Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation

David Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella

Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Playa Vista, CA 90094, USA

Abstract. In order to provide comprehensive listening behavior, virtual humans engaged in dialogue need to incrementally listen, interpret, understand, and react to what someone is saying, in real time, as they are saying it. In this paper, we describe an implemented system for engaging in multiparty dialogue, including incremental understanding and a range of feedback. We present an FML message extension for feedback in multipary dialogue that can be connected to a feedback realizer. We also describe how the important aspects of that message are calculated by different modules involved in partial input processing as a speaker is talking in a multiparty dialogue.

1 Introduction

In order to be human-like in their behavior, intelligent conversational agents need to be able to produce a range of feedback to a speaker during a conversation. Human feedback behavior has a number of features, which translate to a set of requirements for satisfactory virtual agent feedback. Some of these requirements are:

- Human feedback is provided in *real-time*, as the speaker is articulating (or having trouble articulating) her utterance. This means that the feedback mechanism can not wait until after the speaker has finished to calculate the feedback.
- Human feedback is also often *specific* [4], so the feedback mechanism requires interpretation and attempted understanding of what the speaker is saying.
- Taken together, these requirements lead to a third one, that understanding be *incremental* - operating on parts of the evolving utterance and computed in real-time before the utterance has been completed.
- Human feedback is also *expressive* [1], indicating aspects of the current mental state of the feedback giver, including beliefs, goals, emotions, and how the developing utterance is related to these. This means the feedback mechanism must have access to the cognitive aspects and be able to do pragmatic reasoning, including reference resolution, to relate the utterance meaning to the agent's mental state.
- Human feedback is sometimes *evocative* [1], trying to create an impression on or response behavior from the observers of the feedback. This includes intended effects on the main speaker, to regulate the content, timing, and amount of detail of what she is saying, as well as intended effects on other observers, such as adoption of beliefs about the feedback giver, or whether they should take a turn next. Evocative feedback means that the feedback mechanism must have access to the communicative goals, plans and intentions of the agents.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 275–288, 2012.

A model of feedback generation that contains each of these features was presented in [35]. In this paper we describe an FML message specification that supports this behavior generation model, as well as an implemented agent dialogue model that can provide the aspects of semantic and pragmatic understanding for specific expressive and evocative feedback in realtime, This model is similar in many respects to that proposed by [24], however that model worked on typed dyadic dialogue rather than spoken input for multiparty conversation, and focused on expressive feedback, while the model below also includes evocative feedback and participation status of participants.

2 Architecture and Message Specification

Our architecture requires the following components to produce incremental feedback:

- A speech recognizer that can produce incremental, word-by-word results, ideally with confidence scores.
- A natural language understanding component (NLU) that produces semantic representations and predictions of final meaning when given a speech recognition output.
- A meta-NLU component that computes confidence estimates given (partial) speech recognition and NLU outputs.
- A vision component that can recognize speaker behaviors such as gaze direction.
- A domain reasoning component that can model beliefs, tasks, plans, and attitudes toward particular topics.
- A dialogue manager that can compute pragmatic effects of communication as recognized by the above input components and update state and calculate communicative intentions.
- A feedback generator that can produce communicative behaviors given the function specifications from the dialogue manager.

In practice, some of these components can be combined in a single software module. Our architecture combines the NLU and meta-NLU components into a single module, and the dialogue manager and domain reasoning is another single module.

In order to pass the needed information from dialogue manager to the feedback generator, we created the XML backchannel feedback message specification in Figure 1. This message type is meant to be part of the SAIBA framework [23], with most of it being FML content [18]. The message also contains aspects that have been developed by earlier processing components. There is one `participant` element for each participant in the conversation. Participant roles are discussed in Section 8. The `conversation-goal` element contains goals related to maintaining and changing participant status. They are discussed in Section 9. The `dialogue-act` element represents the feedback itself – it can be given either as a backchannel (type=listening) or verbally within a turn (type=speaking), although we have not fully implemented the speaking type as of this writing. The `feedback` element contains information about the utterance that is being spoken, and what the agent thinks about it. Attributes of this element and the `partial-text` element are derived by the speech recognizer, and described in Section 4. The `partial-sem` element contains information from the NLU module, and is described in section 5. This information is augmented by contextual information

```
<act>
 <participant id="[character]" role="[role]"/> (minimum 2)
 <fml>
   <conversation-goal>
     <participation-goal goal="[boolean]"/>
     <comprehension-goal goal="[boolean]"/></conversation-goal>
   <dialog-act type="[listening/speaking]">
     <feedback agent="[character]" speaker="[character]"
          utterance="[id]" progress="[integer]" complete="[boolean]">
        <partial-text>[string]</partial-text>
        <partial-sem confidence="[real]">
           <indicators Correct="[boolean]" High="[boolean]"
                Incorrect="[boolean]" Low="[boolean]" MAXF="[boolean]"
                PF1="[boolean]" PF2="[boolean]" PF3="[boolean]"
                WillBeCorrect="[boolean]" WillBeHigh="[boolean]"
                WillBeIncorrect="[boolean]" WillBeLow="[boolean]"/>
           <predicted_nlu><object name="[id]">
                        ... </predicted_nlu>
           <explicit_subframe><object name="[id]">
                        ...  </explicit_subframe></partial-sem>
        <attitude type="[like/dislike]" target="[id]" stance="[leaked/intended]"
                  intensity="real"/>
        <affect type="[emotion]" target="[id]" stance="[leaked/intended]"
                  intensity="real"/>
      </feedback></dialog-act></fml></act>
```

Fig. 1. Feedback Behavior Generation Message Specification

provided by the dialogue manager, as described in Section 6. Finally, the `attitude` and `affect` elements come from the domain model expected utility calculations and emotions, once the dialogue manager has identified the relevant concepts that are being spoken about. The domain model is discussed in Section 7. Finally, in Section 10, we briefly review the feedback behavior generation component that takes as input messages of the form of Figure 1 (more details are provided in [35]).

The specification and components are domain independent, and have been tested in a few different domains. However, to provide more concreteness in examples we present one domain, *SASO4*, described in the next section. Figures 2 and 3, show a visualization of some of the information from the Feedback message, in a graphical form. These figures show a couple of snapshots 2.0 and 4.6 seconds in the progress of a single 7.4 second utterance in the SASO4 domain.

3 Example: The SASO4 Domain

As our development testbed, we situated this work in the SASO4 domain, which extends the scenario described by [29]: *An American Old West town has been freed from a dangerous outlaw, defeated by a U.S. Ranger with the help of Utah, the local bartender. The Ranger and his Deputy must now leave town to pursue their mission elsewhere. But before leaving, they need to recruit a town sheriff, so they offer the job to Utah. He will need resources – e.g., money to buy guns and to hire men – guaranteed before considering the offer. As owner of the saloon, Harmony is an influential woman in town. She will be present in the discussions, pushing forward her own agenda of demands, part of which she cannot discuss in front of Utah and must be dealt with in private by*

Fig. 2. Visualization of Incremental Speech Processing after 2 seconds

one of the officers. The Ranger and the Deputy have very limited resources, so they must negotiate to reach an agreement by committing as little as possible.

This scenario has many opportunities for feedback, both as the scenario progresses, and within the interpretation of a single utterance. It includes four characters, two played by humans, and two by virtual agents. The scenario starts with no conversation, but the humans could start conversations with one or both agents. It contains situations in which the agent Harmony desires to leave the conversation, and an opportunity for re-entry if she leaves. The agents also have shifting points of view about some of the things discussed as the conversation progresses, e.g. whether Utah will be the Sheriff.

4 Speech Recognition

Our automatic speech recognition (ASR) module is currently PocketSphinx [19]. The ASR is configured with a statistical language model trained on the transcripts in a corpus of user utterances and paraphrases. In the SASO4 scenario, we currently use approximately 1,500 transcripts to train the language model. To enable incremental understanding and feedback based on partial ASR results, after each 200 milliseconds of additional speech from an ongoing user utterance is captured, it is provided to the ASR. The ASR module sets the *utterance, speaker, progress*, and *complete* attributes of the feedback element in the partial message in Figure 1. The *utterance* attribute is a unique id for this session, the *progress* attribute contains the ordinal count of partial interpretations of this utterance. The *complete* attribute signals whether the speaker has stopped speaking. The partial ASR result appears in the `partial-text` element in the feedback message. The partial-text is shown in white in the visualization in Figures 2 and 3.

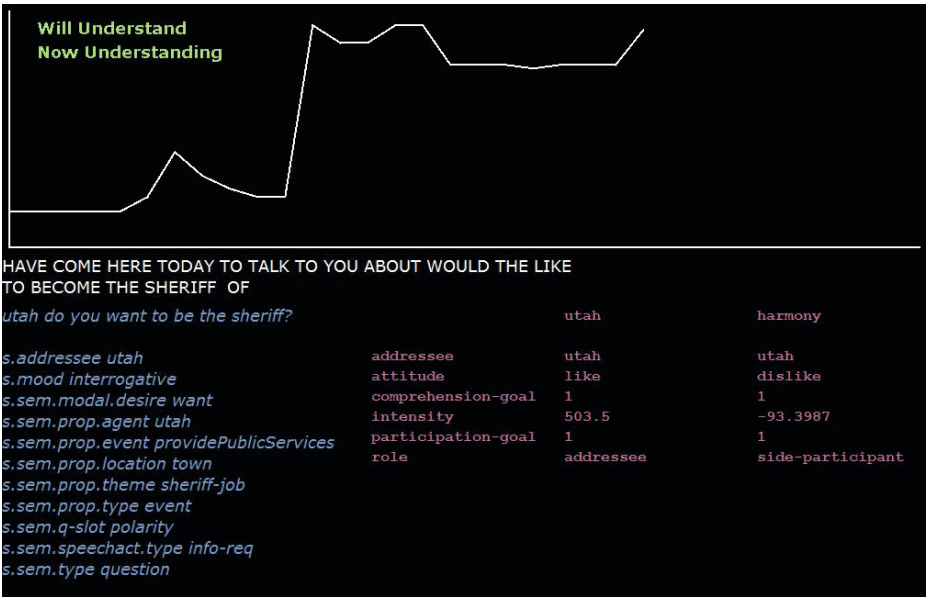

Fig. 3. Visualization of Incremental Speech Processing after 4.6 seconds.

5 Semantic Interpretation

We adopt a detailed framework for incremental understanding and confidence estimation that has been developed in [30,31,8,7]. The key components for listener feedback behavior are the semantic frames and confidence indicators that are produced for each partial ASR result. 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 [16].

As a user utterance progresses, the incremental NLU component produces two semantic frames. The first frame is a prediction of the meaning of the *complete user utterance* (which may not have been fully uttered yet). This prediction is made using a statistically trained maximum entropy classifier [8]. The `<predicted_nlu>` element in the feedback message specification in Figure 1 provides this predicted frame. Examples of such predictions are also shown in blue in Figures 2 and 3, below the ASR partial result, along with a gloss of the meaning of the frame. For Figure 2 the prediction is that this utterance will be a greeting to harmony. In Figure 3, the prediction has changed to a more detailed frame in which the user is asking Utah if he wants to become the sheriff.

The second type of frame produced by the incremental NLU components is an *explicit subframe* that attempts to capture the explicit meaning of only what the user has said so far, without predicting the complete meaning of the user's full utterance. The identification of this subframe can be performed using related statistical classification techniques [17], and the resulting subframe is given in the `<explicit_subframe>` element in the feedback message specification. (This component can optionally be shown in the visualization, but it is not shown in Figures 2 and 3.)

A third component in this incremental processing framework is a set of boolean-valued confidence indicators that can be used to assess, in intuitive terms, the reliability of the predicted frame for the user utterance [7]. The indicators encompass a range of potentially valuable information about how well an utterance is being understood so far, and how much that understanding may improve as the user continues speaking.

All the indicators are ultimately defined in relation to an F-score metric which can generally be used to assess NLU performance. The F-score calculation looks at precision and recall of the attribute-value pairs (or frame elements) that compose the predicted and correct (hand-annotated) frames for each partial ASR result. Precision represents the portion of frame elements in the predicted frame that were correct, and recall represents the portion of frame elements in the gold-standard annotations that were predicted.

Table 1. Metrics for incremental speech understanding

Metric	**Definition**	**Metric**	**Definition**	**Metric**	**Definition**
High_t:	$F_t \geq \frac{1}{2}$	WillBeHigh_t:	$F_L \geq \frac{1}{2}$	PF1_t:	$\text{Correct}_t \vee (\text{Incorrect}_t \wedge \text{WillBeCorrect}_t)$
Correct_t:	$F_t = 1$	WillBeCorrect_t:	$F_L = 1$	PF2_t:	$\text{High}_t \vee (\text{Low}_t \wedge \text{WillBeHigh}_t)$
Incorrect_t:	$F_t < 1$	WillBeIncorrect_t:	$F_L < 1$	PF3_t:	$\text{High}_t \vee (\text{Low}_t \wedge \neg\text{MAXF}_t)$
Low_t:	$F_t < \frac{1}{2}$	WillBeLow_t:	$F_L < \frac{1}{2}$		
		MAXF_t:	$F_t \geq F_L$		

Currently we have been using a set of indicators that are defined in Table 1. In this table, F_t is the F-score of the predicted frame at time $t \in 1...L$, for an utterance that contains L 200 millisecond chunks of audio. F_L is the F-score of the final predicted frame for the complete user utterance. We also are using a set of more complex indicators that may indicate appropriate moments for virtual humans to provide positive feedback. These are defined in the right column of Table 1.

All these indicators are included in the `<indicators>` element in the feedback message specification in Figure 1. In Figure 3, the indicators High ("Now Understanding") and WillBeHigh ("Will Understand") are shown in green at the top left (both true at this point). In Figure 2, only WillBeHigh is true, as the system has low confidence about the current guess. The progression of expected F-score (given in the `confidence` attribute of the `partial-sem` element in the message spec in Figure 1) is shown by the white line at the top of the figures - low for Figure 2 and high for Figure 3.

6 Contextual Pragmatics

The semantic representation provided by the NLU component represents the context-free interpretation of the meaning of the utterance. The next step in processing is interpreting this within the current context to provide a pragmatic meaning, along the lines in [32]. For every partial utterance returned by the NLU component, the following are computed independently by each agent:

- updates on the participant structure
- a set of zero or more dialogue acts that have been performed by the utterance

- resolution of named entities to concepts
- resolution of action and state descriptions in the semantic interpretation to relevant states and tasks in the agents's task model (see Section 7).

We describe each of these briefly (except for addressee and participant structure, described in Section 8). First, primitive concepts that are part of the semantic representation are resolved. Primitive concepts include people (e.g. the participants in the dialogue), objects, locations, action types, attributes, and values. For named concepts, there is a simple look-up table (in most cases, the identity mapping), which allows each agent to have a different internal representation from other agents (if desired), and allows multiple semantic terms to refer to the same internal, domain-specific concept. Slightly more complex is the resolution of typed referring expressions, including indexicals ("I", "we", "you", "here", "there"), anaphors ("he", "she", "it", "this", "that"), and noun phrases that do not uniquely identify the concept for the domain (e.g. "the money"). In this case, the process for reference resolution involves looking up the features of the referring expression (e.g., animacy, gender, location, type) and then finding a list of "candidate" concepts that have these features. Then, if possible, disambiguation is performed by preferring those that have been mentioned most recently. If a single best candidate can not be found, this is motivation for an agent to perform a grounding move, giving feedback of sub-optimal understanding. This might take the form of a clarification request asking which of the possible candidates is meant, or a verification request about one of the candidates, or an open-ended question asking for the disambiguation (though the agent might wait if the agent thinks that they will understand later).

There is also a resolution of complex elements such as actions and states. The agents have a representation of a set of relevant states, whose valence are checked whenever new information comes in, including perceptual information, inference from other internal information, or verbal reports from trusted others. These states are represented as triples of *object, attribute, value*, where each of these are more primitive concepts. A given utterance might uniquely refer to an internal state, but might also not match any state or might match more than one (e.g. if one of the three elements is missing or is an under-constrained referring expression). Likewise, actions are represented, with a set of thematic roles, including *agent, theme, location, destination*, etc. A match is made from the consistency of the semantic frame with the task model frame to get a set of action candidates. Then, depending on the tense and modality, additional matches may be made with action instances in the plan or in the causal history of previous events. Recognized states and actions also provide an additional constraint on concept identification – one candidate concept is preferred over another if it leads to a possible match with a state or action and the other does not.

Finally, a set of dialogue acts are identified that are being performed by the speaker in producing the utterance. These include core speech acts, such as assertions, questions, offers, backward acts, such as answers, acceptances, grounding acts, and other acts that influence the information state of the dialogue. In the version of the semantic frames in the partial-sem attribute of Figure 1, the frames are augmented with reference information, though that is not provided in the visualization in Figures 2 and 3.

7 Domain Reasoner

Once the referred to objects and speech acts have been computed, it is possible to relate the speaker's projected sentiment toward the referenced object with the listener's feelings about the topic. The ability of our agents to interact with humans and other agents is based in their understanding of the goals of each party, the actions that can achieve or thwart those goals, and the commitments and preferences agents have towards competing courses of action. To provide this understanding, our agents use an explicit representation of an agent's current mental state concerning past, present, and future states and actions, their likelihood and desirability, and causal relationships between them. This representation is grounded in a planning representations that has been extended to incorporate representations of decision-theoretic reasoning (i.e, probabilities and utilities), representations to support reasoning about beliefs and intentions and a causal history that expresses the relations between past events to the agent's current beliefs, goals and intentions. We call this representation a causal interpretation.

The agent's valence reactions to its comprehension, including the attitude and affect elements depicted in Figure 1, also rely on this causal interpretation. Specifically, the valence reactions are based on a general computational framework for modeling emotion processes, EMA (Emotion and Adaptation) [14,27]. EMA is based on appraisal theories of emotion that argue that emotion arises from a process of interpreting a person's relationship with their environment; this interpretation can be characterized in terms of a set of criteria (variously called appraisal dimensions, appraisal variables or appraisal checks); and specific emotions are associated with certain configurations of these criteria. To represent the agent's relation to its environment, EMA relies on the agent's decision-theoretic plan based representation. The plan represents a snapshot of the agent's current view of the agent-environment relationship, including its beliefs, desires and intentions. This representation changes moment-to-moment in response to internal and external changes. EMA's appraisal of these changes uses fast feature detectors that map features of the plan into appraisal variables. Appraisals thus provide a continuously updated affective summary of its contents. This is particularly relevant to model the valenced reactions to the dynamically evolving comprehension of partial utterances. The appraisal process in EMA maintains a continuously updated set of appraisal values associated with each proposition in the causal interpretation. A partial list of the variables most relevant to the current discussion include:

Desirability: This characterizes the value of the proposition to the agent (e.g., does it causally advance or inhibit a state of utility for the agent). Desirability has both magnitude and sign – it can be positive or negative. This includes propositions that have intrinsic utility for the agent but also propositions that have extrinsic utility by virtue of causally impacting a proposition that has utility.

Likelihood: This is a measure of the likelihood of propositions.

Causal attribution: who deserves credit/blame.

Controllability: can the outcome be altered by actions under control of the agent.

Changeability: can the outcome be altered by some other causal agent.

Each appraised event is mapped into an emotion instance of some type, such as hope or fear, with some intensity, based on the pattern of appraisals. The intensity is calculated in the form of expected utility based on desirability and utility.

EMA also includes a computational model of coping integrated with the appraisal process. Coping determines, moment-to-moment, how the agent responds to the appraised significance of events. Within EMA, coping strategies are proposed to maintain desirable or overturn undesirable events. As opposed to the more reactive nature of appraisal, coping strategies can be seen as more deliberative attempts to enable or suppress the cognitive processes that operate on the causal interpretation.

With this background, we can characterize how the `affect` and `attitude` elements of the message spec in Figure 1 are calculated. The attitude's type attribute is based on the desirability of a referenced task action. The intensity attribute is derived from the calculation of that action's expected utility. The stance attribute distinguishes between expressive feedback from the appraised desirability (termed "leaked" in the attitude element specification), vs. evocative feedback meant to intentionally realize coping strategies, by conveying a specific affect, which may not be what is really felt (termed "intended"). This latter intentional expression of attitudes is not fully implemented yet in the incremental feedback. In Figure 2, there is no attitude shown, since there is no task model element referred to (yet) given the prediction of a greeting act. In Figure 3, we can see that Utah likes the idea of becoming Sheriff with intensity over 500 while Harmony dislikes the idea with an intensity of about -93.

The `affect` element is tied more directly to the results of the appraisal and coping process. In contrast to the `attitude` element, the `affect` element specifies an emotional category such as anger or fear. It can either express a felt (appraised) emotion or intentionally evoke a reaction by portraying an emotion that might not be felt. Although these appraisal and coping responses are implemented in the agent, the pathway of extracting the appraisals and coping responses based on the partial understanding has not yet been fully implemented.

8 Computing Participant Structure

Participant elements in Figure 1 describe the roles played by all scenario characters that are in contact [34]. Some scenario characters may be out of contact for part of the time, such as when they are in another room. Characters can be played by humans or other agents. The dialogue model tracks two types of roles relevant for participation state. First, there is the *conversational role*, which is either *active-participant* for someone who has recently taken an active part in the conversation, e.g. acting as a speaker or addressee of an utterance that is part of the conversation, or *overhearer* if playing a passive role.

The second type of role is the *utterance role*, which is how the character relates to the particular utterance. Utterance roles are *speaker, addressee, side-participant, overhearer,* and *eavesdropper*. The *speaker* utterance role is the speaker of the utterance that feedback is being given about. Our system currently is given this information for human users via the microphone that is used to pick up the speech or in agent messages used to indicate agent speech. Addressees are computed during message processing, following

the algorithm in [33]. If an explicit name is used in a vocative, then the NLU will recognize the addressee. Otherwise, if the speaker is gazing at someone, then that character is assumed to be the addressee. Otherwise, contextual information is used, including the previous speaker, previous addressee, and other participant status. Active participants in the conversation who are neither speaker nor addressee are assigned the utterance role of *side-participant*. Overhearers in the conversation (who are not speakers or addressees of the current utterance) are assigned the utterance role of *overhearer*. Finally, observers of the utterance who do not have a role in the conversation are assigned the utterance role of *eavesdropper*. We can see some changes in participant status between Figures 2 and 3. In Figure 2, both characters think Harmony will be the addressee, because the NLU component thinks Harmony will be identified in the utterance. Utah is not sure of his role at this point. In Figure 3, both agents now think Utah is the addressee, because of prior context and lack of explicit signals. Harmony thinks she is a side-participant.

9 Evocative Feedback: Conversational Goals

As described in [35], there are two types of conversational goals considered, *comprehension goals* and *participation goals*. Both are linked to participant roles, and both have internal and evocative aspects. The internal aspects refer to the agent's actual goals: for comprehension goal, whether or not to comprehend the current utterance; for participation goal, whether or not to be an active participant in the conversation. The internal aspect influences the agent's cognition and action selection. For a positive comprehension goal, the agent will expend cognitive resources to listen to and understand the utterance. For a negative comprehension goal, the agent will focus attention on other matters, such as planning next actions or utterances, emotional reasoning, or task execution. For a positive participation goal, the agent will look for opportunities to further the conversation with active conversational behavior. A negative participation goal will lead the agent to disengage, perhaps moving further away and out of contact.

The evocative conversational goals are the intention to influence others beliefs and actions related to the agent's goals. Regardless of the true internal goals, agents may want to evoke in others a belief (and resulting behaviors stemming from such a belief) that they have either the same or different conversational and participation goals. In general, it is the evocative conversational goals that are passed to the feedback behavior generation component.

There are default goals that are norms for the different utterance roles, shown in Table 2. These defaults can be overridden, however, by more specific goals or coping strategies of the agent. For instance, if the agent is an overhearer or eavesdropper who wants to join the conversation more actively, or an active participant who wants to leave

Table 2. Normative Goals for utterance participant roles

Role	Comprehension	Participation
Speaker, Addressee, Side-Participant	Yes	Yes
Overhearer	No	No
Eavesdropper	Yes	No

the conversation (as Harmony does if Utah challenges her for disliking the plan to make him Sheriff), they may adopt a participation goal that is contrary to their current status. This may lead to a similar evocative goal, and behaviors indicating the desired new status. Another example is that overhearers and eavesdroppers who hear an action with a strong intensity will decide to join the conversation more actively as it turns to this subject, and adopt a positive participation goal. Likewise, one might want to maintain status as an addressee or side-participant, and keep participation goals, while something more urgent demands attention, thus a negative comprehension goal. In Figures 2 and 3, both participation and comprehension goals are 1 for both characters. However in the accompanying video, we can see that sometimes Harmony has a participation goal of 0.

10 Behavior Generation: A Review

Here we review aspects of feedback behavior generation, first reported in [35]. The generation of nonverbal listening behaviors is controlled by the NonVerbal Behavior Generator (NVBG, [25]) and specifically by an extension to the knowledge incorporated into NVBG. NVBG receives signals of the form of Figure 1 from the virtual human system's dialog module, as well as signals such as head nods and gaze of other agents from the perceptual processing compoents.

10.1 Behaviors

To inform the knowledge used in NVBG, we turned to existing literature that describes listening behaviors depending on a listener's roles and goal. For addressees, gaze and mutual gaze conveys the intent to participate and comprehend as well as continued attention [2]. Addressees also glance at other side-participants to seek social comparison [10] or avert gaze as a signal of cognitive overload when comprehending speech [2,13]. Various nodding behaviors are used to signal that the addressee is attending [28], comprehending [5,9] or reacting to the speaker [20] and thereby to signal participation and comprehension. Head tilts and frowns are used to signal confusion [5], and various facial expressions signal emotional reactions to the content of the speech.

Side-participants exhibit similar behaviors as addressees. However, they may be less committed to comprehend the current dialog. If side-participants do not care about understanding the speaker's utterance (i.e. comprehension goal is false) but the goal is to maintain the participation status, they use glances toward the speaker [2,15]. The glances here are not to further comprehend but rather to act as a ratified participant. Mimicking or mirroring the speaker's behavior [11,26] are also exhibited, in part to hold his/her current conversation role.

Eavesdroppers have the goal to understand the conversation but their status as anonymous eavesdroppers may be threatened if they openly signal their comprehension. Thus, to maintain that role, they should avoid mutual gaze and suppress, or restrain from showing, reactions to the conversation [10]. Furtive glances at the speaker are occasionally used for better comprehension, but gaze is quickly averted to avoid mutual gaze, to prevent providing visual feedback [3] and signs of attention to the speaker [2,3,22].

Overhearers are modeled as having neither goals for participation nor comprehension and have fewer concerns about the conversation. Gaze aversion from conversation

participants is used to prevent mutual gaze [6,12] since gaze may be considered as a request signal to be included into the current conversation [2]. However, in a highly dynamic conversation, an overhearer will have difficulty avoiding attention to, comprehension of, and reactions to the conversation.

In addition to the behaviors associated with the conversation roles, behaviors are also associated with role shifts. One way to signal a change in the conversation role is for behaviors associated with the current role to be avoided and those associated with the new role to be adopted. For example, gazing at the speaker and making mutual gaze signal role shifting from a bystander to a side-participant or an addressee [2,12]. When the role shift involves changes in the participation goal, interpersonal distance is also adjusted by either moving toward or away from the group to join or leave the conversation [21].

10.2 Processing the Signals

Upon receiving input signals from the dialog module, NVBG updates the agent's role and goals and determines whether to generate a role shifting behavior. The role shifting behavior occurs when the agent's updated participation goal differs from the current participation goal. For example, if the agent's current role is overhearer (participation goal is false) and the updated role is addressee (participation goal is true), he will enter the conversation group and generate attendance behavior by gazing at the speaker and nodding. If the agent's participation goal is unchanged, NVBG generates corresponding feedback behaviors depending on the comprehension and current participation goal.

As described in Section **??**, NVBG may also receive affective information. The affective reaction dominates the reactions related to partial understanding of the speaker's utterance: an affective signal will have higher priority than the comprehension information. The affective reactions include behaviors such as smiles for joy and furrowed eyebrows for anger.

To convey the evolving comprehension level in behavior, the confidence attribute, which has range [0.0, 1.0], is used to define three categories of understanding: confusion ([0.0, 0.5)), partial understanding ([0.5, 1.0)), and understanding (1.0). The `MAXF` indicator further determines which specific feedback is generated. Since the partial understanding level may only change slightly between adjacent words, the model processes the dialog signal when the difference between previous and current partial understanding level exceeds a threshold (currently set at 0.2).

11 Conclusion and Future Work

In this paper we have presented a Function Markup Language specification for incremental feedback for multiparty conversation. It takes into account semantic and pragmatic processing and attitude toward the topic of conversation. It supports both expressive and evocative feedback for a variety of conversational roles and goals. It has been implemented, and connected to the behavior realizer developed by [35]. Future work includes linkage to coping strategies for more evocative feedback, as well as evaluating the impact of the feedback on users engaged in the SASO4 and other scenarios.

Acknowledgements. The effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Allwood, J.: Linguistic Communication as Action and Cooperation. Ph.D. thesis, Göteborg University, Department of Linguistics (1976)
2. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
3. Argyle, M., Lalljee, M., Cook, M.: The effects of visibility on interaction in a dyad. Human Relations 21, 3–17 (1968)
4. Bavelas, J.: Listeners as co-narrators. Journal of Personality and Social Psychology 79, 941–952 (2000)
5. Brunner, L.: Smiles can be back channels. JPSP 37(5), 728–734 (1979)
6. Callan, H., Chance, M., Pitcairn, T.: Attention and advertence in human groups. Soc. Sci. Inform. 12, 27–41 (1973)
7. 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: Proceedings of InterSpeech (2011)
8. DeVault, D., Sagae, K., Traum, D.: Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse 2(1) (2011)
9. Dittmann, A., Llewellyn, L.: Relationship between vocalizations and head nods as listener responses. JPSP 9, 79–84 (1968)
10. Ellsworth, P., Friedman, H., Perlick, D., Hoyt, M.: Some effects of gaze on subjects motivated to seek or to avoid social comparison. JESP 14, 69–87 (1978)
11. Friedman, H.S., Riggio, R.E.: Effect of individual differences in non-verbal expressiveness on transmission of emotion. Journal of Nonverbal Behavior 6(2), 96–104 (1981)
12. Goffman, E.: Forms of Talk. University of Pennsylvania Press, Philadelphia (1981)
13. Goodwin, C.: Conversational organization: interaction between speakers and hearers. Academic Press, London (1981)
14. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotion. Journal of Cognitive Systems Research (2004)
15. Gu, E., Badler, N.I.: Visual Attention and Eye Gaze During Multiparty Conversations with Distractions. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 193–204. Springer, Heidelberg (2006)
16. 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)
17. Heintze, S., Baumann, T., Schlangen, D.: Comparing local and sequential models for statistical incremental natural language understanding. In: Proceedings of SIGDIAL (2010)
18. Heylen, D., Kopp, S., Marsella, S.C., Pelachaud, C., Vilhjálmsson, H.H.: The Next Step towards a Function Markup Language. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 270–280. Springer, Heidelberg (2008)
19. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In: Proceedings of ICASSP (2006)
20. Ikeda, K.: Triadic exchange pattern in multiparty communication: A case study of conversational narrative among friends. Language and culture 30(2), 53–65 (2009)

21. Jan, D., Traum, D.R.: Dynamic movement and positioning of embodied agents in multiparty conversations. In: Proc. of 6th AAMAS, pp. 59–66 (2007)
22. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge University Press, Cambridge (1990)
23. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., Vilhjálmsson, H.H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
24. Kopp, S., Stocksmeier, T., Gibbon, D.: Incremental Multimodal Feedback for Conversational Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 139–146. Springer, Heidelberg (2007)
25. Lee, J., Marsella, S.C.: Nonverbal Behavior Generator for Embodied Conversational Agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
26. Maatman, R.M., Gratch, J., Marsella, S.C.: Natural Behavior of a Listening Agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
27. Marsella, S., Gratch, J.: Ema: A process model of appraisal dynamics. Journal of Cognitive Systems Research 10(1), 70–90 (2009)
28. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. AAMAS 20, 70–84 (2010)
29. Plüss, B., DeVault, D., Traum, D.: Toward rapid development of multi-party virtual human negotiation scenarios. In: Proceedings of SemDial 2011, the 15th Workshop on the Semantics and Pragmatics of Dialogue (September 2011)
30. Sagae, K., Christian, G., DeVault, D., Traum, D.R.: Towards natural language understanding of partial speech recognition results in dialogue systems. In: Short Paper Proceedings of NAACL HLT (2009)
31. Sagae, K., DeVault, D., Traum, D.R.: Interpretation of partial utterances in virtual human dialogue systems. In: NAACL-HLT 2010 Demonstration (2010)
32. Traum, D.: Semantics and pragmatics of questions and answers for dialogue agents. In: Proceedings of the International Workshop on Computational Semantics, pp. 380–394 (2003)
33. Traum, D.R., Morency, L.P.: Integration of visual perception in dialogue understanding for virtual humans in multi-party interaction. In: AAMAS International Workshop on Interacting with ECAs as Virtual Characters (May 2010)
34. Traum, D.R., Rickel, J.: Embodied agents for multi-party dialogue in immersive virtual worlds. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 766–773 (2002)
35. Wang, Z., Lee, J., Marsella, S.: Towards More Comprehensive Listening Behavior: Beyond the Bobble Head. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 216–227. Springer, Heidelberg (2011)