The Influence of Prosody on the Requirements for Gesture-Text Alignment

Yingying Wang and Michael Neff

University of California, Davis
{yiwang,mpneff}@ucdavis.edu

Abstract. Designing an agent capable of multimodal communication requires synchronization of the agent's performance across its communication channels: text, prosody, gesture, body movement and facial expressions. The synchronization of gesture and spoken text has significant repercussions for agent design. To explore this issue, we examined people's sensitivity to misalignments between gesture and spoken text, varying both the gesture type and the prosodic emphasis. This study included ratings of individual clips and ratings of paired clips with different alignments. Subjects were unable to notice alignment errors of up to ±0.6s when shown a single clip. However, when shown paired clips, gestures occurring after the lexical affiliate are rated less positively. There is also evidence that stronger prosody cues make people more sensitive to misalignment. This suggests that agent designers may be able to "cheat" when it comes to maintaining tight synchronization between audio and gesture without a decrease in agent naturalness, but this cheating may not be optimal.

1 Introduction

Agents capable of multimodal communication simultaneously express content across multiple channels. These channels include body movement, facial expressions, gesture, spoken text and prosody changes. The need to coordinate these channels puts high demands on the planning and animation subsystems of an agent, yet it is not clear how tight the synchronization must be in order to generate a believable agent. This paper looks at the need for alignment between gesture and spoken text, and how this requirement may vary across the two main factors likely to influence it: the type of gesture performed and variation in the prosody of the accompanying text.

To better understand the need for gesture-text alignment, and how prosody may influence it, we ran two sets of experiments. The first asked people to rate single clips for naturalness that always had strong prosodic emphasis on the lexical affiliate, but varied gesture type and alignment. The second asked people to select a preferable clip from two side-by-side clips and examined variation in the gesture type, prosody and alignment. Results indicate that people have quite low sensitivity to alignment when viewing a single clip. However, when given a second, side-by-side clip, they in general show a lower preference for gestures occurring after the lexical affiliate.

2 Background

According to McNeil in [15], gestures are classified into 4 categories, “beat” - a rhythmic flick of finger, hand or arm to highlight what is being said, “deictic” - a pointing gesture with direction, “iconic” - a representation of a concrete object, or painting with hand and “metaphoric” - explanation of an abstract concept. Besides this, [10, 20, 7] also provide their own taxonomies of gestures.

In virtual character gesture research, solutions for coordinating gestures with other modalities have received a lot of focus. Kopp in [12] presents an incremental production model that can help generate multimodal behaviors from utterance planning. Stone et al.[23] worked on a framework for creating talking characters with both sound and motion data from the real human performance. Cassell et al.[5] takes in text input and generate nonverbal behaviors including both facial expressions and gesture. The system presented in [19] uses plain text as input and their goal is to find a solution for automatically adding gestures.

From previous research, among the multiple communication modalities, prosody is shown to have a close correlation with gesture. According to [24], prosody correlates to emphasis, which makes it coincide with emphasizing gestures like “beats”. Adolphs in [1] and Schroder in [22] mentioned that prosody contains emotive information, which according to [25] and [17] is frequently reflected in non-verbal behavior. Prosody features were previously used to generate facial expression [2, 9], head motion [6], head orientation [21], and gesture [18, 14].

According to McNeil [15], gesture strokes end at or before, but never after the prosodic stress peak of the accompanying syllable. Results from [26] showed that 90% gestures cooccur with lexeme syllable in the speech, and 65% to 75% cases contain a prosody accent. Experiments in [11] indicate that gestures which are performed 0.2 second or 0.6 second earlier w.r.t. the accompanying text get higher ratings for their naturalness.

3 Experiment 1: Perception of gesture misalignment with single clip

Our main hypotheses are that the perceived gesture misalignment varies based on gesture type and also prosodic emphasis. To verify the effect of gesture type and prosodic emphasis, our experiment design include four different types of gestures: “deictic” (D), “metaphoric” (M), “iconic” (I) and “beat” (B) – being placed on utterances with weak (W) prosody or strong (S) prosody.

3.1 Gesture Form

We designed the following four utterances to include the "deictic", "metaphoric", "iconic" and "beat" gesture types. The designed gestures are associated with the highlighted lexeme. A "pointing" gesture was used as the "deictic" gesture, indicating "you" in the text; a "progressive" gesture symbolizes the progress of the conversation; the "iconic" gesture shapes the size of the box; and a dismissal flip of hand motion was used as one-peak "beat" gesture to express the negative content. The gesture motions were generated by editing motion captured data.

T1(D): I know that **you** took it.
T2(M): The conversation **ran** for a long time.
T3(I): I bought a **box** at the store.
T4(B): This is **not** the case.

3.2 Prosody Emphasis

We generated two utterances for each text, with weak prosody and strong prosody. Based on previous research [6, 4, 9, 26, 14, 13], we describe prosody using pitch and intensity. The utterances were recorded from a male adult with different variations of flat and emphasized prosody. We used the Praat speech analysis tool [3] to analyze the recorded utterance, and guarantee no significant intensity or pitch variations in the weak prosody utterances, while for strong prosody cases, there was at least a 10dB change of intensity and a 100Hz change of pitch for the highlighted lexeme. To differentiate the scenario, we use x-y notation, where x indicates the gesture, and y indicates the prosody. Thus we have 8 utterances, D-W, D-S, M-W, M-S, I-W, I-S, B-W and B-S.

3.3 Alignment Timing

To include the misalignment of gesture w.r.t. the utterance, gestures were placed on the lexeme, with 7 different offsets: -0.6s, -0.4s, -0.2s, 0s, +0.2s, +0.4s, +0.6s. Thus 7 motion clips were used in total for each utterance to verify the perceived naturalness. Each clip lasts about 10 seconds and contains one gesture in the utterance.

3.4 Experiment Execution

For our experiment, a male virtual character rendered in Maya[8] is used to match the voice for the gesture performance. The character's face is blocked, and thus neither asynchronous lip movement nor facial expression will affect subjects' judgment. A front viewpoint was chosen which displayed the character from the knee to head. The clips are generated at size 640 x 480 with high

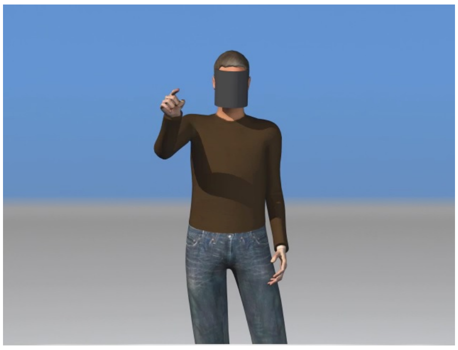


Fig. 1. "Deictic" gesture in the utterance T1.

quality, see Figure (1).

We conduct our experiments by putting all the movie clips on mechanical turk. Subjects watched each clip and were then asked to rate the naturalness of the character's behavior. A 7-point Likert rating scale was used, with 1 indicating "least natural", and 7 indicating "most natural". We encouraged subjects to try to use the entire rating scale. Subjects were allowed to replay the clips as many times as they wanted.

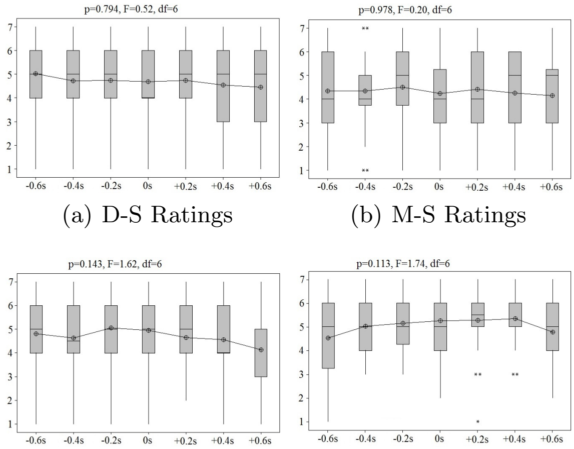
3.5 Result

40 subjects participated in our perception study. For each utterance, we ran the ANOVA to check the perceived difference of the gesture timing factor. No significant rating differences were found over the alignment timing factor for all the 8 utterances. Results for strong prosody cases are illustrated in Figure 2. Thus we conclude, given a single clip, subjects do not have adequate sensitivity to differentiate misaligned gestures.

4 Experiment 2: Perception of gesture misalignment with paired clips

4.1 Experiment Design

We maintained the gesture type, prosodic emphasis and alignment timing, and re-organize the single clips from Experiment 1 into comparison pairs. According to previous research [11, 15, 16], we assume -0.2s is the most natural gesture timing (gesture occurring slightly before the word), and pair clips of "-0.2s" with clips of "-0.6s", "-0.4s", "0s", "+0.2s", "+0.4s" and "+0.6s", given the same



(a) D-S Ratings

(b) M-S Ratings

(c) I-S Ratings

(d) B-S Ratings

Fig. 2. Naturalness ratings for single clips with strong prosody.

utterance. Paired clips were shown in different left and right orders to eliminate an order effect. For each utterance, 12 pairs of movie clips were generated, see Figure (3).

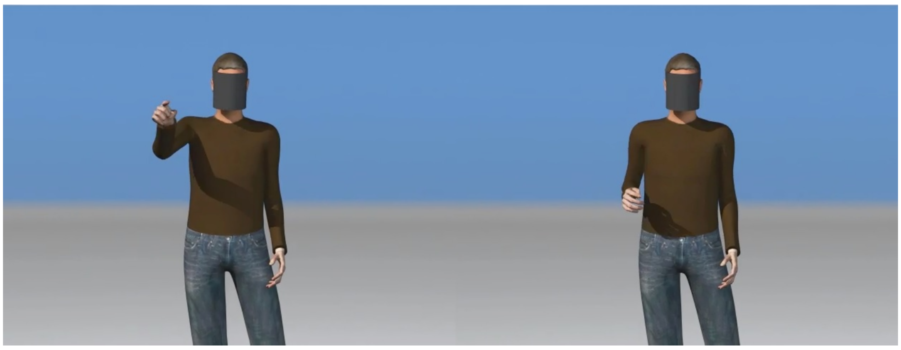


Fig. 3. Side-by-side comparison of “deictic” misalignment.

We conducted our experiments by putting the paired movie clips of all utterances on mechanical turk. To evaluate the perceived naturalness, subjects were asked to choose the more natural one by selecting a relative naturalness rating between the two clips. Five options were offered, “Strongly prefer the left clip”, “Slightly prefer the left clip”, “Almost the same”, “Slightly prefer the right clip” and “Strongly prefer the right clip”, with numeric scores -2, -1, 0, 1 and 2. A naturalness rating was calculated from subject’s choice, 1 point will be added if the clip is slightly preferred, 2 point will be added if strongly preferred, and 0 point if not preferred or subjects cannot tell the difference.

4.2 Result

40 subjects were recruited for the perception task. For paired clips in the movie, we run the 2-sample T test on their ratings to check subjects' preference. Due to the use of multiple T-tests, Bonferroni correction was used, which sets the significance at level 0.001. In general, people were more likely to detect late gestures as being unnatural, and tend to give higher natural ratings to early gestures. However, the detailed situation differs based on gesture types and prosodic emphasis, which we will discuss separately. The ratings from different left/right orders for the same clip are combined, as our ANOVA does not detect significant effect on left/right ordering.

D-W and D-S: For "deictic" gesture, results for weak prosody and strong prosody are listed in Table 1 and Table 2. With weak prosody, subjects can only detect the difference when "deictic" gesture is more than +0.2s later than the lexeme, otherwise, they don't differentiate the gesture alignment with the -0.2s alignment. Given strong prosody, the rating difference is significant between -0.2s alignment and 0s alignment, which indicates subjects strongly prefer the "deictic" gesture being placed 0.2s earlier than the utterance rather than exactly on the utterance. In both prosody settings, subjects prefer earlier "deictic" gesture, but do not really care too much how much earlier it is within the time window -0.6s to -0.2s.

M-W and M-S: For "metaphoric" gesture, results for weak prosody and strong prosody are listed in Table 3 and Table 4. Subjects can detect the all the late alignment of "metaphoric" gestures given the weak prosody. However, when using the strong prosody utterance, subjects no longer detect the difference between the -0.2s alignment and the 0s alignment.

I-W and I-S: For "iconic" gesture, results for weak prosody and strong prosody are listed in Table 5 and Table 6. Given weak prosody, the rating difference between -0.2s alignment and early -0.4s alignment and all late placements are significant. However, when using strong prosody utterance, subjects do not differentiate -0.2s alignment with 0s alignment, nor do they notice the difference between -0.2s, -0.4s and -0.6s alignment. In weak prosody settings, subjects indicate preference for earlier "iconic" gesture placement, and the -0.4s alignment is more preferable than -0.2s alignment.

B-W and B-S: For "beat" gesture, results for weak prosody and strong prosody are listed in Table 7 and Table 8. Subjects' sensitivity to gesture alignment does not vary too much given 2 different prosody settings. Late "beat" gesture can be detected, while earlier placement does not seem to differ too much according to subjects' ratings.

Between Gesture Types: Our findings in the experiments verify that gestures performed earlier than its spoken text are perceived as more natural. Between different gesture types, the situations are not exactly the same. The influence of prosody on the perceived naturalness varies for different gesture types. Strong prosody can sharpen subjects' sensitivity to gesture misalignment. For the "deictic" and "beat" gestures, subjects were more able to detect misaligned clips when the prosody was strong. For the "metaphoric" and "iconic" gestures, in the weak prosody case, there was an example of an earlier -0.4s alignment being preferred over the -0.2s alignment. This did not happen in the strong prosody case. So the prosody signal may have weakened the preference for an early gesture. It can be explained as "metaphoric" and "iconic" gestures are more likely to contain information not tightly coupled with the prosodic emphasis, while "deictic" and "beat" have sharper forms which are more related to the voice.

5 Discussion and Conclusion

This paper summarizes a series of studies exploring people's sensitivity to gesture alignment with text and how this varies as prosody changes. When shown a single clip with a strong prosody signal, our subjects on Mechanical Turk were not able to reliably detect misalignment, even given relatively large misalignments of 0.6s. However, when shown side-by-side clips, subjects generally viewed gestures occurring later, greater than 0.2s after the lexical affiliate, less favorably. Results were more mixed for gestures occurring early, and in some cases, gestures occurring 0.4s before the lexical affiliate were preferred to those occurring 0.2s before the lexical affiliate, our presumed best alignment.

With regards to prosody, the picture is complex. We had hoped to find a clear indication that when prosody was strong, people had higher demands for alignment. This does appear to be true for "deictic" and "beat" gestures, but the opposite picture emerged for the "metaphoric" and "iconic" examples. This is perhaps not surprising as "deictic" and "beat" gestures may both have sharper forms, more tightly coupled with the emphasis in the voice, whereas "metaphoric" and "iconic" gestures are more likely to contain information not copied in the speech. This may be an area worth further study.

In terms of agent design, it would appear that it is not necessary to maintain tight alignment, as people seem to have low sensitivity to this, especially if seeing only one clip. Where variation from the speech is allowed, it seems clear that it is preferable to move the gestures earlier in time, not later.

Table 1. Average ratings and T test results from paired clips for utterance D-W.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.431) -0.6s (0.444) T=-0.14, P=0.892	-0.2s (0.284) -0.4s (0.432) T=-1.59, P=0.114	-0.2s (0.465) 0s (0.31) T=1.56, P=0.12	-0.2s (0.696) +0.2s (0.261) T=3.91, **P<0.0001**	-0.2s (0.922) +0.4s (0.195) T=6.87, **P<0.0001**	-0.2s (1.139) +0.6s (0.153) T=9.11, **P<0.0001**

Table 2. Average ratings and T test results from paired clips for utterance D-S.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.608) -0.6s (0.405) T=1.81, P=0.072	-0.2s (0.239) -0.4s (0.451) T=-2.4, P=0.018	-0.2s (0.639) 0s (0.449) T=4.61 **P<0.0001**	-0.2s (0.973) +0.2s (0.151) T=8.01, **P<0.0001**	-0.2s (1.141) +0.4s (0.141) T=9.08, **P<0.0001**	-0.2s (1.040) +0.6s (0.133) T=8.34, **P<0.0001**

Table 3. Average ratings and T test results from paired clips for utterance M-W.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.284) -0.6s (0.608) T=-3.22, P=0.002	-0.2s (0.227) -0.4s (0.507) T=-3.01, P=0.003	-0.2s (0.581) 0s (0.230) T=3.69, **P<0.0001**	-0.2s (0.693) +0.2s (0.187) T=5.27, **P<0.0001**	-0.2s (0.760) +0.4s (0.263) T=4.49, **P<0.0001**	-0.2s (0.895) +0.6s (0.250) T=5.53, **P<0.0001**

Table 4. Average ratings and T test results from paired clips for utterance M-S.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.5) -0.6s (0.474) T=0.23, P=0.818	-0.2s (0.321) -0.4s (0.385) T=-0.68, P=0.498	-0.2s (0.52) 0s (0.267) T=2.58, P=0.011	-0.2s (0.72) +0.2s (0.267) T=4.23 **P<0.0001**	-0.2s (0.933) +0.4s (0.160) T=7.05, **P<0.0001**	-0.2s (0.960) +0.6s (0.213) T=6.77 **P<0.0001**

Table 5. Average ratings and T test results from paired clips for utterance I-W.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.347) -0.6s (0.533) T=-2.06, P=0.041	-0.2s (0.186) -0.4s (0.5) T=-3.45, **P=0.0001**	-0.2s (0.568) 0s (0.149) T=5.17, **P<0.0001**	-0.2s (0.84) +0.2s (0.16) T=7.3, **P<0.0001**	-0.2s (1.054) +0.4s (0.108) T=10.37, **P<0.0001**	-0.2s (1.068) +0.6s (0.137) T=8.77, **P<0.0001**

Table 6. Average ratings and T test results from paired clips for utterance I-S.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.349) -0.6s (0.538) T=-2.22, P=0.028	-0.2s (0.418) -0.4s (0.439) T=-0.2, P=0.839	-0.2s (0.422) 0s (0.328) T=0.94 P=0.347	-0.2s (0.721) +0.2s (0.246) T=4.0 **P<0.0001**	-0.2s (0.723) +0.4s (0.292) T=3.63, **P<0.0001**	-0.2s (0.885) +0.6s (0.262) T=4.87, **P<0.001**

Table 7. Average ratings and T test results from paired clips for utterance B-W.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.346) -0.6s (0.538) T=-1.94, P=0.054	-0.2s (0.25) -0.4s (0.438) T=-2.05, P=0.042	-0.2s (0.539) 0s (0.276) T=2.75, P=0.007	-0.2s (0.797) +0.2s (0.114) T=8.21, **P<0.0001**	-0.2s (1) +0.4s (0.132) T=8.28, **P<0.0001**	-0.2s (1.208) +0.6s (0.091) T=11.31, **P<0.0001**

Table 8. Average ratings and T test results from paired clips for utterance B-S.

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.449) -0.6s (0.487) T=-0.37, P=0.712	-0.2s (0.244) -0.4s (0.423) T=-2.22, P=0.028	-0.2s (0.519) 0s (0.247) T=2.91 P=0.004	-0.2s (0.857) +0.2s (0.065) T=9.57 **P<0.0001**	-0.2s (1.039) +0.4s (0.143) T=8.68, **P<0.0001**	-0.2s (1.138) +0.6s (0.125) T=9.47, **P<0.0001**

6 Acknowledgments

Financial support for this work was provided in part through NSF grants IIS 1115872 and IIS 0845529.

References

1. Adolphs, R.: Neural systems for recognizing emotion. Current opinion in neurobiology 12(2), 169–177 (2002)
2. Albrecht, I., Haber, J., Seidel, H.P.: Automatic generation of non-verbal facial expressions from speech. In: Advances in Modelling, Animation and Rendering, pp. 283–293. Springer (2002)
3. Boersma, P.: Praat, a system for doing phonetics by computer. Glot international 5(9/10), 341–345 (2002)
4. Bregler, C., Covell, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. pp. 353–360. ACM Press/Addison-Wesley Publishing Co. (1997)
5. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 477–486. ACM (2001)
6. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. ACM Transactions on Graphics (TOG) 24(2), 331–347 (2005)
7. Efron, D.: Gesture and Environments. King's Crown Press (1941)
8. Inc., A.: Maya, 3d computer graphics software (2008)
9. Ju, E., Lee, J.: Expressive facial gestures from motion capture data 27(2), 381–388 (2008)
10. Kendon, A.: Current issues in the study of gesture. The Biological Foundations of Gestures: Motor and Semiotic Aspects pp. 23–47 (1986)
11. Kirchhof, C.: On the audiovisual integration of speech and gesture. The 5th Conference of the International Society for Gesture Studies (ISGS) (2012)
12. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer animation and virtual worlds 15(1), 39–52 (2004)
13. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Transactions on Graphics (TOG) 29(4), 124 (2010)
14. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. In: ACM Transactions on Graphics (TOG). vol. 28, p. 172. ACM (2009)
15. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
16. McNeill, D.: Gesture and thought. University of Chicago Press (2008)
17. Montepare, J., Koff, E., Zaitchik, D., Albert, M.: The use of body movements and gestures as cues to emotions in younger and older adults. Journal of Nonverbal Behavior 23(2), 133–152 (1999)

18. Morency, L.P., Sidner, C., Lee, C., Darrell, T.: Head gestures for perceptual interfaces: The role of context in improving recognition. Artificial Intelligence 171(8), 568–585 (2007)
19. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) 27(1), 5 (2008)
20. Rimé, B., Schiaratura, L.: Gesture and speech. (1991)
21. Sargin, M.E., Erzin, E., Yemez, Y., Tekalp, A., Erdem, A., Erdem, C., Ozkan, M.: Prosody-driven head-gesture animation. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. vol. 2, pp. II–677. IEEE (2007)
22. Schröder, M.: Speech and emotion research
23. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: Creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics (TOG) 23(3), 506–513 (2004)
24. Terken, J.: Fundamental frequency and perceived prominence of accented syllables. The Journal of the Acoustical Society of America 89, 1768 (1991)
25. Wallbott, H.G.: Bodily expression of emotion. European journal of social psychology 28(6), 879–896 (1998)
26. Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S.: The timing of speech-accompanying gestures with respect to prosody. In: Proceedings of the International Conference: From Sound to Sense. vol. 50, pp. 10–13 (2004)