Paralinguistic Behaviors in Dialog as a Continuous Process

David Novick

Department of Computer Science, The University of Texas at El Paso, El Paso, TX, USA

novick@utep.edu

Abstract

Prior research on gaze, turn-taking, and backchannels suggests that the speaker's gaze cues the listener's paralinguistic responses, including feedback behaviors. To explore how conversants use feedback cues and responses, I studied a corpus of face-to-face conversational interaction, primarily using a conversation-analytic approach. Analysis of the dialogs suggests that paralinguistic behaviors express meaning at a level of granularity often smaller than dialog control acts. Behaviors such as gaze and nodding can be seen as continuous rather than discrete actions. Moreover, speaker gaze shift toward the listener is a polysemous expression that can cue a range of behaviors in the listener, including continued attention, head nods as backchannels, utterances as backchannels, and turn-taking. The analysis also suggests that gaze, from both speakers and listeners, can express a state rather than a discrete act.

Index Terms: dialog, grounding, feedback, gaze, nod

1. Introduction

Humans and embodied conversational agents appear to converse more effectively when the agents appear to sense and produce paralinguistic behaviors such as gaze shifts and head nods. Conversants use nods more often when an agent's feedback indicates that it perceives the nods [1]. Agents using human-like patterns of interaction are better appreciated by human conversants and contribute to more efficient interaction [2]. Consequently, more natural models of feedback behaviors should lead to even better interaction.

The dialog functions of paralinguistic behaviors, such as gaze and nods, can be expressed in terms of dialog control acts analogous to speech acts. Both David Traum [2] and I [4] have described act-based models that include dialog control acts such as "take turn." These models' discreteness makes them useful for computational representation and implementation, and they can be applied to action at a sub-utterance level. For example, gaze can be modeled as grounding at the level of intonation phrases, where speakers actively monitor for positive evidence of understanding [5].

Models of conversation, such as Suchman's model of joint action have, all along, described these processes as continuous:

> Closer analyses of face-to-face communication indicate that conversation is not so much an alternating series of actions and reactions between individuals as it is a joint action accomplished through the participants' continuous engagement in speaking and listening [references omitted]. [6, p. 71]

Suchman's model, though, was continuous at the level of the conversants' contributions—a succession of discrete, interacting verbal responses rather than a moment-by-moment interplay of verbal and non-verbal. Successful embodied conversational agents will have to possess the ability to perceive, understand, and communicate through genuinely continuous processes that reflect the fine-grained dynamics of actual conversation and the moment-by-moment judgments of speakers about listeners' understanding.

An incremental approach to interaction has been implemented at least on the generation side [7]. That is, the agent displays multimodal paralinguistic behaviors even though it cannot sense these behaviors in the human conversant. But it is the listener who moderates the speaker's production, using nonverbal means. As Heylen pointed out,

> [T]he behaviors displayed by auditors is an essential determinant of the way in which conversations proceed. By showing displays of attention, interest, understanding, compassion, or the reverse, the auditor/listener, determines to an important extent the flow of conversation, providing feedback on several levels. [8, p. 82]

The research on the relationships between gaze, turn-taking, and backchannels suggests that the speaker's gaze cues the listener's paralinguistic responses, including feedback behaviors. Speakers often use gaze to cue turn-exchanges by shifting gaze to the listener [9, 10], and speakers can use gaze shifts to cue backchannels, both verbal and nonverbal [11].

To improve my understanding of feedback cues and responses as a continuous process, I studied a corpus of face-to-face conversation. Because I was interested in exploring the micro-components of paralinguistics, my analysis was more in the tradition of conversation analysis than discourse analysis, complemented by reviewing all of the conversations between Americans in the entire corpus. My analysis of these interactions suggests that:

- Conversants vary widely with respect to feedback behaviors.
- Both speakers and listeners can produce multiple paralinguistic behaviors within single intonation phrases.
- While listeners sometimes nod while the speaker is looking away, they typically nod when or shortly after the speaker looks at the listener.
- Listener gaze aversion (i.e., the end of continued attention) can signal understanding.
- Speaker gaze shift toward the listener is a polysemous expression that can cue a range of behaviors in the listener, including continued attention, head nods as backchannels, utterances as backchannels, and turn-taking.

- Gaze, from both speakers and listeners, can express a state rather than a discrete act.

Moreover, if paralinguistic behaviors are really expressing states rather than acts, and if the behaviors are still to be viewed in the perspective of speech acts (along the lines of meta-acts or dialog control acts), then speech-act theory will have to accommodate expression of being. In the balance of this paper, I present the evidence—mostly conversation analytic—for these conclusions and discuss their implication for conversation-act models.

2. Observations

To explore how conversants use feedback cues and responses as a continuous process, I turned to the UTEP-CIFA corpus [12] of face-to-face conversational interaction. These conversations were recorded as part of a study of proxemics and trust, comparing behaviors between native speakers of American English and native speakers of Iraqi Arabic. For the purposes of this research, I limited my study to the twelve dialogs conducted by the eight American conversants. Each dialog was about four minutes long, for a total of about 48 minutes of conversation.

2.1. Differences among conversants

When the UTEP-CIFA corpus was collected, our research team annotated the dialogs for gaze, hand movements, and head nods. Analysis of the annotations indicates that American conversants produced, on average, 6.90 nods per minute, with a standard deviation of 2.17 nods per minute. As the standard deviation would suggest, the variation in nod rates among the dialogues was high, with two dialogs fewer than 4 nods per minute and three dialogs with more than 10 nods per minute. The variation among dialogs reflects variation among the individual conversants, each of whom participated in three dialogs. The mean, minimum and maximum rate of nods per minute across the conversants were 6.90, 4.84 and 10.07, respectively.

Analysis of the annotations also disclosed similarly wide differences with respect to the amount of time that the conversants gazed at their conversational partner. The mean amount of gaze time per minute was 16.64 seconds, but the standard deviation was 7.66 seconds, and the minimum and maximum gaze time per minute across all of the dialogs were 4.47 seconds and 32.42 seconds, respectively. In other words, there were dialogs where one of the conversants almost never looked at the other conversant, and there were dialogs where one of the conversants looked at the other conversant about half the time. Again, the differences among the dialogues reflect differences among the individual conversants, whose average gaze per minute varied from a minimum of 9.08 seconds to a maximum of 2.74 seconds.

These differences among conversants were immediately apparent when viewing the corpus. Some conversants were animated listeners, nodding more or less continuously; others were impassive, rarely nodding, even after the speaker shifted gaze. Some conversants engaged with gaze much of the time; others steadfastly kept their gaze away.

2.2. Multiple head gestures within single intonation phrases

I turn now from discourse analysis to something more along the lines of conversation analysis. I focused on segment of about 30 seconds in dialog P5 of the corpus; I transcribed the verbal and nonverbal actions of the segment by hand, viewing each moment of the conversation perhaps a dozen times. Figure 1 shows my transcript of this dialog segment.

A Verbal	A Nonverbal	Time	B Nonverbal
um the uh	gaze: away	00:15:10	gaze: A
the most recent uh biggest influence on my			
life	gaze: B	00:19:10	succession of four little nods
the group			
um	gaze: away	00:20:10	
and when I heard the word group I used to uh			
I just finished a	brief head tilt away	00:23:00	
six-year tour with the	gaze: B	00:24:00	
nine-oh-second military	brief head tilt away	00:25:05	
intelligence		00:26:10	gaze: away, and two nods
group	gaze: away	00:26:25	succession of three diminishing nods
at Fort Meade			
and that was my first			
chance to work	gaze: B	00:30:15	
with the military folks	two little nods	00:31:10	succession of three little nods
and um	gaze: away		
army and I got such			
a sense of dedication	gaze: B	00:37:05	
and loyalty to the United States		00:38:05	succession of four or five little nods

Figure 1. *Partial transcript of dialog P5.*

The transcript, especially at 00:24:00-00:25:00 and 00:30:15-00:31:10, shows complex bursts of paralinguistic behaviors from both conversants. Moreover, my transcription does a poor job of conveying the continuous, animated quality of the interaction from the speaker more or less all the time, and from the listener when—aside from the case I discuss in the next subsection—the speaker's gaze is directed at him. In any event, these combinations of activity, within a single intonation phrase unit, include behaviors such as shifting gaze, tilting the head to side away from the other speaker, and, untranscribed because the actions are rather subtle, the suggestion of a couple of nods—all within about a second. On the part of the listener, the combinations are less complex but typically include successions of small nods, or nodding plus gaze aversion.

2.3. Gaze shift as a cue for nodding

Consistent with the behaviors described in [10, 13], the listeners in the transcribed segment and in the overall corpus generally nodded when the speaker shifted gaze to the listener. While this was subject to the variation among conversants with respect to overall frequency of nodding, when listeners did nod it was almost always just after a gaze shift toward the listener, and rarely otherwise. For example, when the speaker shifts his gaze to the listener at 00:19:10, the listener immediately produces a succession of small nods.

This pattern has a plausible, if prosaic, explanation: if the speaker is not looking at you, it does not do much for you to nod because the speaker may not (cf. peripherally) see your action. So if you want to signal grounding via head nods, your first opportunity to do so is when the speaker shifts gaze to you.

Moreover, if you want the speaker to continue but do not want to or cannot signal grounding of the speaker's preceding speech, then you really should not nod. This gives the speaker the opportunity to elaborate or clarify, after which the listener can then nod if he or she wants to signal grounding.

Figure 2 shows the dialog at exactly this sort of point. At 00:24:15, Conversant A, on the left, has just shifted his gaze to Conversant B, and conversant B is still looking at A without

nodding (in contrast to the immediate nod responding to the speaker's toward-listener gaze shift at 00:19:10). My interpretation of the dialog at 00:24:15 is that the speaker's fragmentary utterances ("the group um and when I heard the word group I used to uh I just finished a six-year tour with the") have left the listener in a position where he is struggling to understand the listener's meaning. So when the speaker shifts his gaze to the listener, the listener does not immediately nod. Rather, the speaker continues production of the utterance ("nine-oh-second military intelligence"). This apparently helps the listener grasp the speaker's meaning, and the listener then, two seconds after the speaker's gaze shift, nods.

Figure 2. *Conversant A (on left) shifts gaze to Conversant B, who waits about two seconds before nodding.*

Actually, the listener not only nods, but he averts his gaze as he does so, as shown in Figure 3. This behavior has a logic to it. The listener's non-nodding continued attention was signaling non-understanding, which is not the usual case for continued attention, which Clark and Schaefer [14] listed as a weak form of acceptance. So to signal understanding the listener has to change his gaze behavior and thus averts his gaze while nodding. In other words, the lack of initial nod transforms the listener's continued attention into lack of acceptance, and so to signal acceptance the listener has to end his continued attention. In fact, this pattern occurred across different pairs of conversants.

2.4. Gaze shift as a polysemous cue

In the corpus, I observed the speaker's gaze shift toward the listener cue backchannel nodding. But I also observed the same sort of gaze shift lead to a range of listener paralinguistics: continued attention (as in Section 2.3), nodding as backchannels (also as in Section 2.3), verbal backchannels, turn exchanges. That is to say, the speaker's gaze shift toward the listener is a polysemous cue, in that it can cue any one of these four behaviors in the listener. This suggests that gaze shift is an action rather than act: it is a nonverbal behavior that has meaning as a dialog control act in the context of the interaction and of the conversants' respective intents, much in the same way that an individual word or expression is not an act in itself but rather becomes an act when interpreted in context.

Part of the context for assigning meaning to gaze shifts consists of the speaker's prosody, which may differentiate the nonverbal action into more specific acts through, for example, prosodic patterns for backchannel cues (see, e.g., [15]). Another part of the context involves the conversants', and especially the listener's, state of mind with respect to acceptance and grounding: no matter how clear the speaker's cue, a listener who is not understanding the speaker would usually be ill-served by signaling acceptance. And part of the context involves the actual content of the dialog: if the speaker has apparently completed a contribution to the conversation, the listener can take the turn.

Figure 3. *Conversant A continues to gaze to Conversant B, who averts his gaze and nods twice.*

2.5. Gaze as an expression of state

The meta-act or dialog-control-act model of interpreting paralinguistic behaviors still has both utility and intuitive appeal, as it explains what conversants are *doing*. At the same time, though, even the 30-second segment of the dialog corpus analyzed here leads to questions about the discreteness of the model:

- If a listener is continuously gazing at the speaker, what is or was the listener's act? Did the act occur when the speaker first gazed at the listener? Is there still an act some seconds later when the listener remains gazing at the speaker?
- If a speaker shifts gaze to the listener and holds this gaze, what is or was the speaker's act? Did the speaker produce an "invite backchannel" act when he or she shifted gaze? How can the invitation still be in force as the speaker continues to gaze at the listener, as at 00:24:15 of dialog P5?
- If a listener, after not following through on an invitation to backchannel, continues to gaze at the speaker, presumably responding with an invitation for the speaker to clarify or elaborate, what is or was the listener's act? Does the act still continue as long as the listener holds his gaze under these circumstances?

In light of these questions I suggest that gaze, and probably other paralinguistic behaviors such as continuous nodding, can be understood as expressing a state of being rather than expressing an act. The state of the continuously gazing accepting listener is something like "I am following what you are saying and invite you to continue." This is a continuous rather than a discrete phenomenon; the state has a beginning and an end, but for its duration it is a continuing proposition. Similarly, the state of the speaker, once having shifted gaze to the listener, holding his gaze toward the listener (perhaps in search of feedback), is

something like "I am speaking and would like to see from you a positive signal of understanding." Again, the speaker's proposition is a continuous one. Finally, the state of the continuously gazing non-accepting listener is something like "I am hearing you but not yet able to ground your current contribution." This, too, is a state of being rather than an act.

For representation of paralinguistic behaviors as dialogue control acts, then, the act model will have to be extended to include continuous states of being. In other words, at each moment that a listener is normally gazing at a speaker, the listener's state is not just the ongoing action of "continued attention" but rather a state given meaning by the context, for example, "I invite you to continue speaking."

2.6. Implications for speech-act theory

If, as I suggested above, dialog acts should be extended encompass to states of being, in addition to discrete acts, then perhaps traditional speech acts should be extended similarly. For example, if a speaker says "I am hereby ready to sign the contract," it is the case the speaker has not just commented on his or her own status but actually stands ready to sign. In other word, the speaker is now in a state of being willing to sign the contract, and this state remains in effect until ended by another act from the speaker or some relevant change in circumstances. Or, even more to the point, imagine that the speaker says "I am hereby ready to sign the contract" and extends a hand while holding a pen. While the hand remains extended, the speaker appears to be in a state of willingness to sign. The speaker's state has a sort of continuing illocution. When the hand is withdrawn, the state of willingness to sign appears to end, and the illocution ends with it.

3. Conclusions

Analysis of face-to-face dialogs suggests that paralinguistic behaviors express meaning at a level of granularity often smaller than dialog control acts such as "take turn." Behaviors such as gaze and nodding can be seen as continuous rather than discrete actions.

While it is true that conversants vary widely with respect to the extent they use feedback behaviors such as gazing and nodding, both speakers and listeners can produce multiple paralinguistic behaviors within single intonation phrases.

While listeners sometimes nod while the speaker is looking away, they typically nod when or shortly after the speaker looks at the listener. But when continued gaze without nodding means that the listener is not accepting the speaker's current contribution, gaze aversion by the listener (i.e., the end of continued attention) can be part of the listener's signaling of understanding.

As is apparent from the interaction in the corpus, speaker gaze shift toward the listener can cue a range of behaviors in the listener, including continued attention, head nods as backchannels, utterances as backchannels, and turn-taking. In other words, gaze shift can have multiple meanings and effects, and these meanings and effects depend on prosody, context, and intention.

Gaze, and probably other paralinguistic behaviors such as continuous nodding, from both speakers and listeners, can be understood as expressing a state of being rather than expressing a discrete act. The model of dialog control acts, and perhaps speech act theory more generally, may have to be extended to accommodate expression of state of being.

In future work, we plan to extend the analysis of the corpus—and probably other available corpora that record naturally occurring interaction—to verify more systematically the observations of this paper that arose from a conversation-analytic approach. We also plan to test the finer-grained or continuous model of dialog control through experiments with embodied conversational agents.

4. References

[1] Morency, L.-P., Context-based visual feedback recognition, Computer Science and Artificial Intelligence Laboratory Technical Report MIT-CSAIL-TR-2006-075, Massachusetts Institute of Technology, 2006.

[2] Heylen, D.K.J. and van Es, I. and Nijholt, A. and van Dijk, E.M.A.G., "Controlling the gaze of conversational agents," Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic Publishers, 245-262, 2005.

[3] Traum, D. R. and Hinkelman, E. A. "Conversation acts in task-oriented spoken dialogue," Computational Intelligence, 8(3): 575–599, 1992.

[4] Novick, D., "Controlling interaction with meta-acts," Conference on Human Factors in Computing Systems (CHI 91), New Orleans, LA, May, 1991, 495, 1991.

[5] Nakano, Y.I., Reinstein, G., Stocky, T., and Cassell, J., "Towards a model of face-to-face grounding," Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL '03), Stroudsburg, PA, USA, 553-561, 2003.

[6] Suchman, L. A., Plans and situated actions. Cambridge: Cambridge University Press, 1987.

[7] Kopp, S., Stocksmeier, T., Gibbon, D.: "Incremental multimodal feedback for conversational agents, Pelachaud, C. et al. (eds.): Intelligent Virtual Agents '07, LNAI 4722, Springer-Verlag, 139-146, 2007.

[8] Heylen, D., "Multimodal backchannel generation for conversational agents, Workshop on Multimodal Output Generation," MOG 2007, Aberdeen, Scotland, January 25-26, 2007, 81-92, 2007.

[9] Novick, D., Hansen, B., and Ward, K., "Coordinating turn-taking with gaze," Proceedings of ICSLP-96, Philadelphia, PA, October, 1996, 3, 1888-91, 1996.

[10] van Es, I., Heylen, D., van Dijk, B., and Nijholt, A., "Making agents gaze naturally - Does it work?" Proceedings AVI 2002: Advanced Visual Interfaces, Trento, Italy, May 2002, 357-358, 2002.

[11] Timmerman, A., "Backchannels must be seen," 13th Twente Student Conference on IT, Enschede, The Netherlands, June 21, 2010.

[12] Flecha-Garcia, M., Novick, D., and Ward, N., Differences between Americans and Arabs in the production and interpretation of verbal and non-verbal dialogue behaviour, Speech and Face-to-Face Communication Workshop, Grenoble, France, October 27-29, 2008, 47-48, 2008.

[13] Huang, L., Morency, L.P., and Gratch, J., "A multimodal end-of-turn prediction model: Learning from parasocial consensus sampling," Tenth International Conference on Autonomous Agents and Multiagent Systems, May 2011.

[14] Clark, H., and Schaefer, E., "Contributing to discourse," Cognitive Science, 13, 259-294, 1989.

[15] Rivera, A., and Ward, N., "Prosodic features that lead to back-channel feedback in Northern Mexican Spanish," Proceedings of the Seventh Annual High Desert Linguistics Society Conference, Albuquerque, NM, 19-26, 2008.