Synchronizing Dialogue Contributions of Human Users and Virtual Characters in a Virtual Reality Environment

Norbert Pfleger, Markus Löckelt

German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
{pfleger,loeckelt}@dfki.de

Abstract

Synchronizing user and system actions in a real-time virtual reality environment is a challenging task. Key components of a dialogue system like speech recognition, discourse processing, speech generation and synthesis all contribute significant delays to the response time. With human interlocutors, however, a continuous flow of conversation is important as any implausible gap may cause confusion with respect to floor management. In this paper, we describe some cases of how to make sensible use of information available at each processing step to be able to give early useful feedback even while processing.

1. Introduction

We present some practical solutions with respect to real-time synchronization issues in multi-party dialogues. The solutions are being implemented in the multimodal dialogue system VirtualHuman[1] , which is currently under development with a first working prototype. In VirtualHuman, dialogue participants can be human users or virtual characters pursuing individual task goals in a 3D scene. A human user can interact with the system using speech and gestural input; the actions of the virtual characters comprise synthesized speech, gestures, and facial expressions. An overall description of the component architecture can be found in [1]. We will use the following dialogue fragment as an illustration for the rest of the paper. The dialogue is an excerpt from a football quiz with three participants: Two virtual characters (the moderator and a contestant called Frank), as well as one human contestant ("User"):

(1) *Moderator*: [looks at both candidates] *"The next question: Who scored the last goal at the world championship in 1990?"*

(2) *User*: [looks at moderator] *"Franz Beckenbauer"*

(3) *Moderator*: [looks at User; Frank shakes his head and raises his index-finger] *"well, . . . "* [looks at Frank] *"yes, Frank?"*

(4) *Frank*: [looks at moderator] *"He was the coach . . . "* [Moderator nods] *"The correct answer is Andreas Brehme."*

(5) *Moderator*: [looks at Frank] *"Yes, that will be one point"* [points at Frank] *"for Frank!"*

The issues we address here are the following: (a) The initial question in (1) is directed at two addressees, Frank and the human user. They both compete for the response turn. To avoid confusion, overlapping speech should be prevented in this situation. This requires immediate feedback as soon as one participant attempts to take the turn for a response. (b) It must be distinguished whether an utterance during or in response to (1) is an actual answer to the question, or just (backchannel) feedback signaling that it was understood. (c) When dialogue moves impose constraints on the range of expected follow-ups, such as a question demanding an answer on the same subject, the recognition and analysis components can be adapted to aid the recognition rate. However, a flexible interaction requires that unexpected utterances should also be recognized. (d) As for floor management, generating utterances for virtual characters takes time, and so does analyzing a user utterance. To avoid barge-in during an utterance, as well as interruptions during the generation phase (*"barge-before"*), the system can (i) produce cues for the user that an utterance is being analyzed or produced, e. g. insert filler phrases (e. g. filled pauses or interjections), and (ii) adapt the behavior of the virtual characters to reflect that they are about to say something.

We begin by describing issues related to recognizing and processing user input and then turn to the generation of interactional behavior by the virtual characters. We conclude with some remarks how our approach is integrated in our system, and how we plan to improve it in future work.

2. Processing User Input

A key aspect of human-computer interaction is related to the synchronization of the actions of the user and the system. Usually, it takes some time until a spoken utterance of the user is recognized, analyzed and interpreted before an appropriate system reaction can be triggered. However, the delays that emerge from this sequential processing chain cause serious discontinuities that ruin the effect of an interactive conversation. Even through incremental approaches to recognition and generation, this limitation cannot be fully resolved as the actual system reaction cannot be rendered before the final results of the incremental processes are available. For convincing interactions with embodied virtual characters, the time span between the user utterance and a first system reaction has to be much shorter. To achieve this, we decided to better exploit the available information at each processing step and tweak the communication between those modules that are involved in the recognition and analysis part of our dialogue system.

We argue that two additional interfaces between processing components of a dialogue system plus additional contextual information can provide key cues for a reactive processing of

[1] http://www.virtual-human.org

user contributions during a conversational dialogue. Besides the well-known contextual factors like expectations about the next user contribution we use the notion of a conversational context that provides some additional information thereby improving the recognition and processing of user contributions. In what follows, we focus on that contextual information that supports automatic speech recognition (ASR). However, many of the following observations would also hold for gesture recognition.

2.1. Contextual Information Supporting ASR

Perhaps the most obvious and often applied contextual information is the use of context dependent language models (LMs). This means that after each system contribution the ASR is updated with a language model that predominantly covers the most likely and thus expected user contributions. This increases the likelihood of correctly recognizing anticipated user input but at the same time hampers the ability to recognize unexpected input. We use this technique and plan to further refine it by introducing different degrees of expectedness, as will be shortly outlined in the discussion section.

To avoid the limitation introduced by specialized LMs, one can also use the expectations to enhance the processing and scoring of recognition hypotheses during the analysis of already recognized speech input . For example, the natural language component [2] of the multimodal dialogue system SmartKom employs fine-grained expectations about expected and possible user contributions to score and rank the lattice of different interpretation hypotheses. [3] describes how virtually all analysis components of such a system benefit from the combination of these approaches.

2.2. Reacting to User Contributions

The decision whether a contribution is meant to take the turn or whether it simply provides backchannel feedback [4] calls for rapid decisions that need to be drawn on a moment by moment basis. An approach often taken in telephone based VoiceXML systems is to interpret any user input during a system contribution as a turn request and to stop the ongoing utterance at that very moment. Even though this might be a suitable approach in the context of a system-driven, dyadic interaction it is not applicable for multi-party conversational dialogue systems where both the users and the virtual characters are treated as equal interlocutors, and the initiative can change dynamically.

Turns (1) and (2) exemplify the need for quick reactions: In turn (1) the moderator directs his question to both attendees thereby initiating a competition about who is getting to be the next speaker. Usually, situations like these are resolved on a simple first come, first served basis. The participant starting to speak first gets the turn. Thus, if one participant starts to speak first (in this case the user) the other participant needs to back off and wait until he gets the turn.

This kind of floor competition requires a permanent monitoring of the other participants' actions. This means that the individual actions of the other participants need to be interpreted with respect to their impact on the own, planned actions. Whereas this is a task that can be solved for virtual characters within our architecture of autonomous agents representing the individual characters (see section 4.1), it is different when human users come into play. What we would need is a recognition engine that supports real-time processing of user contributions. If the user started to talk before the virtual character Frank started, the system should stop the output of any planned utterance. Unfortunately, we are not aware of any currently publicly available ASR engine that supports an open microphone recognition and that also provides recognition results in near real-time. Because of this, we developed an approach that enables the interpretation of user behavior even before the final recognition result is available.

What we do is that we break what is normally viewed as a monolithic dialogue system up into two distinct types of conversational dialogue engines (CDEs): (i) *user CDEs* covering the input and analysis components and (ii) *character CDEs* covering the dialogue manager and output components (see Fig. 1). The components communicate by means of an abstracted, ontology-based representation of the world in which the virtual characters are displayed (see section 4.1 for more details). We integrated the silence detection of the ASR into the subsequent processing chain. When the silence detection detects some input event, the virtual characters are informed that a user just started to speak: a "turn-taking" or "backchannel" act is sent to all overhearing characters, which will in turn decide on the basis of their current conversational role what will be an appropriate next move. During turn (2) of our example dialogue, the previous speaker (the moderator) responds by looking at the new speaker and the other virtual character (Frank), who was also addressed by the previous question, and immediately cuts off the ongoing preparation of his answer.

2.3. Interpreting User Contributions

As mentioned in the previous section, there is other contextual information than expectations that supports the processing of user contributions. In general, conversations are not just a continuous flow of speech, but also comprise silent or filled pauses within and in between turns as well as nonverbal behavior. Especially nonverbal actions of the characters can be used to reveal a lot about their intentions, interests, feelings and ideas. Besides augmenting the propositional content of utterances, nonverbal behavior can contribute to the interactional organization of the discourse, e. g., by gazing. However, to understand the function of these actions, the context of use needs to be considered (see, e. g., [5, 6]).

Thus, when we aim to process the full variety of possible conversational behavior, we need the immediate as well as the wider context of the utterance in order to draw valid conclusions. Pauses between two subsequent turns are often used for strategic purpose, e. g., for taking or yielding the turn. Such pauses can mark *transition relevance places* (TRPs) where a turn exchange is possible and the hearer may take over the floor [7]. However, a pause marks a TRP only if the nonverbal behavior (gaze and gestures) displayed by the speaker at the same time suggests that he is willing to yield the turn (see Fig. 2).

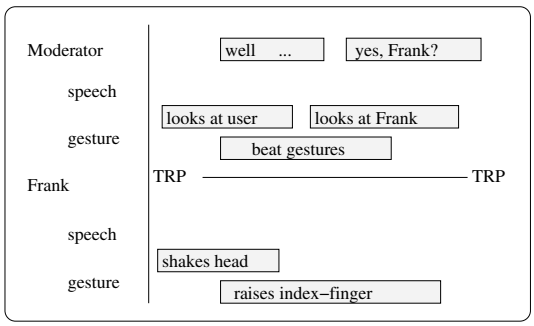

Figure 2: TRPs in relation to the actions of the characters

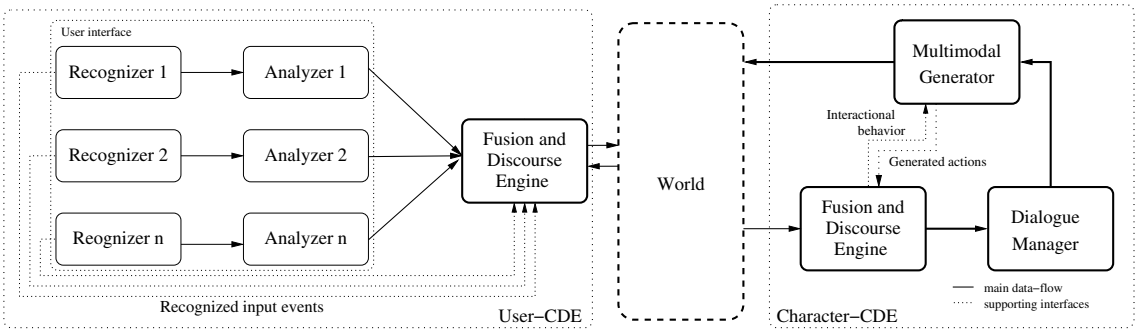

Figure 1: Abstracted architectures of user and character CDEs

Our approach is based on the notion of an immediate *conversational context*. By conversational context we mean those contextual aspects that are defined by the current state of the conversation, e. g., the current role of each participant in the conversation—*speaker, hearer, overhearer*—and the current nonverbal behavior displayed by the individual participants—*who is gazing at whom, pointing gestures, metaphoric and emblematic gestures*—but also facial expressions [8].

These cues provide valuable information for processing recognized input, so that we are able to identify whether a contribution is meant as a claim for the speaking turn, or just to provide backchannel feedback [9]: If, for example, the user just started to speak, this decision depends on whether the previous speaker has already finished his turn and displayed turn-yielding cues, or whether he is still holding the turn. The former situation clearly suggests to interpret the just started spoken utterance as a new turn so that the other participants should remain silent until this interpretation is disproved. Conversely, the latter situation suggests to rather interpret the utterance as backchannel feedback, meaning that the current speaker should continue. However, if the subsequent processing of the recognized speech reveals that it marks a new turn, the speaker should either yield the turn or display *attempt-suppressing signals* [10]. The turn-management policy is implemented by means of a production rule system that is part of the fusion and discourse engine (FADE).

3. Generating Interactional Behavior

3.1. Feedback and Pauses from Virtual Characters

A smooth exchange of turns requires signaling clearly when a character attempts to take the turn, so that the human user and the other participants will not interrupt while the utterance is planned and generated (see section 2.2).

Silent and filled pauses are often used between human interlocutors for these strategic purposes, e. g., taking, holding and yielding the turn [5]. Consider, for example, turn (3) of our example dialog where the moderator takes the turn in order to evaluate the answer of the user. But since he is still busy with planning the utterance he uses *"well, . . . "* at the beginning of the turn to underline his claim for the turn.[2] But in addition to this verbal marker, he displays a nonverbal turn-taking cue by looking at the previous speaker.

When generating longer utterances, the generation component exhibits a similar behavior. The kind of verbal filler used for the pauses depends on the dialogue act type to be generated; additionally, the dialogue manager can provide explicit cues about the semantic content, e. g. whether an answer to a question is positive or negative. When the generator determines that it needs to fill a pause, it draws from a collection of set pieces to be inserted. This can include general utterances (like "hmm . . . ") as well as more specific utterances depending on the discourse context ("one moment please . . . ", "what I want to know is . . . "). We differentiate two classes of interactional contribution that can be generated by the virtual characters:

1. *Direct nonverbal feedback* is triggered via FADE and comprises actions like gazing behavior (if another participant begins to speak) or canceling ongoing speech output (if a character's attempt to take the turn is suppressed by the previous speaker).
2. *Filled pauses, interjections and short phrases* are triggered via the dialog manager. These units are used to fill the gap during which the actual utterance is generated and synthesized. They may already reveal some aspects of the subsequent utterance in terms of a general attitude towards the previous statement.

Both classes have in common that their effect depends on a precise timing. If, for example, the delay between the end of the turn of the previous speaker and the beginning of the new turn is too long, another participant might step in and take the turn. We adopted this interactional behavior via the aforementioned additional connection between the recognition components and FADE (in user-CDEs) and an additional connection between FADE and the multimodal generation component (in character-CDEs). If the user, represented by an user-CDE, attempts to take the floor, the internal representation of the world gets updated about this fact. All other CDEs are able to recognize this fact immediately and will take it into account for their next actions. A character-CDE that attempts to take the speaking turn will update the internal representation of the world accordingly; this will take effect when the contribution is rendered.

3.2. Game Recognition

Each dialogue agent holds a set of dialogue games defining the sequences of dialogue acts it can recognize and participate in. A game active between interlocutors specifies—among other

[2] The example cited here is somewhat more complicated: the moderator does not get to produce the actual utterance, since the other participant claims and gets the turn while the moderator is still planning.

things—rules that state which dialogue acts are "acceptable" (or "expected") at each stage in the game (e. g. a question may be expected to be followed by an answer or a refusal to answer), and constraints about the beliefs of participants, (e. g. an interlocutor is usually obliged to believe the propositional content of a statement he just made). We will not describe dialogue games here (see e. g. [11] for a general discussion or [12] for some formal aspects), but just use the notion that they can be used to predict future utterances.

To determine which game an utterance occurs in, the move recognition component can assume that it is likely that an active game is continued with a move that is legal in the current game situation, or it can match dialog moves against game-initial moves, to start or embed a new game. When matching, game constraints can be used to narrow down the choice based on context, e. g. if a question is posed, and the addressee knows that the questioner already knows the answer, he may infer that it is a rhetorical question, and therefore, no reply is needed – unless the question occurs in an "examination" context.

Both user and character CDEs use information about which dialogue game is currently active. User CDEs need to convert user utterances into dialogue acts and will prefer such acts that are legal game moves (e. g. a *response* following a *question* when an *infoSeek* game is currently active). Likewise, the deliberative unit of character CDEs will try to produce legal moves. Thus, types and constraints on the legal followup move types in active dialogue games provide expectation information to aid the analysis modules.

4. Conclusion

4.1. System Integration

Our implemented system consists of concurrent modules and employs a blackboard communication architecture. Each dialogue participant is represented by an independent conversational dialogue engine (CDE); *user CDEs* for human interlocutors and *character CDEs* for virtual characters.

A CDE consists of several sub-components depending on its type (see Fig. 1). Each user CDE contains input recognition components (speech and gesture) and transforms the input to an ontological representation in terms of dialogue acts, which is used in intra-CDE communication. A Fusion and Discourse Engine (FADE) implements multimodal fusion, maintains the dialogue history, and triggers reflexive behavior such as gazing and gestures (for character CDEs). A dialogue manager recognizes the roles of contributions in terms of moves in dialog games and uses it to produce the expectations (see Section 3.2). The dialogue manager component of a character CDE implements the deliberate behavior to achieve the goals of the character. It uses the predictions to trigger dialogue contributions that are legal moves in the current game. Character CDEs also include a multimodal generation component that receives the semantical representation of the dialogue contributions of a character and transforms them into multimodal presentation instructions (spoken utterances, gestures, and facial expressions) for a multimodal player that renders the scene for the human user.

4.2. Future Work

Context Based Language Models: To improve the context based selection of language models, a first step would be to use a general purpose language model as a fallback when the contextually selected model fails. A more complex solution would comprise series of LMs that are ordered based on either dialogue history recency or semantic relatedness to the current task. This also seems to bear some psychological plausibility. The feasibility of this approach is dependent on the structure of the interaction: Task-oriented dialogues are more adequate than relatively free conversational dialogues with less structure.

Evaluation: Although it seems intuitive that the additional feedback provided by the characters improves the general fluency of interactions between human users and virtual characters, we need empirical data to quantify the effect. Therefore, we plan to evaluate the system's performance by a user study comparing the system as described in this paper with a baseline version stripped of the synchronization functionality.

5. Acknowledgments

This research is funded by the German Ministry of Research and Technology (BMBF) under grant 01 IMB 01A. The responsibility lies with the authors.

6. References

[1] M. Löckelt and N. Pfleger, "Multi-Party Interaction With Self-Contained Virtual Characters," in *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue (DIALOR)*, Nancy, France, 2005, pp. 139 – 143.

[2] R. Engel, "SPIN: Language understanding for spoken dialogue systems using a production system approach," in *Proceedings of 7th International Conference on Spoken Language Processing (ICSLP-2002)*, Denver, USA, 2002, pp. 2717–2720.

[3] W. Wahlster, "Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression," in *Proceedings of the 26th German Conference on Artificial Intelligence*, A. Günter, R. Kruse, and B. Neumann, Eds. Berlin, Heidelberg: Springer, September 2003, pp. 1 – 18.

[4] V. H. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting*. Chicago Linguistics Society, 1970, pp. 567–577.

[5] A.-B. Senström, *An Introduction to Spoken Interaction*. Longman Group UK, 1994.

[6] A. Kendon, *Gesture: Visible Action as Utterance*. The Press Syndicate of the University of Cambridge, 2004.

[7] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversations," *Language*, vol. 50, no. 4, pp. 696 – 734, 1974.

[8] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction*. Wadsworth Publishing - ITP, 2002.

[9] N. Pfleger and J. Alexandersson, "Modeling non-verbal behavior in multimodal conversational systems," *W. Wahlster (ed.): Special Journal Issue "Conversational User Interfaces" it - Information Technology*, vol. 46, no. 6, pp. 342–245, 2004.

[10] S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[11] J. Hulstijn, "Dialogue games are recipes for joint action," in *Proceedings of Gotalog Workshop on the Semantics and Pragmatics of Dialogues*, Gothenburg, Sweden, 2000.

[12] P. McBurney and S. Parsons, "Dialogue games in multi-agent systems," *Informal Logic*, vol. Special Issue on Applications of Argumentation in Computer Science, 2002.