GESTURAL ALIGNMENT AND ACCOMMODATION IN SPEAKER-LISTENER HEAD GESTURES

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

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

ABSTRACT

The timing of both manual co-speech gestures and head gestures is sensitive to prosodic structure of speech. However, head gesters are used not only by speakers, but also by listeners as a back-channeling device. Little research exists on the timing of gestures in back-channeling. To address this gap, we compare timing of listener and speaker head gestures in an interview context. Results reveal the dual role that head gestures play in speech and conversational interaction: while they are coordinated in key ways to one's own speech, they are also coordinated to the gestures (and hence, the speech) of a conversation partner when one is actively listening to them. We also show that head gesture timing is sensitive to social dynamics between interlocutors. This study provides a novel contribution to literature on head gesture timing and has implications for studies of discourse and accommodation.

Keywords: Co-speech gesture, gesture alignment, speech timing, discourse, accommodation

1. INTRODUCTION

Literature on the timing alignment of speech and co-speech gesture demonstrates that gestures are sensitive to the prosodic signal, including stress and intonational phrase boundary [1]. In particular, a number of studies have shown that gestures are timed to stressed vowels or pitch accents (see [2] for an overview) and gestures are retracted, or shifted earlier, at intonational phrase boundaries [3, 4], with similar results across manual and head gestures.

While manual gestures are crucially timed with one's own speech, head gestures are employed by both speakers and listeners during conversational interaction: while speakers are known to align their head gestures with prominent syllables in their own speech, listeners use head gestures as a back-channeling device to indicate engagement with the speaker [5]. Listener head gestures are known to be denser at points of overlapping discourse [6]; however, how head gestures are timed relative to the discourse, e.g., whether gestures are still timed to speech or other aspects of the interaction, is unstudied. Furthermore, back-channeling cues are subject to accommodation, where the social dynamic between interlocutors influences speech patterns such as the timing [7], acoustics, and prosody of speech [8, 9]. Interlocutors that accommodate to their speech partners are viewed as more likable and the resulting conversations are perceived as more natural and successful [10, 11, 12, 13, 14]. Whether social roles affect head gesture coordination between speakers remains unknown.

We analyze how interlocutor roles—speaker vs. listener—affect head gesture timing. Furthermore, we analyze how timing differs across speakers depending on the social role of the interlocutor by analyzing differences in the interviewer and subjects using a corpus of Salon Talks interviews. We hypothesized that, consistent with previous literature, the timing of one's own gestures would align with stressed syllables in one's own speech and that this alignment would be sensitive to proximity to an intonational boundary. Given the importance of head gestures in listener back-channeling, we also predicted that participant head gestures would align closely with an interlocutor's speech when the participant was actively listening to their interlocutor and that timing of the interviewer's gestures would be more sensitive to the interview subject's gestures than the other way around. Our findings confirm that listener gestures are highly sensitive to the timing of the interlocutor's gestures. Furthermore, we find that the interviewer shows greater accommodation to the timing of the subject's gestures than the subjects to the interviewer.

2. METHODS

2.1. Data

To analyze gestural turn-taking, we used interviews from Salon Talks retrieved from YouTube. Salon

Talks interviews were optimal for analyzing gestural turn-taking because the videos had few camera angles and had long durations shot from a camera angle in which speakers' heads and upper bodies were simultaneously visible (see Figure 1). We examined clips from four interviews conducted by Mary Elizabeth Williams with the subjects Zoey Deutch (posted Mar 21, 2018), Jesse Eisenberg (posted Mar 11, 2019), Randall Park (posted Jun 6, 2019), and Mira Sorvino (posted Nov 7, 2018). The clips were approximately two minutes in length and consisted of a single camera angle where both speakers were visible for the duration of the clip.

salontalks
salon
"ALWAYS BE MY MAYBE" FOLLOWS CHILDHOOD SWEETHEARTS WHO RECONNECT AFTER 15 YEARS
@salon | @salon | salon.com

Figure 1: Still image from Salon Talks interview with Mary Elizabeth Williams and Randall Park.

2.2. Gesture and speech annotation

Gestures were coded by a team of researchers trained in gesture coding using ELAN [15] following the MIT Gesture Studies Coding Manual [16], which outlines several phases of the gesture including preparations, strokes, holds, and recoveries based on [17]. This method was updated for head gestures following [2] and [3]. The videos were coded for gestural phase, apex, and interlocutor role. The apex of the gesture was coded as the point of maximum extension. Because ELAN does not permit annotation of a single point in time, but instead requires interval annotations, the point of maximum extension was annotated as the end point of the apex (T2) and extended two frames prior to the apex. Calculations regarding apex used the T2 of the apex interval. For each video, there were approximately 250 apexes coded per subject. The interlocutor role was coded as *speaker*, *listener*, or *speaker/listener*, where both the interviewer and subject were speaking simultaneously. Gestures that were coded as *speaker/listener* were not analyzed in this study as gestures could not be reliably attributed to the speech of a single speaker, and thus, the interlocutor role was ambiguous. A transcript of the interview clip was made by a team of researchers and was aligned using the FAVE Forced Aligner [18]. The forced alignment data was then hand-corrected by a team of researchers.

2.3. Speech to gesture and gesture to gesture timing

To analyze the timing between gestural apexes and phones, the lag time between apex and phone was calculated as the T2 of the apex minus the start time (T1) of the phone. Phones were coded as either stressed vowel, unstressed vowel, or consonant. To analyze how intonational boundary affected lag time between vowels and gestures, the distance to nearest vowel was calculated following the same method. Likewise, intonational boundary was coded as either initial, for the first word of an intonational phrase, final for the final word of an intonational phrase, or else medial. Together, these calculations allowed us to analyze the alignment between gestural apexes and prosodic elements including, phone, stress, and intonational phrase boundary (see Section 3.1).

To analyze the timing between gestures across interlocutors, the apex lag time was calculated as the time from the T2 of one apex to the T2 of another apex in a sequence of alternating gestures between interlocutors, as illustrated in Figure 2. In other words, in a sequence of Gesturer A^1 - Gesturer B - Gesturer A^2, Gesturer B's lag time was calculated as Gesturer B minus Gesturer A^1.

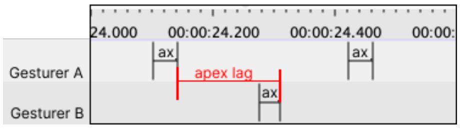

Figure 2: Example of alternating gestures between speakers used to calculate apex lag. Endpoint of ax intervals coincides with the point of maximal head extension.

This measure of apex lag was used to analyze the timing relationship between interlocutors, i.e., *speaker* and *listener*, across the dyads, e.g., Mary Elizabeth Williams and Randall Park, and based on participant role, i.e., *interviewer* and *subject*. The results of this analysis are discussed in Section 3.2.

3. RESULTS

3.1. Timing of speech and gestures

Overall, the results of the analysis on the timing between gestures and speech are consistent with previous work on the alignment of speech and co-speech gesture [2, 3, 4]. We found that gestural apexes were more likely to align with stressed vowels than with other segments. The correlation between stressed vowels and gesture apex alignment was analyzed using a Pearson's Chi-squared test ($p < 0.001$). There is a significant correlation between the alignment of gesture apexes and stressed vowels, as illustrated in Figure 3. Positive or 'attraction' relationships are shown in blue and negative or 'repelling' relationships are shown in red, where the darkness of the shade of blue or red indicates the strength of the relationship. The size of the bubble indicates the contribution of the variable to the model, where larger bubbles indicate variables that have a larger effect in the model. As indicated by the large, dark blue bubble for stressed vowels aligned with apexes, there is a strong attraction relationship between stressed vowels and gestural apexes, consistent with [2].

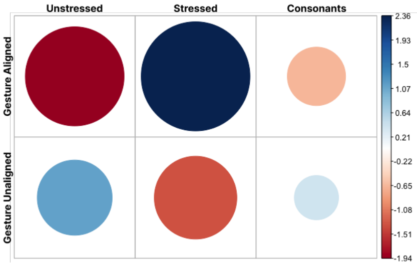

Figure 3: Correlation plot between segment type and gesture aligned vs. unaligned segments.

We also examined intonational phrase boundaries and apex alignment by analyzing the time between vowels and aligned apexes and found that gestures retract, or occur earlier with respect to the vowel, at intonational phrase boundaries (Figure 4). We hypothesize that this result was obscured for Zoey Deutch because she had fewer gestures occurring at phrase boundaries.

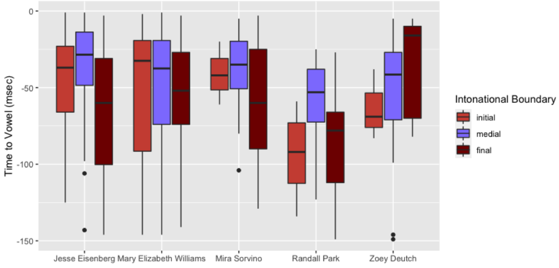

Figure 4: Time between apexes and vowels at intonational boundaries, plotted by subject. Value of 0 corresponds to perfect alignment between gesture apex and vowel onset.

A linear mixed effects model (lmer function of the lme4 package [19]) of the time between vowels and apexes shows a significant effect, consistent with the findings in [3] (Initial vs Final: β= 15.082, t= 2.157, p< 0.05; Medial vs Final: β= 25.626, t=5.446, $p < 0.001$). All models were originally fit with the maximal random effect structure with random slopes for all predictors, but where the fit was singular, we removed random slope terms that eliminated singularity [20].

Together, these results show that head gestures are sensitive to prosodic elements of speech including both stress and intonational boundary. Gestural apexes tend to coincide with stressed vowels, but are retracted at intonational phrase boundaries.

3.2. Timing between conversation partner gestures

Head gestures, in addition to being used to augment one's own speech, are likewise used by listeners as a back-channeling cue, indicating interaction and engagement. Furthermore, speakers provide timing cues to listeners both in the timing of their speech and in the timing of their own gestures. Thus, in addition to analyzing the timing between speech and head gestures, we also analyzed the timing between interlocutor gestures, focusing on the difference in timing between listeners and speakers at points in the video where there was no overlap between speakers, and as such, speech role was unambiguous. Overall, listeners time their apexes to their interlocutor's gestures, while speakers time their gestures to their own speech.

To analyze the influence of interlocutor gestures on apex timing, we calculated the lag between each participant's head gesture apex relative to the immediately preceding apex of their conversation partner. This provides a measure of alignment between gestures across participants. We found that overall lag time was shorter when participants were listening to their interlocutor than when they were speaking. A linear regression demonstrates that this effect is significant (Interlocutor Role (speaker): β= 0.25, t= 3.01, p< 0.01; Participant Role (subject): β= 0.28, t=3.44, $p < 0.001$).

Figure 5 plots the lag between speakers for each of the interview subjects alongside the interviewer and shows that apex lag times were shorter when the subject was listening than speaking for all subjects except Mira Sorvino. The results likely differ for Sorvino because she gestured less while listening than the other subjects, thus providing a relatively limited sample and obscuring the trend. The plot additionally shows that the interviewer's lag time was consistently shorter than her subject's relative to her own and that the difference in lag between her apexes as speaker vs. listener was smaller.

On the other hand, gesture-to-speech timing

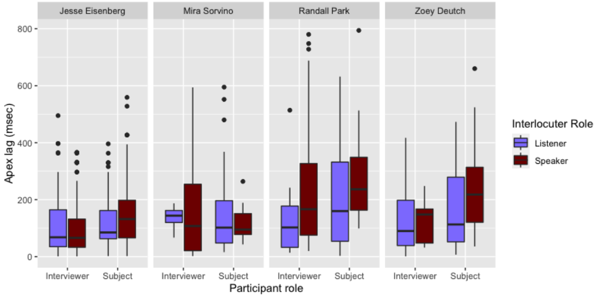

Figure 5: Time between the gestural apex of Speaker A to the gestural apex of Speaker B plotted for four dyads with Mary Elizabeth Williams as the interviewer across all four dyads.

showed a very different pattern: all five participants showed shorter apex to vowel lag times between apexes and phones when speaking compared to listening, shown in Figure 6; this is consistent with the prediction that participants would time their gestures with their own speech when in the speaker role. A linear mixed effects model (lmer function of the lme4 package [19]) predicting time between apexes and vowels from interlocutor role with participant as a random intercept shows that this effect was significant (Interlocutor role (speaker): β= 12.433, t= 2.381, p= 0.05).

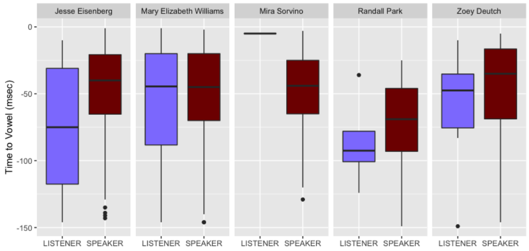

Figure 6: Time between apexes and vowels by interlocutor role, where lag time is shorter when participants are speaking compared to listening.

Overall, these results show that speakers time their gestures to speech while listeners time their gestures to their interlocutor's gestures. Furthermore, the social role of the interlocutor influences the degree of this effect, where the interviewer showed greater accommodation to the interview subject's gestures regardless of the interviewer's interlocutor role.

4. DISCUSSION

The results of this study on the timing between head gestures and speech are overall consistent with the findings of previous literature [2, 3, 4], where gestures are sensitive to prominence in the speech signal. Namely, gestural apexes are timed to stressed vowels and are retracted from stressed vowels at intonational phrase boundaries. However, a novel contribution of this study is the clear difference in gestural timing depending on interlocutor role. In particular, in analyzing the lag time of alternating apexes between interlocutors, there is a difference between the timing of gestural apexes in listeners compared to speakers, where apex-to-apex lag times are shorter for listeners than for speakers. This is contrasted with the lag time between apexes and phones, where lag times are shorter between apexes and phones for speakers compared to listeners. This asymmetry is indicative of the dual role that head gestures play, coordinating with a participant's own speech when they are speaking, and with an interlocutor's gestures when listening.

In addition, the social role of the participants influenced gestural timing. In particular, the interviewer had shorter apex lag times both when speaking and when listening, and furthermore, shorter lag times than the interview subjects. These results are consistent with literature showing interlocutors who accommodate to their conversation partner are viewed as more likable and the interactions are viewed as more successful [10, 11, 12, 13, 14]. Mary Elizabeth William's role as the interviewer positions her to accommodate more to her interview subjects, and thus, her gestural apexes more closely align to the interview subjects' gestures. These results are an essential contribution to literature on gestural timing and accommodation.

5. CONCLUSION

This study confirms the findings that speaker gestures are timed to prominent elements of the speech signal. In addition, we demonstrate that gestural alignment differs depending on whether the gestures are produced by speakers or listeners. Namely, while speakers' gestures are more closely timed to aspects of the speech signal, listeners' gestures are more closely timed to their interlocutor's gestures. Furthermore, patterns in timing differ depending on the social role of the interlocutor, where the interviewer shows greater accommodation to the interview subjects. These results provide novel insights into gestural alignment, showing that listeners use the visual cues of an interlocutor's gestures to plan their own gestures and that listeners engage with and accommodate to their interlocutor's gestures.

6. ACKNOWLEDGMENTS

The authors would like to thank the graduate and undergraduate researchers who helped annotate and process the data discussed in this paper, including 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. 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] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012. [Online]. Available: https://doi.org/10.1515/lp-2012-0006

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

[3] N. Esteve-Gibert, J. Borràs-Comes, and E. Asor, "The timing of head movements: The role of prosodic heads and edges," *The Journal of the Acoustical Society of America*, vol. 141, pp. 4727–4739, 2017.

[4] J. Krivokapič, M. Tiede, and M. E. Tyrone, "A kinematic analysis of prosodic structure in speech and manual gestures." in *ICPhS*, 2015.

[5] J. B. Bavelas and J. Gerwing, "The listener as addressee in face-to-face dialogue," *International Journal of Listening*, vol. 25, no. 3, pp. 178–198, 2011.

[6] S. G. Danner, J. Krivokapič, and D. Byrd, "Co-speech movement in conversational turn-taking," *Frontiers in Communication*, vol. 6, 2021.

[7] Štefan Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[8] Y. Lee, S. G. Danner, B. Parrell, S. Lee, L. Goldstein, and D. Byrd, "Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task," *PLoS ONE*, vol. 13, 2018.

[9] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[10] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[11] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech accommodation theory: The first decade and beyond," *Annals of the International Communication Association*, vol. 10, no. 1, pp. 13–48, 1987.

[12] J. B. Hirschberg, A. Nenkova, and A. Gravano, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL/HLT*, 2008.

[13] R. L. Street Jr, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[14] S. Stoyanchev and A. Stent, "Lexical and syntactic adaptation and their impact in deployed spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 189–192.

[15] "ELAN (version 6.4) [computer software]," 2022. [Online]. Available: https://archive.mpi.nl/tla/elan

[16] "MIT Speech Communication Group Gesture coding manual." [Online]. Available: http://scg.mit.edu/gesture/coding-manual.html

[17] A. Kendon, *Gesticulation and Speech: Two Aspects of the Process of Utterance*. Berlin, New York: De Gruyter Mouton, 1980, pp. 207–228.

[18] "FAVE (Forced Alignment and Vowel Extraction) suite." [Online]. Available: https://www.research.ed.ac.uk/en/publications/fave-forced-alignment-and-vowel-extraction-suite

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, 2014.

[20] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.