Auton Agent Multi-Agent Syst
DOI 10.1007/s10458-012-9215-8

Multi-party, multi-role comprehensive listening behavior

Zhiyang Wang · Jina Lee · Stacy Marsella

Abstract Realizing effective listening behavior in virtual humans has become a key area of research, especially as research has sought to realize more complex social scenarios involving multiple participants and bystanders. A human listener's nonverbal behavior is conditioned by a variety of factors, from current speaker's behavior to the listener's role and desire to participate in the conversation and unfolding comprehension of the speaker. Similarly, we seek to create virtual humans able to provide feedback based on their participatory goals and their unfolding understanding of, and reaction to, the relevance of what the speaker is saying as the speaker speaks. Based on a survey of existing psychological literature as well as recent technological advances in recognition and partial understanding of natural language, we describe a model of how to integrate these factors into a virtual human that behaves consistently with these goals. We then discuss how the model is implemented into a virtual human architecture and present an evaluation of behaviors used in the model.

Keywords Artificial intelligence · Listener feedback · Context based feedback · Nonverbal behavior

1 Introduction

Two people are having a heated conversation in a cafe. Around the cafe, various bystanders are listening to the interaction. Some avert their gaze, pretend to do something else, hoping not to become participants in the interaction but nevertheless eavesdropping on the exchange. They are hopelessly drawn to the unfolding scene, glancing at the main protagonists to glean

Z. Wang (✉)· J. Lee · S. Marsella
Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive,
Playa Vista, CA 90094, USA
e-mail: zhiyang86.wang@gmail.com; zhiyangw@usc.edu

J. Lee
e-mail: jinal@usc.edu

S. Marsella
e-mail: marsella@ict.usc.edu

Published online: 22 January 2013

information on the interaction from their dialog and nonverbal behavior, but careful to avoid the mutual gaze that might draw them into the interaction. Meanwhile, the owner of the cafe, wanting to calm the situation, is signaling his intention to join the interaction.

Developing virtual humans that can handle such ranges of participation has become an increasingly important area of research, more so as work has sought to realize more complex dramatic scenarios [17]. Work on listening behavior has tackled various aspects of this challenge. For example, there is work on dyadic interactions between human and *rapport* agents that have an implicit, fixed goal of establishing rapport but often have limited understanding of the content of the speaker's utterance [14]. The agents rather rely on low level analysis of the nonverbal and perceptual features of the human speaker's behavior that are correlated with listener feedback, such as pauses in the speaker's utterance.

Although effective in establishing rapport, this approach suffers from several limitations. First, such approaches only provide *generic feedback* [3] signaling such factors that the agent is attending. They cannot provide *specific feedback* [3], feedback tied to a deeper understanding of, and reaction to, the personal relevance of what the speaker is saying as the utterance unfolds. Another limitation is the fixed, implicit goal of establishing rapport. In practice, however, people can have very different kinds of stances towards a conversation, including even their lack of interest in understanding the speaker or a desire to leave the conversation. One approach to addressing this limitation is to have the listener's behavior be conditional on attitudinal factors [4]. Finally, the focus for listening behavior has been largely on dyadic conversations, where the listener agent is main and sole addressee, though there have been notable exceptions [21].

In this work, our interest is to realize this richer form of interaction in a multiparty setting where there may be several virtual humans interacting with one or more humans, playing a variety of roles (e.g. main addressee, side-participants, overhearer, bystander, etc.) with varying degrees of participation in, and commitment to, the conversation. The question that interests us is how these characters respond nonverbally according to their current role in the conversation, their desire to participate, their understanding of the speaker's partial utterance, as well as behavioral signals from the speaker.

This raises technical challenges of how to integrate the various factors that influence a listener, including the perception of the speaker's verbal/nonverbal behavior as well as the listener's reactions to the speaker in light of their goals for participation. In this article, we review relevant literature on listener feedback and propose a model that tailors behaviors based on how the various roles of participants influence their nonverbal behaviors and how those behaviors can signal their goals to change roles. To provide both generic and specific feedback, the model integrates information from perceptual and comprehension processes. We then discuss how the model is implemented into a virtual human architecture, relying on prior work to provide perceptual processing of the nonverbal and prosodic features of speaker behavior [34] as well as to provide natural language understanding of a speaker's partial utterance [7] and emotional reaction [32] to it. Finally, we present a preliminary evaluation of behavioral signals used in the model and discuss future directions.

2 Related work

Listener's feedback [41] has been studied both in social science and humanities research on human behavior as well as in technology work on the design of virtual agents. This section discusses the virtual agent work. Literature on human behavior that has informed this work is discussed and referenced in subsequent sections.

Research on listening behavior for virtual agents has largely focused on dyadic interactions between virtual agent and human, where the virtual agent is the main addressee. The rapport agent created by Gratch et al. [14] provides listening feedback based on nonverbal and prosodic features of the speaker's behavior, such as pauses. They demonstrated that mimicry of the speaker's behavior, including head movements and gaze aversion, improves the human speaker's sense of rapport and speech fluency. The work of Morency et al. [34] learned a model that predicts listener's nonverbal feedback from the human speaker's multimodal output features (e.g., prosody, spoken words and eye gaze).

Poppe et al. [35] focused on the timing of feedback, evaluating six different multimodal rule-based strategies based on speaker's speech and gaze to define backchannel timings for artificial listeners. Their experiment shows that the number, timing, and type (nod, vocalization, or both) of backchannel were important in how natural the behaviors were perceived. De Kok et al. [26] compared and analyzed two different methods to collect multiple perspectives of listener responses to study the appropriate and inappropriate timings to provide listener backchannels.

Because such designs are driven by the speaker's behavior and more importantly do not incorporate the listener's (virtual human) interpretation and reaction to the utterance, they are arguably more important for generic feedback as opposed to specific feedback [3]. To drive listener's specific backchannel behaviors, the virtual agent needs to interpret utterances and generate feedback based on personal relevance, as the human speaker's utterance is in progress. Research has sought to address this technological challenge in several ways. Jónsdóttir et al. [22] collected human listeners' feedback data, summarized a set of speaker's key phrases in a limited topic domain, and built a system to generate virtual listener's feedbacks when input utterance match those lexical feedback markers (key phrases). Kopp et al. [27] designed an event-based feedback model for their virtual agent Max. The model generates listener's feedback and multi-level perception and understanding by measuring the speaker's pauses and lexical information. DeVault et al. [7] used a classifier to classify partial utterances in terms of semantic frames that the agent understands.

In addition to such work on dyadic conversation, there also has been work in multiparty conversation. Jan and Traum [21] involves movement for modeling agents' participation restriction with group conversation. They developed a social force model to control the distance between agent and group center. The force pushes two agents apart if they were too close to each other, while the virtual bystander may be dragged towards the group if he/she was outside the circular participation domain.

In contrast to prior work, the focus of this paper is on a model for generating listener nonverbal feedbacks for multiparty conversations that includes both generic and specific feedback, as well as taking into account that there may be a variety of participants with varying roles and goals for their participation.

3 Conversational roles and goals

In this section we discuss the relationships between conversation roles and goals which we later use when mapping listener feedback behaviors to our model. First we define the various conversation roles by adopting the terminology used by Goffman [12]. In a conversation, the core participants are the speaker and nonspeaking participants (ratified participants), which includes the *addressee* ("addressed recipient") and the *side-participants* ("unaddressed recipients"). In addition, unofficial-participants are called *bystanders*. Goffman identifies two types of bystanders: *eavesdroppers,* who purposefully listen to the conversation, and *overhearers,* who accidentally and unintentionally hear the conversation. However, these conversation roles are not static and can change during social interaction [16,40].

We can characterize these various roles from the perspective of the goals that the role normatively presumes. Here we define two types of conversation goals: *participation goal* and *comprehension goal*. Since addressees and side-participants are part of the core conversation participants, they hold positive participation goals and to maintain this status they must act appropriately. However, bystanders (overhearers and eavesdroppers) normatively have a negative participation goal (i.e. they are not or do not want to be perceived as participants) and should act in ways that do not increase their level of participation. The conversation roles can also be further distinguish based on the comprehension goals. Eavesdroppers have stronger intentions to understand the conversation, whereas overhearers do not intend to comprehend the conversation. In contrast, addressees and side participants are expected to have positive comprehension goals and to behave consistently with those goals.

We can then summarize the relationships between conversation roles and goals as the following. Addressees have positive participation and comprehension goals; side-participants have positive participation goal and either positive or negative comprehension goal; eavesdroppers have negative participation goal but positive comprehension goal; overhearers have both negative participation and comprehension goals.

Several aspects of this classification must be stressed. First, we assume that all of the agents, regardless of their roles, have freedom to change their participation or comprehension goals. For example, although side-participants are part of the conversation group, they may want to leave the conversation at any time. Second, there is a distinction between having a goal and openly appearing (or signaling) that one has a goal. For instance, eavesdroppers may wish to comprehend the conversation (and thus have a positive comprehension goal), but because they do not want to *participate* in the conversation, it is important not to appear so since they could be drawn into the conversation and endanger their role as eavesdroppers. Third, these goals are the norm for the roles. For example, side-participants are presumed to be committed to participate and comprehend the conversation and should act consistently, but in reality they may not be concerned with understanding the content of the conversation. For this reason, it is important to consider the individual's goals for participation and comprehension distinct from the role, since the realization of behaviors may depend on both. In this paper we simplify the goals to have a binary value (positive or negative), but one can also imagine the goals having numerical values to specify the strength of the individual's desire to participate or comprehend.

4 Modeling the impact of roles and goals on behavior

The literature describes various listening behaviors depending on the conversation roles and goals, which we use to inform the knowledge used in our model. Table 1 categorizes the knowledge currently used in the model. The behaviors are categorized according to the agent's goal to participate in the conversation and its desire to comprehend the speech content. In this section, we discuss that knowledge and in the next section we cover how that knowledge is used by the model.

For addressees, gaze and mutual gaze is used to signal goals of participation and comprehension as well as continued attention [1]. This also helps addressees to get clearer visual and vocal (nonverbal/verbal) information from the speaker [24]. Addressees also glance at other side-participants to seek social comparison [10] or avert gaze as a signal of cognitive overload when comprehending speech [1,13]. In addition, various forms of nodding behaviors signal

Table 1 Relationship between conversation goals, roles, and listener feedbacks

Conversation Goals		Conversation Roles	Rule and Behavior
Participating	Not Specified Comprehending	Addressee or Side-participant	*Attendance*: gaze speaker [1,15] and head nod [34].
			Mimicry: mimic gaze direction: listener mimics speaker's gaze direction when the speaker has gazed away for a long period. Mimic head gesture: Listener repeats speaker's shaking or nodding behavior. [31]
		Switch from Eavesdropper/Overhearer to Addressee/Side-participant	*Enter group*: decrease distance by moving towards the group [21]
		Addressee or Side-participant	*Respond feedback request*: respond to other participant's communication request by gazing at the speaker [20]; Glance at speaker to indicate continued attention and willingness [1]
			Mirror Emotion: adjust its own emotion to group's emotion status [11]
	Comprehending	Addressee or Side-participant	*Understand*: head nod [5,8]
			Think: gaze aversion [1]
			Gather Information (addressee/ side-participant): glance at speaker to study speaker's facial expressions and direction of gaze [1,24] or generate social comparison behavior [10].
			Confusion: head tilt and frown [5]
			Emotion reaction: different head movement, gaze behavior and facial expression according to different emotion types.
Not Participating		Eavesdropper	*Gather Information(eavesdropper)*: glance at speaker but with faster speed and less magnitude [1] and avoid mutual gaze [2]. Show less reaction [10] .
	Not Comprehending	Overhearer	*Avoid mutual gaze*: gaze aversion [6,12]
		Switch from Addressee/Side-participant to overhearer	*Leave group*: increase distance by moving away from group [21]

that the addressee is attending [34], comprehending [5,8] or reacting to the speaker [20] and thereby to signal participation and comprehension. On the other hand, head tilts and frowns are used to signal confusion [5] and various facial expressions are shown to signal emotional reactions to the content of the speech.

Side-participants are also ratified by the speaker and 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 negative) but the goal is to maintain the participation status, they use glances toward the speaker [1,15]. The glances here are not to further comprehension but rather to act as a ratified participant. Mimicking or mirroring of the speaker's behavior [11,31] is also exhibited to hold one's current conversational 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, they avoid mutual gaze and restrain from showing (or even suppress) 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, prevent providing visual feedback [2] and signs of attention to the speaker [1,2,24].

Overhearers have neither goals for participation or 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 [1]. However, in a highly dynamic conversation, an overhearer may have difficulty avoiding attention to, comprehension of, and reactions to the conversation.

In addition to the behaviors associated with the conversation roles, there are behaviors associated with role shifts. To signal a change in the conversation role, behaviors associated with the current role are avoided and those associated with the new role can 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 [1,12]. To shift from an overhearer to an eavesdropper, increased glances at the speaker is adopted to show a desire for better comprehension. 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].

Finally, note that we have not discussed the varieties of turn-taking behaviors associated with the participant seizing the dialog turn or a speaker relinquishing his role as speaker. Such behaviors are more common components of virtual human systems so we have not discussed them here.

5 Implementation

In this section, we describe the listener feedback model that operates within the Austin virtual human system developed at the USC Institute for Creative Technologies (e.g., [39]). The virtual human system has been used to realize a range of scenarios. In the most recent scenario, SASO4, users engage in face-to-face interactions with virtual humans and practice multi-party negotiation skills. Figure 1 shows human trainees interacting with two virtual humans Utah and Harmony. The SASO4 scenario consists of multiple components to support real-time interactions including cognitive processing, perceptual processing, non-verbal behavior generation and realization of behaviors through animations. These components communicate with each other by exchanging information through a messaging system.

The listener feedback model has been developed as an extension to the (NVBG) [29], the behavior planner of our virtual human system, by constructing a set of feedback rules in addition to the existing rules and extending the set of input messages NVBG processes. The listener feedback model in particular processes information passed down from the cognitive and perceptual processing modules. The entire virtual human system (including the listener feedback model) operates in real-time. In the following sections, we first provide an overview of the NVBG then discuss the details of the listener feedback model, including the input signals and the mapping to various behaviors.

5.1 Nonverbal behavior generator

The NVBG [29] is a tool that automates the selection and timing of nonverbal behaviors for virtual agents. It uses a rule-based approach based on the literature of psychological research [1,9,18,23,25,33] and our own study of corpora of human nonverbal behaviors (see [29] for details). NVBG realizes a robust process that does not make any strong assumption about the markup of the agent's communicative intent (e.g. affective state, emphasis points, and

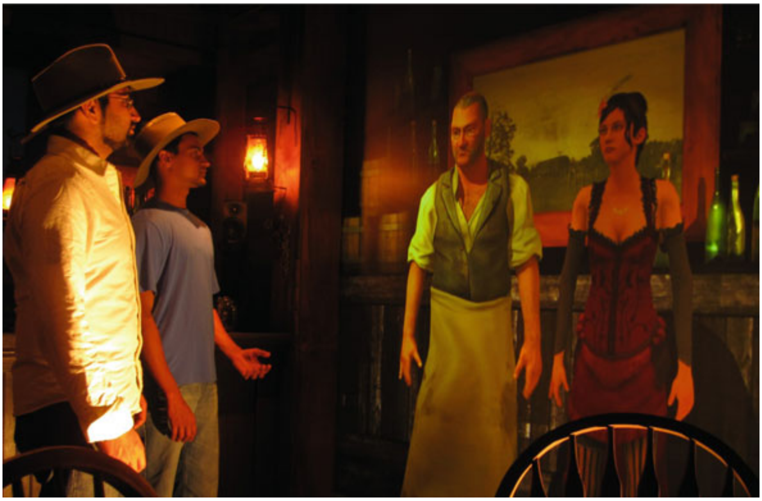


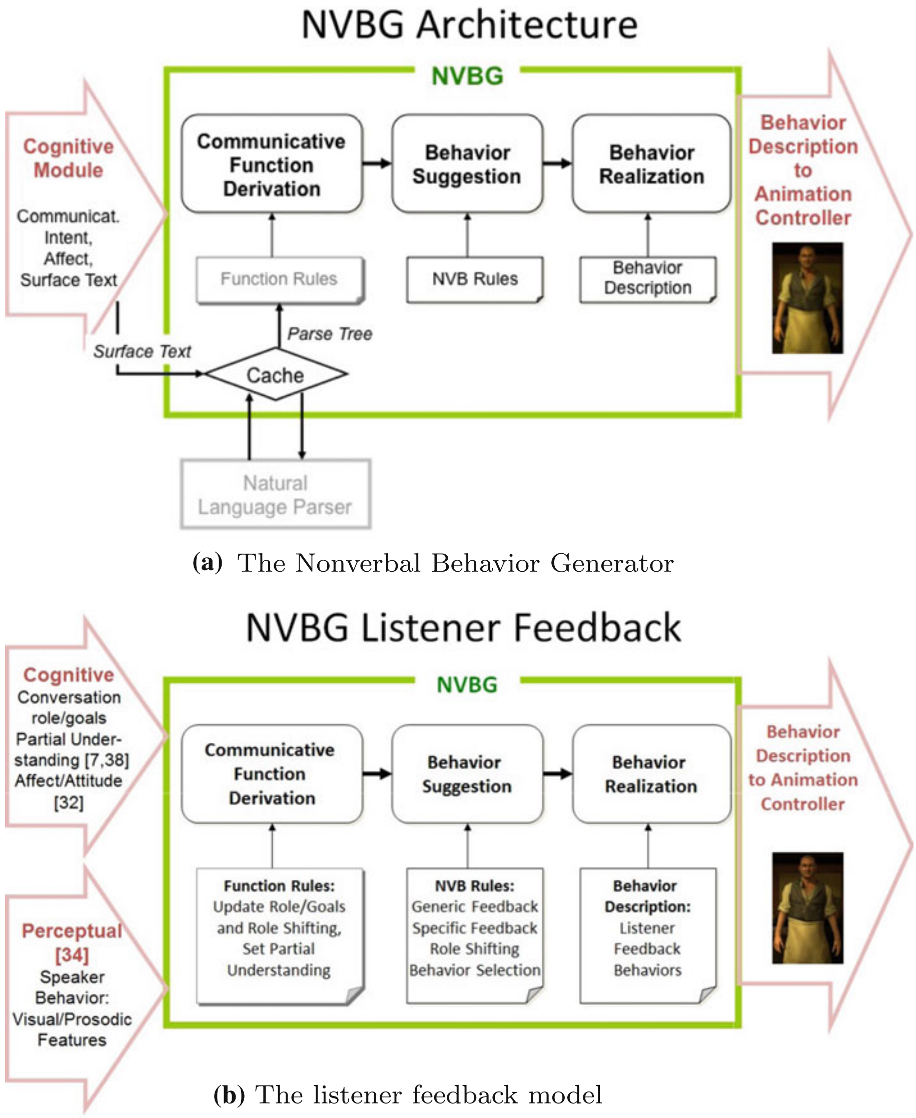
Fig. 1 SASO4 Gunslinger scenario. Users participate in a multi-party negotiation with virtual humans in a mixed reality environment

attitude) in the surface text. In the absence of such markup, NVBG can extract information from the lexical, syntactic, and semantic structure of the surface text and can support the generation of believable nonverbal behaviors.

The architecture of the NVBG is shown in Fig. 2a. Information about the agent's communicative intents, emotional state and surface text is passed from the agent's cognitive module to NVBG. It then uses the given information as well as information obtained through analysis of the syntactic and semantic structure of the surface text to infer communicative functions the agent intends to deliver. Some examples of these communicative functions include affirmation, inclusivity, intensification, etc. (see [29] for details). NVBG then goes through a behavior suggestion stage, in which a set of *nonverbal behavior rules* that map between communicative functions to specific nonverbal behaviors are triggered to suggest candidate behaviors. If there are two or more rules overlapping with each other causing conflict, NVBG resolves the conflict by filtering out the rule with lower priority. The priority value of each rule has been set through our earlier study of human behaviors using video corpora. The final set of behaviors are described using Behavior Markup Language [28] and passed to the animation system.

5.2 Listener feedback model

The listener feedback model extends the NVBG framework by constructing additional rule sets to the existing NVBG rules. Figure 2b shows the architecture of the model and the information flow specific to the listener feedback model. In this model, we make a distinction between generic feedback and specific feedback, handling them using different rule sets. The listener feedback model receives input signals from external modules that provide information about the agent's roles/goals and the comprehension of the speech as a listener as well as perceptual information about the speaker's behaviors. The model analyzes this information and triggers relevant listener feedback rules, which are mapped to various nonverbal behaviors. The input signals also govern conflict resolution when more than one listener feedback


NVBG Architecture
NVBG
Cognitive Module
Communicat. Intent, Affect, Surface Text
Communicative Function Derivation
Behavior Suggestion
Behavior Realization
Function Rules
NVB Rules
Behavior Description
Parse Tree
Surface Text
Cache
Natural Language Parser
Behavior Description to Animation Controller
NVBG Listener Feedback
NVBG
Cognitive
Conversation role/goals
Partial Under-standing [7,38]
Affect/Attitude [32]
Perceptual [34]
Speaker Behavior: Visual/Prosodic Features
Communicative Function Derivation
Behavior Suggestion
Behavior Realization
Function Rules: Update Role/Goals and Role Shifting, Set Partial Understanding
NVB Rules: Generic Feedback Specific Feedback Role Shifting Behavior Selection
Behavior Description: Listener Feedback Behaviors
Behavior Description to Animation Controller

(a) The Nonverbal Behavior Generator

(b) The listener feedback model

Fig. 2 Architectures of the nonverbal behavior generator (NVBG) and the listener feedback model incorporated within the NVBG. The *bottom figure* shows information flow specific to the listener feedback model

rules are triggered. The following sections discuss the details of the input signals and the different rule sets.

5.2.1 Inputs

In addition to the input signals NVBG processed previously, NVBG has been extended to receive and process streams of signals required for the listener feedback model, mainly from the virtual human system's cognitive module. These signals are broadly classified as cognitive processing and perceptual processing signals, as shown in Fig. 2b. The cognitive processing signal provides (a) the virtual human's current conversational role as well as participation and comprehension goals, (b) incremental partial understanding information and (c) affective and attitude signals.

The conversational role and goals are sent by the virtual human's dialogue module at the start of the conversation and are updated as the interaction between participants unfold. The listener's incremental interpretation of partial utterances is realized by DeVault et al.'s classifier [7,38], which provides a semantic interpretation as well as a measure how confident the agent is of their current understanding and a measure of whether the agent believes it will understand better if it continues listening.

The affective signal is the agent's valenced reactions to its evolving interpretation of the speaker's utterance. The signal comes from the system's domain independent computational model of emotion, **EM**otion and **A**daptation (EMA [32]). EMA is based on appraisal theories of emotion [36] that argue emotion arises from a person's subjective interpretation of their relationship with their environment. This interpretation is in terms of a set of criteria (variously called appraisal dimensions, variables or checks), such as whether an event is desirable, who is responsible for the event and the degree to which the person has control over the event. Specific emotions are associated with certain configurations of these criteria. For example, if the agent's interprets the speaker's partial utterance as deliberately proposing an action to harm the agent, then the agent's reaction will be anger.

We have formatted the input signals in functional markup language, <fml>, messages [19]. The cognitive processing signal is termed *vrBCFeedback* and the perceptual processing signal is termed *vrVision*. Table 2 specifies the format of this message and Fig. 3 provides a sample message.

The perceptual processing signal is provided by the virtual human's perceptual model which includes information about the *speaker's* behaviors such as the head movements, gaze direction, pitch accents, and speech pauses. It also includes predictions of the listener's backchannel nods, based on the work of Morency et al. [34]. When the feedback model triggers the *Attendance* rule (see Table 1), the model will propagate these predicted head nods. Table 3 specifies the format of this *vrVision* message and Fig. 4 provides a sample.

5.2.2 Mapping to listening behaviors

Upon receiving the input signals, function derivation 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=negative) and the updated role is *addressee* (participation goal=positive), he will enter the conversation group and generate attendance behavior by gazing at the speaker and nodding. The role shifting behaviors refer to rule *Enter group* and *Leave group* in Table 1.

If the agent's participation goal is unchanged, behavior suggestion's generic and specific rules generate corresponding feedback behaviors depending on the comprehension goal. In particular, the cognitive processing signals are handled by the specific feedback rules and the perceptual processing signals are handled by the generic feedback rules. In our model, both rule sets are active, generating feedbacks concurrently. However, one might instead argue for a more complex interaction. For example, once the partial understanding has achieved high confidence, specific feedback may dominate generic feedback.

The generic feedback rules generate behaviors when the agent's participation goal is positive and the comprehension goal is not positive, since comprehension feedback has higher

Table 2 Cognitive processing message format

Name	Parameter Type	Description
participant	element	participant name and role
agent	attribute	agent name
role	attribute	speaker, addressee, side-participant, eavesdropper, overhearer
feedback	element	feedback related info.
agent	attribute	feedback agent name
participantion-goal	attribute	positive or negative goal
utterance	attribute	dialog message id
progress	attribute	time mark of feedback word
time-point	attribute	real time in seconds
complete	attribute	reached the last word of entire utterance or not
affect	element	affective feedback
type	attribute	joy,fear,surprise,disgust,anger,sadness
stance	attribute	presentation type: leaked, hide
instensity	attribute	instensity of affect
attitude	element	attitude feedback
type	attribute	agree, disagree, like dislike, insterested, uninterested
stance	attribute	presentation type: leaked, hide
instensity	attribute	instensity of attitude
dialog-act	element	utterance info.
type	attribute	listen, speak
partial-text	element	sub-element of <dialog-act>, partial utterance content
partial-sem	element	sub-element of <dialog-act>, feedback on partial utterance
confidence	attribute	partial understanding state
maxf	attribute	future understanding capability

```
<act>
<participant agent=“utah” role=“addressee”/>
<participant agent=“harmony” role=“side-participant” />
<participant agent=“ranger” role=“overhearer”/>
<fml>
<feedback agent=“harmony” participation-goal = “1”
comprehension-goal=“1” utterance=“elder-al-hassan266” progress=“T3”
time-point =“0.34” complete=“no”/>
<dialog-act type=“listening”>
<partial-text> these people</partial-text>
<partial-sem confidence=“0.6” maxf=“1”/>
</dialog-act>
<affect type=“Fear” target=“people” stance=“leaked” intensity=“0.8”/>
<attitude type=“dislike” target=“these” stance =“hide” intensity=“0.7”/>
</fml>
</act>
```

Fig. 3 Example vrBCFeedback message for Harmony agent

priorty to generate when it is positive. They process the speaker’s perceptual information and generate behaviors such as gazing at the speaker, head nods, or mimicking the speaker’s gaze direction and facial expressions. This includes listener feedback rules such as *Attendance, Respond feedback request*, and *Mirror Emotion* in Table 1.

The specific feedback rules process affective or attitudinal information as well as the comprehension information. In this model, the agent’s emotional reaction is stronger than the reactions related to the partial understanding of the speaker’s utterance, therefore any incoming affective or attitudinal signal will have higher priority than the comprehension information. The affective reactions include behaviors such as smiles for joy and furrowed eyebrows for anger (rule *Emotion reaction*).

Table 3 Perceptual processing message format

Name	Parameter Type	Description
participant	element	sub-element of <act>
id	attribute	agent name
role	attribute	speaker
attribute	element	speaker's visual and prosidic feature
name	attribute	feature category: head, face, gaze, vocal
type	attribute	feature type: nod (head), smile (face), pause (vocal)
direction	attribute	gaze direction
magnitude	attribute	magnitude of speaker's behavior
probability	attribute	probability of suggested speaker's behavior
predict	element	predicted listener generic feedback
name	attribute	feature category: head, face, gaze, vocal
type	attribute	feature type: nod (head), smile (face), pause (vocal)
direction	attribute	gaze direction
magnitude	attribute	magnitude of speaker's behavior
probability	attribute	probability of suggested behavior for listener to generate

```
<act> <participant id="ranger" role ="speaker"/>
<fml>
<attribute name="head" type="nod" magnitude="0.3" duration="0.8"/>
<attribute name="face" type="smile" magnitude="0.3" duration="0.8"/>
<attribute name="gaze" direction="POLAR 45" magnitude="5"/>
<attribute name="vocal" type="pause"/>
<predict name="head" type="nod" magnitude="0.5" probability="0.6"/>
<predict name="face" type="smile" magnitude="0.5" probability="0.6"/>
<predict name="gaze" direction="POLAR -45" magnitude="0.5" probability="0.6"/>
</fml>
</act>
```

Fig. 4 Example vrVision message

Table 4 Selection of comprehension feedback rules

			Feedback Rules with Different Roles		
Input	Confidence	maxf	addressee side-participant	eavesdropper	overhearer
	[0.0, 0.5)	0	Confusion	Idle	Idle
		1	Attendance		
	[0.5, 1.0)	0	Partial Understand/ Think/ Idle	Gather Info.(eavesdropper)	
		1	Partial Understand/ Think/ / Attendance/ Gather-Info. (addressee/side-participant)	Idle	
	1.0	0	Understand		
		1			

The comprehension information contains two parameter values: *confidence* (range [0.0, 1.0]) and *maxf* (0 or 1). The confidence value indicates how confident the agent believes it understands the speaker's utterance and maxf indicates whether the agent believes it will understand the utterance more if it keeps on listening. We define three categories of understanding based on the confidence value: confusion ([0.0, 0.5)), partial understanding ([0.5, 1.0)), understand (1.0). The *maxf* value further determines which specific feedback is generated. Table 4 shows how the *confidence* and *maxf* values determine the selection of listener feedback rules.

Since the cognitive processing signals are sent out by the natural language understanding (NLU) module after each word (roughly around 400 ms/message) whereas it takes substantially longer time to realize a feedback behavior, the model needs to determine which signal to process and when to generate feedback behaviors. In our model, a new behavior is generated only after the previous behavior has been completed by the animation system. Furthermore, since the agent's 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 certain threshold (currently set at 0.2). Similarly, perceptual processing signals are processed only after the previous behavior has been realized by the animation system.

Since the listener feedback model has been implemented in the NVBG platform [29], the output message is in behavior markup language <bml> format [28].

6 Example

We now go through an example scenario to demonstrate how the listener's nonverbal backchannel behaviors are generated. The listener feedback model, including roles, goals and feedback behaviors, has been fully implemented in our virtual human system and integrated with the incoming cognitive processing and perceptual processing signals as well as the SmartBody character animation system [37].

The example is from the mixed reality Gunslinger scenario (see Fig. 1) [17]. The interactive experience takes place in a 19th century American old west small town saloon. The human user plays the Ranger, who has the task of bringing the murderous gunslinger Rio Lane to justice. After killing Rio in a gunfight, the Ranger's final task before leaving the scenario is to appoint a local sheriff that will maintain law and order in the town.

In this example, the human user (Ranger) tries to convince Utah, the bartender, to take the job of sheriff. Utah is in favor of this offer. On the other hand, Harmony, the owner of the saloon and friend of Utah, hears the conversation first as an overhearer and shows negative reactions to Ranger's proposal then switches her role to a side-participant to join the conversation.

Below we take an excerpt from the scenario when Ranger offers the job to Utah and describe the input signals and corresponding output behaviors along seven different points in the utterance. We represent the agent's conversational roles and goals as "*Role(participation goal, comprehension goal)*." For example, "*Eavesdropper(0,1)*" denotes that the role is eavesdropper with negative participation goal and positive comprehension goal. Table 5 presents the feedbacks according to different input signals for each agent. The columns are the index for the seven points, input signals and output feedback. Figure 5 shows the screenshot for listeners' feedback behaviors on each stroke point.

Ranger (Speaker):

"Utah①, it's time for me to move on② and the③ town will need a strong leader④ like yourself⑤ to⑥ maintain law and order⑦."

From Table 5 we can see that even with the same input perceptual and partial understanding signals, the agent's feedback is significantly different according to the different conversation roles and goals. This example demonstrates that the feedback model enables the agents with a rich set of reactions that go beyond simply establishing rapport.

7 Behavior assessments

An evaluation of the full system has not yet been performed. However a preliminary question has been explored concerning whether people can interpret, or decode, the behaviors the model employs, especially the behaviors related to the comprehension goal: gathering infor-

Table 5 Feedback behaviors for the example utterance. Here we list the feedback rules for each feedback points

Index	Input Signal				Output Feedback			
	Perceptual	Dialog			Utah		Harmony	
	(Common)	(Common)	(Utah)	(Harmony)	Generic	Specific	Generic	Specific
①	Predict: nod prob.(0.6)	PU: maxf(0), confidence(1.0)	Role(P,C): A(1,1)	Role(P,C): O(0,0)	Response feedback request	Role shifting: Enter group, Attendance	-	Idle
②	-	PU: maxf(1), confidence(1.0)	-	-	-	Understand	-	Idle
③	-	PU:maxf(0), confidence([0.5,1.0))	-	Role(P,C): E(0,1)	-	Partial Understand/ Think/ Idle	-	Role shifting: Avoid Mutual Gaze
④	-	PU: maxf(1), confidence(1.0)	-	-	-	Understand	-	Idle
⑤	-	PU: maxf(1) confidence(1.0)	Affective: surprise	Affective: surprise	-	Emotion (surprise)	-	Idle
⑥	-	PU: maxf(1) confidence(1.0),	-	Role(P,C): SP(1,1)	-	Understand	-	Role shifting: Enter Group, Attendance
⑦	-	PU: maxf(1) confidence(1.0)	Attitude: like	Attitude: dislike	-	Attitude: (like)	-	Attitude: (dislike)

Refer to Table 1 for the specific behaviors. “Idle” indicates idle behavior

P participation goal; *C* comprehension goal; *1* positive; *0* negative; *PU* partial understanding; *A* addressee; *SP* side participant; *E* eavesdropper; *O* overhearer

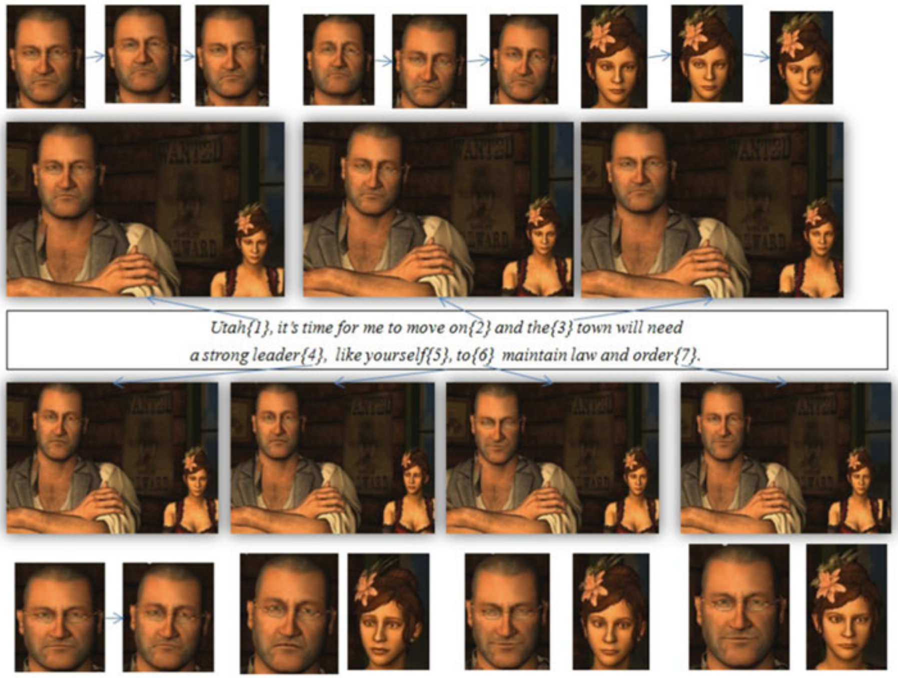

Utah{1}, it's time for me to move on{2} and the{3} town will need
a strong leader{4}, like yourself{5}, to{6} maintain law and order{7}.

Fig. 5 Screenshots of the example scenario: mapping between the utterance and the listening behaviors. { } are stroke points and the *arrows* indicate the listeners' behaviors at that moment. The male agent is Utah, the female agent is Harmony

mation, eavesdropping, thinking, understand and confusion. As opposed to the decoding of emotional states that has been extensively studied, there is less evidence that the behaviors we posit for these states can be effectively decoded. If the behaviors are highly ambiguous to observers, it undermines the rationale for employing the partial understanding component of the model.

To do an initial assessment of this, we created seven video clips of virtual listener nonverbal feedback, based on the rules and behaviors listed in the "Comprehending" signal row of Table 1. In each video, there is a hidden speaker (behind the camera) talking to a virtual human in front of the camera who provides nonverbal feedback (e.g. head nods, facial expressions, etc.) to the speaker. Each subject watched all seven videos. The speech played in the background is the same for each video, while the agent's behaviors were different. The speech is gibberish (nonsense content), so the subject is not influenced by the utterance content itself. After watching each video, the subject was given a forced choice questionnaire that asked him/her to select the best interpretation from a list of the alternative comprehension goals.[1] We recruited 15 subjects to participate in the experiment. Table 6 shows the results. The rows are the rules and behavior exhibited in the video and the columns are the subject's interpretation of the behavior with each cell listing how many subjects chose that interpretation. The hypothesized interpretation is in bold.

[1] The forced choice obviously simplifies this decoding task for the observer but the use of gibberish makes it harder.

Table 6 Behavior assessments results

Rule/Behavior	**Interpretation**					**Recog. Rate**
	Confusion	Think	Gather-Info.	Eavesdrop	Understand	
Confusion/Head Tilt & Frown	**11**	2	1	0	1	73.33%
Think/Gaze Avert ①	2	**11**	0	0	2	73.33%
Think/Gaze Avert ②	1	**11**	2	1	0	73.33%
Gather-Info(ratified)/Scan①	0	1	**13**	1	0	86.67%
Gather-Info(ratified)/ Scan②	0	4	**10**	0	1	66.67%
Gather-Info(eavesdropper)/ Glance at speaker	0	1	3	**10**	1	66.67%
Understand/Nod	1	0	0	0	**14**	93.33%

Think ① and ②: gaze aversion with different direction, magnitude, speed and duration. Gather information (ratified participant)①: glance between speaker's head and chest; ②: glance between speaker's head and lumbar

The result shows that for every category, the dominant choice was the hypothesized interpretation. However, some behaviors clearly could be improved if our goal was to reduce decoding ambiguity further. Of course, this is an assessment of just one aspect of the design. We discuss additional evaluation goals in the next section.

8 Conclusion and future work

In this paper, we have described the Listener Feedback Model for virtual agents in multi-party conversations. The vision behind this model is that the agent will generate both generic feedback and specific feedback conditioned on a variety of factors, including the speaker's behavior, the listener's role and the desire to participate in the conversation as well as the unfolding comprehension of partial utterances. The model has been implemented within the nonverbal behavior generation component of our virtual human system and drives the agent to perform feedback automatically and dynamically.

This work will be extended in several ways. A range of extensions to the model itself are being considered. In particular, we are interested in incorporating other factors which may influence listener's feedback, such as interpersonal relationship, personality, and culture. There are alternative ways in achieving this; the current listener feedback rules could be further added to and modified according to the varying factors or a data-driven approach (e.g., [30]) could be employed to learn models using different sets of data reflecting variations of those factors. Also, as mentioned earlier, there are alternative approaches to how the generic and specific feedback interact that need to be assessed.

One pressing empirical question concerns how the specific feedback influences the human-virtual human interaction. There have been studies looking at the impact of the generic feedback of rapport agents, but the kind of specific feedback we are discussing here may have a more profound impact. The feedback might facilitate the interaction, providing the human with important information to guide the interaction. On the other hand, the virtual human's reaction to its partial understanding of the utterance, such as a look of anger, could also conceivably cause pauses or disfluency in the human speaker. This in turn may well throw off speech recognition/natural language understanding, thereby impacting the virtual human's ability to recognize and understand the utterance. Regardless, we expect the feedback to impact the human user's impression of, and expectations about, the virtual human as well as impact potentially a range of relational factors such as trust. Overall, the design of the virtual human may have to fundamentally change to take into account this finer grain of interactivity.

Acknoledgments This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also would like to thank our colleagues Drs. David DeVault, LP Morency and David Traum for all their help in implementing this model.

References

1. Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.
2. Argyle, M., Lalljee, M., & Cook, M. (1968). The effects of visibility on interaction in a dyad. *Human Relations*, *21*, 3–17.
3. Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*, 941–952.
4. Bevacqua, E., Pammi, S., Hyniewska, S. J., Schroder, M., & Pelachaud, C. (2010). Multimodal backchannels for embodied conversational agents. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents* (pp. 194–200). Philadelphia: IVA.
5. Brunner, L. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology*, *37*(5), 728–734.
6. Callan, H., Chance, M., & Pitcairn, T. (1973). Attention and advertence in human groups. *Social Science Information*, *12*, 27–41.
7. DeVault, D., Sagae, K., & Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, *2*(1), 143–170.
8. Dittmann, A., & Llewellyn, L. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, *9*, 79–84.
9. Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology* (pp. 169–248). Cambridge: Cambridge University Press.
10. Ellsworth, P., Friedman, H., Perlick, D., & Hoyt, M. (1978). Some effects of gaze on subjects motivated to seek or to avoid social comparison. *Journal of Experimental Social Pscyhology*, *14*, 69–87.
11. Friedman, H. S., & Riggio, R. E. (1981). Effect of individual differences in non-verbal expressiveness on transmission of emotion. *Journal of Nonverbal Behavior*, *6*(2), 96–104.
12. Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.
13. Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.
14. Gratch, J., Wang, N., Gerten, J., Fast, E. & Duffy, R. (2007). Creating rapport with virtual agents. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*. Paris: IVA.
15. Gu, E. & Badler, N. (2006). Visual attention and eye gaze during multipartite conversations with distractions. In *Proceedings of the 6th, International Conference on Intelligent Virtual Agents*. Marina Del Rey: IVA.
16. Hanks, W. F. (1996). *Language and communicative practices*. Boulder: Westview Press.
17. Hartholt, A., Gratch, J., Weiss, L., & Team, T. G. (2009). At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents* (pp. 500–501). Berlin/Heidelberg: Springer.
18. Heylen, D. (2005). Challenges ahead: Head movements and other social acts in conversations. In *Social presence cues symposium*. Hatfield: University of Hertfordshire.
19. Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., & Vilhjlmsson, H., (2008). The next step towards a functional markup language. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (pp. 270–280). Berlin/Heidelberg: Springer.
20. Ikeda, K. (2009). Triadic exchange pattern in multiparty communication: A case study of conversational narrative among friends. *Language and Culture*, *30*(2), 53–65.
21. Jan, D., & Traum, D. R. (2007). Dynamic movement and positioning of embodied agents in multiparty conversations. In *Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems* (pp. 59–66). Toronto: AAMAS.
22. Jónsdóttir, G. R., Gratch, J., Fast, E., Thórission, K. R. (2007) Fluid semantic back-channel feedback in dialogue: Challenges and progress. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*. Paris: IVA.

23. Kendon, A. (1972). Some relationships between body motion and speech. In A. Seigman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–216). Elmsford, New York: Pergamon Press.
24. Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge: Cambridge University Press.
25. Kendon, A. (2002). Some uses of the head shake. *Gesture*, *2*(36), 147–182.
26. Kok, I., & Heylen, D. (2011). Appropriate and inappropriate timing of listener responses from multiple perspectives. In H. Vilhjlmsson, S. Kopp, S. Marsella, & K. Thrisson (Eds.), *Intelligent virtual agents*. Lecture Notes in Computer Science (vol. 6895). Berlin/Heidelberg: Springer.
27. Kopp S., Allwood J., Grammer, K., Ahlsen, E., & Stocksmeier, T. Modeling embodied feedback with virtual humans. In *Modeling communication with robots and virtual humans* (Vol. 4930, pp. 18–37). Berlin/Heidelberg: Springer.
28. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thrisson, K., & Vilhjlmsson, H. (2006). *Towards a common framework for multimodal generation: The behavior markup language* (Vol. 4133, pp. 205G217). Berlin/Heidelberg: Springer.
29. Lee, J., & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents* (pp. 243–255). IVA: Marina del Rey.
30. Lee, J., & Marsella, S. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Transactions on Multimedia*, *12*(6), 552–562.
31. Maatman, R., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. In *Proceedings of the 5th International Conference on Intelligent Virtual Agents* (pp. 25–36). IVA: Kos. .
32. Marsella, S., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, *10*(1), 70–90.
33. McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(24), 855–878.
34. Morency, L.-P., de Kok, I., & Gratch, J. (2008). A probabilistic multimodal approach for predicting listener backchannels. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (pp. 70–84). IVA: Tokyo.
35. Poppe, R., Truong, K., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents*. Lecture Notes in Computer Science (Vol. 6356, pp. 146–158). Berlin Heidelberg: Springer.
36. Smith, C. A., & Lazarus, R. (1990). Emotion and adaptation. In L. A. Pervin (Ed.), *Handbook of personality: Theory & research* (pp. 609–637). New York: Guilford Press.
37. Thiébaux, M., Marshall, A., Marsella, S., & Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. In *Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (pp. 151–158). Estoril: AAMAS.
38. Traum, D., DeVault, D., Lee, J., Wang, Z., & Marsella, S. (2012). Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent virtual agents*. Lecture Notes in Computer Science (Vol. 7502, pp. 275–288). Berlin/Heidelberg: Springer.
39. Traum, D., Marsella, S., Gratch, J., Lee, J., & Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (pp. 117–130). IVA: Tokyo.
40. Vertegaa, R., der Veer, G. C. V., & Vons, H. (2000). Effects of gaze on multiparty mediated communication. *Proceedings of Graphics, Interface* (pp. 95–102). New York: ACM Press.
41. Yngve, V. (1970). On getting a word in edgewise. In *Papers from the 6th regional meeting* (pp. 567–578). Chicago: Chicago Linguistic Society.