Towards predicting dialog acts from previous speakers' non-verbal cues

Matthew Roddy and Naomi Harte
ADAPT Centre, School of Engineering
Trinity College Dublin, Ireland

I. Background

In studies of response times during conversational turn-taking, a modal time of 200 ms has been observed to be a universal value that exists across languages and cross-culturally [1][2]. This 200 ms value is also seen as the limit of human response times to any stimulus (e.g the response time to a starting-gun in a race). It has also been shown that human language production is slow and can take up to 1500 ms to generate even a short clause [3]. Due to these two observations, it is necessary for a person to start formulating their turns long before the end of their interlocutor's turn. To do this we must predict elements of what a person will say in order to formulate our responses and sustain the flow of conversation. In this sense, the end of a person's turn can be viewed as a trigger for a prepared response [2]. This model of human language production informs incremental approaches to the design of dialog systems, where dialog options are evaluated incrementally, while the system processes user utterances [4].

One way we can form our predictions is by reading the non-linguistic signals that are produced by our interlocutor. For example, prosodic information such as pitch inflection can be used to infer whether a question is being asked or a statement is being made [5]. Pitch and intensity information can also be used to infer whether a backchannel is an appropriate response [6]. These backchannel prediction models based on non-linguistic cues can be used by conversational agents to carry out more fluid interactions with users [7]. The development of better prediction models that exploit the social signals that humans use will lead to agents that can reproduce the interaction behaviors of humans more effectively.

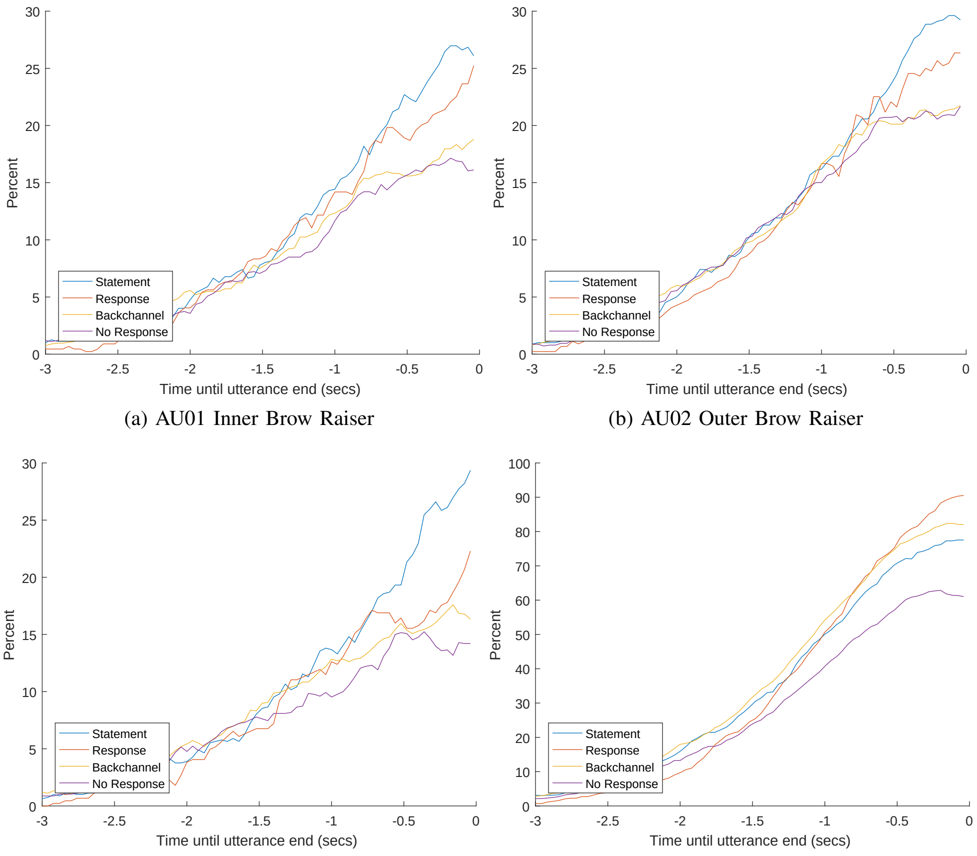
In this analysis we look at non-verbal speaker signals that can be used to predict the appropriate dialogue act that will follow the speaker's utterance. We define three categories of dialogue acts: (1) response (as in a response to a question), (2) statement (a general turn switch which does not include other dialog act types), and (3) backchannel (vocalizations encouraging the speaker to continues speaking). In addition we define a fourth category, no-response, which is not strictly a dialogue act but is a relevant category for agent interactions. We identify four types of non-verbal signals that can be used to predict the appropriate type of response dialogue act: inner eyebrow movement, outer eyebrow movement, blinks, and gaze. We analyze the behavior of these four signals in the vicinity of the dialogue acts.

II. Methods

The data set used in this study is the IFA Dialog Video Corpus [8] which consists of dyadic, face-to-face conversations in Dutch. We use a subset of 9 conversations, each lasting 15 minutes. The data set is annotated for utterances (IPUs) as well as dialog acts such as backchannels, questions, and responses. The amount of instances of each dialogue act type were: 797 statements, 444 responses, 1347 backchannels, and 1260 no-responses. The modes of the turn-lengths directly preceding the new dialogue acts were approximately 1.2 seconds for all types of dialogue act responses. The data set also includes binary gaze annotations for when each person is looking at their interlocutor. We use these annotations for our analysis of gaze.

The free-form conversations were recorded using two video cameras, one camera per person, focused on their faces. We use OpenFace [9] to automatically estimate facial action units (AUs) of the subjects. The values for the AUs are binary values that represent the presence or absence of the given AU. In this analysis we investigate three AUs: AU01 (inner brow raiser), AU02 (outer brow raiser), and AU45 (blinks). It is worth noting that while the gaze annotations can be considered reliable as they were hand annotated, the extracted action units should be considered less reliable due to factors such as lighting conditions, camera angles, and differences between the training data used to create the automatic system and our data set.

To analyze the four different features we first locate instances of each dialogue act and then examine the behavior of the features in the other person leading up to that (plots shown in Fig. 1). So if person B produces a backchannel (given that person A is the first speaker and person B is the second) we analyze the behavior in person A's turn directly leading up to that backchannel. We use a three second window leading up to the end of person A's utterance. In our calculations we only include the presence of a given feature during person A's last utterance within that three second window. Features that existed during earlier utterances that still lie within the 3 second window are not counted as they could perform different communicative functions. For example, when we are calculating the frequency plots for responses, we take the end of all the questions which elicited a response as our reference points. We then take the features (sampled every 40 ms) from the previous 3 seconds leading up to the reference points



(a) AU01 Inner Brow Raiser

(b) AU02 Outer Brow Raiser

(c) AU45 Blink

(d) Gaze at addressee

Fig. 1. The percentage of frames in which each non-verbal feature is observed in speakers leading up to the end of their utterance. The different dialogue acts correspond to those of the subsequent speaker.

and align them to calculate the frequency of the features at each time point within the window. However, if the speaker produces an utterance other than the question during the three seconds, this is not included in the frequency calculations.

III. Discussion

A general observation that can be made about all four features is that the graphs for the dialogue types tend to diverge somewhere between the -1 to -0.5 second marks. This supports the previously discussed statements in the literature that we formulate our responses in advance of our turn. For example, in both of the brow features the statement and response percentages diverge from the no-response and backchannel graphs around the -0.7 second mark. This could imply that the intention to relinquish a speaking turn is manifested by brow movement somewhere in that region. Another general observation is in the order that the different dialog acts appear in the four different feature graphs: the eyebrow graphs have their highest percentage associated with statements, whereas gaze is associated with responses and backchannels. The graphs provide evidence that eyebrows and gaze have different functions in dialogue and could potentially be good predictive features.

It is interesting to note the hierarchy of the dialogue acts in the gaze feature. Responses are associated with the most gaze while backchannels, statements, and no-replies are associated with less (in that order). This suggests that gaze may be an indication of a solicitation of a reply by a speaker. There is also a noticeable dip in the gaze trajectory of the no-response category 0.2 seconds before the turn end. A possible explanation for this dip is speakers indicating that they wish to continue their turn in the next utterance by looking away. Interestingly, a similar dip is also noticeable in the inner brow no-response plot at a similar point in the graph.

A notable aspect of the graphs of the brow features is that they suggest that brow movement occurs less in questions than in normal turn-endings. This has been observed in previous studies such as [10]. Some research has associated eyebrow movement with questions [11][12], which would have manifested itself in higher response percentages. In our graphs, the higher levels of eyebrow movement in statements may be related to the movement's function as a signal of surprise or astonishment [11] during feedback.

Blinks appear to be a good feature for the detection of turn endings. This finding was also observed in [13]. There is also a notable dip in the trend of the response graph around -0.5 sec mark. This could be caused by both participants avoiding blinking due to mutual gaze that is observed during turn-switches [14]. This -0.5 seconds value is similar to the timing of mutual gaze windows reported by Bavelas in [15]. The amount of blinks sharply increases around -0.2 sec which indicates that after the critical point where mutual gaze has occurred, and turn-taking has been agreed upon, blinking is then possible. This also provides evidence of utterance planning as the end of the turn is anticipated.

This preliminary analysis is an investigation into the behavior of these features in conversation. In subsequent work we plan on using these features in conjunction with features from other modalities (e.g prosody, movement, linguistic) to perform dialogue act predictions. We are also interested in the temporal aspects of dialogue act predictions. How soon can we know, within a degree of confidence, that a given dialogue act is appropriate? This information could aid in the design of conversational agents that can plan their responses in advance using incremental approaches.

REFERENCES

[1] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon, and others, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[2] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, Jun. 2015.

[3] S. C. Levinson, "Turn-taking in Human Communication – Origins and Implications for Language Processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, pp. 6–14, Jan. 2016.

[4] D. Schlangen and G. Skantze, ""A General, Abstract Model of Incremental Dialogue Processing"," *Dialogue & Discourse*, vol. 2, no. 1, pp. 83–111, 2011.

[5] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[6] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue." in *INTERSPEECH*, 2009, pp. 1019–1022.

[7] I. de Kok and D. Heylen, "Integrating backchannel prediction models into embodied conversational agents," in *Intelligent Virtual Agents*. Springer, 2012, pp. 268–274.

[8] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV Corpus: A Free Dialog Video Corpus." in *LREC*, 2008, pp. 501–508.

[9] T. Baltrušaitis, P. Robinson, L.-P. Morency, and others, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[10] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English," *Speech Communication*, vol. 52, no. 6, pp. 542–554, Jun. 2010.

[11] I. Eibl-Eibesfeldt, "Similarities and differences between cultures in expressive movements." 1972.

[12] R. J. Srinivasan and D. W. Massaro, "Perceiving Prosody from the Face and Voice: Distinguishing Statements from Echoic Questions in English," *Language and Speech*, vol. 46, no. 1, pp. 1–22, Mar. 2003.

[13] F. Cummins, "Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals," *Language and Cognitive Processes*, vol. 27, no. 10, pp. 1525–1549, Dec. 2012.

[14] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[15] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *Journal of Communication*, vol. 52, no. 3, pp. 566–580, 2002.