroke phase in both semi-automatic and manually coded methods. Further research on the quantitative profile of the stroke phase is needed in order to automate stroke identification as co-speech gestures are complex and varied in both their speed and x, y coordinate profile, making it difficult to automate this part of the process. Nevertheless, training machine learning models to differentiate between both gestural phases and types may provide a fruitful path to further automation of this technology and our understanding of co-speech gesture. While the implementation of this approach would require a substantial corpus for training, the technological capabilities currently available make this a feasible possibility.

Overall, semi-automated methods of gesture annotation provide a reliable means for streamlining gesture research that is comparable to traditional methods of manual annotation. Furthermore, alignment between computationally coded apexes and phones is not only comparable to the alignment between traditionally coded apexes and phones, but in fact, provides a slight improvement in the accuracy of alignment to phones, making the semi-automatic annotation method a valid method in co-speech gesture timing research. Thus, this study provides an important contribution to co-speech gesture research as it provides a means to expand co-speech gesture corpora without the use of motion capture technology, allowing for increased research on co-speech gesture and the further 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/