Toward a Model for Incremental Grounding in Spoken Dialogue Systems

Thomas Visser[1], David Traum[2], David DeVault[2], and Rieks op den Akker[1]

[1] University of Twente
Enschede, The Netherlands
thomas.visser@gmail.com, h.j.a.opdenakker@utwente.nl

[2] USC Institute for Creative Technologies
Playa Vista, CA, USA
{traum,devault}@ict.usc.edu

Abstract. Recent advances in incremental language processing for dialogue systems promise to enable more natural conversation between humans and computers. By analyzing the user's utterance while it is still in progress, systems can provide more human-like overlapping and backchannel responses to convey their level of understanding and respond more quickly. In this paper, we look at examples of several overlapping response types in human-human dialogues, and present an initial computational model of the incremental grounding process in these responses. Additionally, we describe an implementation of this model in a virtual human dialogue system that can provide backchannels, head nods, frowns, completions and low latency responses.

Keywords: spoken dialogue systems, incremental language processing, grounding

1 Introduction

Effective and fluent conversation requires joint effort from both interlocutors [1], and in spoken human dialogue, this effort is often manifested in real time as speech is happening. While speaking, we monitor the listener's reaction to what we say, and as listeners, we give frequent feedback on what we perceive and understand. Such feedback often overlaps the speaker's ongoing utterance and can take the form of head nods, verbal backchannels, interruptions, and other overlapping responses.

These overlapping responses are important for efficient conversation, and emphasize the incremental nature of human-human communication [2, 3]. For a spoken dialogue system to understand and generate such behaviors, it needs to process speech incrementally. This requires that the processing of user input and planning of system responses occurs frequently, not only while the user speaks, but also while the user listens. While traditional systems employ a rigid turn-taking model, in which overlapping speech is not supported, recent research has

begun to develop some of these incremental processing and response capabilities in implemented systems (e.g., [4–7]). This work has shown that incremental response capabilities can achieve positive effects on user interactions, including user preference over non-incremental systems and increases in perceived human-likeness and efficiency [8, 9], and even increased fluency of user speech [10].

To date, however, implemented systems that model the process of *grounding* in dialogue [11], the process by which interlocutors work to add understood content to their common ground, have not closely linked such incremental response behaviors directly to the grounding model. The system presented in [8] is capable of incremental grounding behavior, but, as pointed out by [12], the domain lacks a notion of utterances and a meaning beyond the surface text. We believe that a grounding model should include the intention and conversational meaning of utterances. In this paper, we take up this project, and present an initial computational grounding model that can connect some of these incremental response behaviors to an incrementally evolving grounding state. We begin in Section 2 by looking at examples of incremental grounding behavior in spoken conversations between human interlocutors. In Section 3, we present our model of the incremental grounding process. Section 4 discusses its implementation within a working spoken dialogue system.

2 Incremental Grounding Behavior in Human Dialogue

Dialogue excerpt (1), from the AMI Meeting Corpus [13], includes two types of incremental grounding behavior.

(1) C: We could just go with um
D*: Yeah
A*: Normal coloured buttons
B: Well do you want colour differentiation here?
C: ...

In the middle of C's sentence, C appears to struggle with how to continue his utterance, uttering a verbal hesitation "um". A then utters "Normal coloured buttons" as a completion of C's partial utterance. The dialogue continues without correction by C, so it is reasonable to assume that this was indeed what C intended to communicate (or was close enough). Meanwhile, D gives a simultaneous backchannel acknowledgment of C's utterance.

An ability to predict the meaning (or perhaps even the surface form) of a partial utterance seems to be on display in such examples. Further, such examples raise a question about the grounding status of the partial utterance and predicted completion. In our model and implementation, we call the content of the partial utterance *explicit* and the content of the utterance completion *predicted*.[3] For grounding purposes, note that A's completion not only demonstrates his understanding of the explicit content of C's utterance, but also of the predicted

[3] It is sometimes useful to distinguish further between the explicit or predicted surface form, as opposed to the explicit or predicted meaning.

content. We assume that a theory of incremental grounding should explain the grounding status of such explicit and predicted content as a dialogue progresses.

Attempted utterance completions do not always exactly match a speaker's intended content or surface form, as in dialogue excerpt (2).

(2) B: That would probably not be in keeping with the um the
C*: *laugh* Technology
B: fashion statement and such, yeah.
C*: Yeah.

In this dialogue, B and C are reflecting on the features and design of the remote control they created. When B shows hesitation ("... with the um"), C decides to help and offers "Technology" as a completion of B's utterance.[4] B however continues his utterance by saying "fashion statement and such", revealing perhaps more precisely what he intended to say. C then issues an overlapping acknowledgment of B's continuation with "fashion statement", by saying "Yeah".

In this example, C's predicted content "Technology" apparently does not exactly match B's original intention. However, it does provide some evidence of understanding of the explicit content of B's partial utterance. We assume that a theory of incremental grounding should explain the status of explicit and predicted content in such cases, and also describe how the grounding state is updated by an overlapping acknowledgment such as C's utterance of "Yeah", here, acknowledging B's continuation with "fashion statement...".

3 A Model For Incremental Grounding

We take as our starting point Traum's computational model of grounding [14], which we summarize briefly here. This model defines seven *grounding acts*: initiate, continue, acknowledge, request repair, repair, request acknowledgment and cancel. Every behavior, either verbal or non-verbal, can convey one or more grounding acts relating to one or more Common Ground Units (CGU). The processing of an utterance in order to update the common ground consists of two steps. The system first has to determine the grounding acts that are being conveyed by the utterance and to which CGUs they apply. Then, the grounding status of the corresponding CGUs is updated.

Traum's theory uses a finite state model that assigns each CGU to one of several states at each point as a dialogue progresses.[5] In general, a CGU is placed into the starting state upon being initiated by a speaker; eventually (if all goes well), the CGU moves into a final state signifying that the CGUs content has entered the common ground. In the meantime, various patterns of continue, repair, acknowledgment, and other grounding acts may occur. Throughout this process, speaker and addressee information is used to determine which role, either initiator or responder, the participants have with respect to each CGU.

[4] In this paper, we are ignoring the evidence of understanding that laughter can convey.
[5] See [14], p. 41.

In this paper, we adapt this model for the purposes of more fine-grained incremental processing. The core of our approach is to allow CGUs to be created and updated incrementally, while an utterance is in progress. These incremental updates can affect both the grounding states and the contents of the CGUs. They can also result in the creation of new CGUs.

To achieve this, we assume a model of incremental speech understanding that delivers a finite sequence of incremental outputs as an utterance progresses, and that each of these outputs estimates both the explicit and predicted content of the utterance at each point in time. (We describe our implementation of such an understanding capability for user speech in Section 4.) Suppose that N outputs are delivered during a spoken utterance. We will denote the sequence of outputs by $\mathcal{O} = \langle (E_1, P_1, C_1), ..., (E_N, P_N, C_N) \rangle$, where E_i is the explicit content and P_i is the predicted content for the i^{th} incremental output. At each point in time, we assume further that the incremental speech understanding model is able to assign a confidence level C_i that describes the reliability of its estimates E_i and P_i. We discuss confidence levels further below.

We begin by describing some of the recognition conditions and the effect of various grounding acts in this incremental grounding model, and then discuss some of the potential realizations of these grounding acts in verbal and non-verbal behavior.

3.1 Grounding acts

Initiate acts generally occur when a speaker begins a new utterance which does not include a continue, request repair, repair or cancel act.[6] Initiate acts create a new open CGU, whose content will be the ungrounded explicit content of the evolving utterance.

Continue acts occur when a new speaker utterance serves as a continuation of an ungrounded CGU that was previously initiated. As a rule, when an interlocutor begins to speak, if there is an open CGU by the same speaker, and that speaker either initiated or recently repaired the CGU, and the utterance does not convey a repair, the utterance is treated as a continue act. For continue acts, each new incremental output E_i is used to update the content of this open CGU.

Acknowledgments generally transition CGUs into the final grounded state, and move the content of the CGUs into the common ground. Of particular interest for incremental grounding is the case of overlapping acknowledgment. If an overlapping acknowledgment is performed by a listener, e.g. a head nod or "Yeah", the CGU being acknowledged will move into the final state, and its content, representing the explicit content of the utterance so far, will become grounded. Generally, the speaker will continue their utterance. On our approach, the continuing speech will be treated as a new initiate act, creating a new CGU to hold the ungrounded content of the continuing utterance. Correct completions are treated as an acknowledgment of the complete utterance, both the explicit

[6] Sometimes, an utterance that includes an acknowledgment will also proceed to initiate a new CGU (as in "okay, so let's talk about the other matter").

and predicted part, since that is the intention of the completing party. The completion makes the predicted content part of the CGU, which so far only contained the explicit content.

Requests for repair may be detected incrementally, as in the case of an overlapping request for repair. These relate to the most recent CGU that was not initiated by the interlocutor making the request, or that was recently repaired by another interlocutor, if such a CGU exists.

Repairs can occur in two cases. The first case is when the utterance was preceded by a request repair addressed to the speaker. The second case is when the speaker repairs a previous utterance by him/herself that conveyed an initiate, continue or repair act. In both cases, content from the CGU is removed, and the explicit content of the new repair utterance is added.

Cancel acts move the relevant CGU into a special canceled state. No special logic is needed to handle this in the incremental grounding model.

3.2 Grounding through Verbal and Non-verbal Behavior

Grounding acts can be realized through a range of verbal and non-verbal behaviors. We now survey a sample of these behaviors, focusing on those that have been addressed in our implementation work to date. An **acknowledging head nod** conveys an acknowledgment grounding act. It is an alternative to a verbal acknowledgment. A **verbal backchannel** (e.g. "okay", "right", "uh-huh") can also be used to perform an acknowledgment act. During a speaker's utterance, a listener whose understanding is progressing adequately may signal continued attention with an **attentive head nod**, inviting the speaker to proceed with their utterance. This is distinguished from an acknowledging nod in that it does not convey an acknowledgment grounding act. A **frown** can be used to realize a request for repair. As discussed in Section 2, **completion** can be used to acknowledge understanding of both explicit and predicted content. An example can be found in Dialogue Excerpt (1), where A completes C's unfinished utterance. This behavior conveys an acknowledge act for the full predicted utterance it is completing. Completions will generally occur when understanding confidence is high, although provisional completions may be used in cases of lower confidence.

4 Implementation

The incremental grounding model is being implemented as part of a virtual human spoken dialogue system, which has been designed to allow trainees to practice their negotiation skills by engaging in face to face negotiation with multiple virtual humans [15]. The specific setting is the SASO4 scenario, which extends the scenario described in [16]. In the SASO4 scenario, two human users play the role of a U.S. Ranger and his deputy, and negotiate with two virtual humans, called Utah and Harmony, to try to convince them that Utah should become the new sheriff of a town.

The system has a fairly typical set of processing components for virtual humans or dialogue systems, including automatic speech recognition (ASR, mapping speech to words), natural language understanding (NLU, mapping from words to semantic frames), dialogue interpretation and management (DM, handling context, dialogue and grounding acts, reference and deciding what content to express), natural language generation (NLG, mapping frames to words), nonverbal generation, and synthesis and realization.

In exploring the incremental grounding model, we build on our existing framework for predictive incremental understanding and confidence estimation [17–19]. This incremental understanding framework captures utterance meanings using a frame representation, where attributes and values represent semantic information that is linked to a domain-specific ontology and task model [20]. Our implementation of the incremental grounding model makes use of several modules for incremental language processing.

Incremental understanding. As a user utterance progresses, every 200 milliseconds, the incremental NLU component produces two semantic frames. The first frame (P_i) is a prediction of the meaning of the *complete user utterance*, which may not have been fully uttered yet; see [17]. The second frame (E_i) is an *explicit subframe* that attempts to capture the explicit meaning of only what the user has said so far; see [19].

Incremental confidence modeling. We also make use of an incremental confidence estimation technique that assigns specific confidence levels (C_i) to the output of the predicted utterance meaning at each point in time. (To date, we do not assign confidence levels to the explicit subframes.) The confidence metrics make qualitative distinctions about the system's level of understanding, and can judge the current understanding level to be *low*, *high*, *incorrect*, and *correct*, among others; see [18].

Collaborative utterance completion. Finally, we use an existing capability for collaborative utterance completion, which allows the virtual humans to complete certain user utterances when the understanding confidence level is adequate and a pause is detected in user speech; see [17].

Building on these models, we have implemented an initial version of the incremental grounding model described in Section 3. The implementation initializes and extends CGUs incrementally, as users are speaking, to maintain an incremental grounding state. To generate grounding behaviors in our virtual humans, we have also designed and implemented an overlapping behavior policy for the virtual humans in SASO4, which we summarize in Figure 1. The behaviors are selected from existing work on feedback models for virtual agents [21, 7], describing various types of nods and facial expressions to signal understanding or confusion.

This policy allows our virtual humans to provide frequent feedback of their level of understanding. For instance, after a short pause in user speech, when there is ungrounded content in a CGU, three kinds of incremental feedback may be provided. If NLU is fully confident that its predicted understanding is *correct*, a verbal backchannel is generated. If the NLU confidence level is *high*

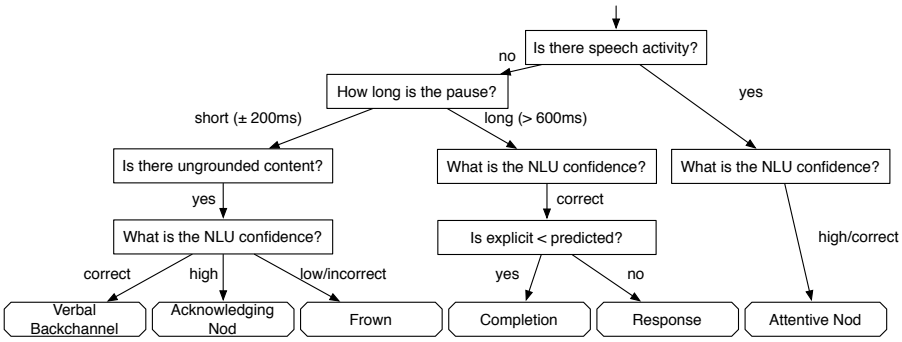

Fig. 1. An overview of the overlapping behavior policy. Every behavior is a leaf in the decision tree, every node a condition testing the user state, common ground or NLU confidence.

(but the NLU is not confident that its understanding is perfectly correct), an acknowledging nod is generated. If the NLU confidence level is *low* or *incorrect*, they generate a frown, signaling a request for repair. Similar rules enable the virtual humans to generate utterance completions or to simply respond to the user's utterance during longer pauses in user speech. A response is chosen in cases when the user's utterance is "finished" in the sense that the explicit content is equal to the predicted content. (In such cases, no completion is necessary.)

While an initial version of this model is implemented, and the virtual humans' responses often seem appropriate, several aspects of the implementation still need to be extended and improved. We also have not yet evaluated the incremental grounding behavior in interactions with users. Exploring the possibility of leveraging more comprehensive and already evaluated behavior policies such as [7] will be part of future work.

References

1. Clark, H.: Using language. Volume 4. Cambridge University Press Cambridge (1996)
2. Tanenhaus, M., Brown-Schmidt, S.: Language processing in the natural world. Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1493) (2008) 1105–1122
3. Oviatt, S., Cohen, P.: Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. Computer Speech & Language **5**(4) (1991) 297–326
4. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: Proceedings of SIGDIAL 2009, London, UK (2009)
5. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. In: Proc. of the 12th Conference of the European Chapter of the ACL. (2009)

6. Morency, L.P., Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems **20** (January 2010) 70–84
7. Wang, Z., Lee, J., Marsella, S.: Towards more comprehensive listening behavior: beyond the bobble head. In: Intelligent Virtual Agents, Springer (2011) 216–227
8. Skantze, G., Schlangen, D.: Incremental dialogue processing in a micro-domain. In: Proceedings of the 12th Conference of the European Association for Computational Linguistics (EACL). (2009)
9. Skantze, G., Hjalmarsson, A.: Towards incremental speech generation in dialogue systems. In: Proceedings of the SIGDIAL 2010 Conference, Tokyo, Japan, Association for Computational Linguistics (September 2010) 1–8
10. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., Morency, L.P.: Virtual Rapport. In Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P., eds.: Intelligent Virtual Agents. Volume 4133. Springer Berlin Heidelberg, Berlin, Heidelberg (2006) 14–27
11. Clark, H., Schaefer, E.: Contributing to discourse. Cognitive science **13**(2) (1989) 259–294
12. Buß, O., Baumann, T., Schlangen, D.: Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics (2010) 233–236
13. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. Language Resources and Evaluation **41**(2) (2007) 181–190
14. Traum, D.R.: A Computational Theory of Grounding in Natural Language Conversation. PhD thesis, University of Rochester, Rochester, NY (1994)
15. Traum, D.R., Marsella, S., Gratch, J., Lee, J., Hartholt, A.: Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In: IVA. (2008) 117–130
16. Plüss, B., DeVault, D., Traum, D.: Toward rapid development of multi-party virtual human negotiation scenarios. Proceedings of SemDial (2011)
17. DeVault, D., Sagae, K., Traum, D.: Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse **2**(1) (2011)
18. DeVault, D., Sagae, K., Traum, D.: Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system. In: The 12th Annual Conference of the International Speech Communication Association (InterSpeech 2011). (2011)
19. Traum, D., DeVault, D., Lee, J., Wang, Z., Marsella, S.: Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In: Intelligent Virtual Agents, Springer (2012)
20. Hartholt, A., Russ, T., Traum, D., Hovy, E., Robinson, S.: A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture. In: Language Resources and Evaluation Conference (LREC). (May 2008)
21. Huang, L., Morency, L., Gratch, J.: Virtual rapport 2.0. In: Intelligent Virtual Agents, Springer (2011) 68–79