J Multimodal User Interfaces (2011) 4:97–118
DOI 10.1007/s12193-011-0060-x

ORIGINAL PAPER

Continuous interaction with a virtual human

Dennis Reidsma · Iwan de Kok · Daniel Neiberg · Sathish Chandra Pammi · Bart van Straalen · Khiet Truong · Herwin van Welbergen

Received: 5 February 2011 / Accepted: 29 April 2011 / Published online: 27 May 2011

Abstract This paper presents our progress in developing a Virtual Human capable of being an attentive speaker. Such a Virtual Human should be able to attend to its interaction partner while it is speaking—and modify its communicative behavior on-the-fly based on what it observes in the behavior of its partner. We report new developments concerning a number of aspects, such as scheduling and interrupting multimodal behavior, automatic classification of listener responses, generation of response eliciting behavior, and strategies for generating appropriate reactions to listener responses. On the basis of this progress, a task-based setup for a responsive Virtual Human was implemented to carry out two user studies, the results of which are presented and discussed in this paper.

Keywords Virtual humans · Attentive speaking · Listener responses · Continuous interaction

1 Introduction

Continuous interaction is one of the fundaments underlying Attentive Speaking and Active Listening for Virtual Humans (VHs). Attentive Speaking and Active Listening require that a Virtual Human be capable of *simultaneous perception/interpretation and generation* of communicative behavior. A Virtual Human should be able to signal its attitude and attention while it is listening to its interaction partner, and be able to attend to its interaction partner while it is speaking—and modify its communicative behavior on-the-fly based on what it observes in the behavior of its partner.

This paper presents our progress in developing a Virtual Human that supports continuous interaction. We discuss our work on perception capabilities, involving development and evaluation of automatic classifiers for vocal listener responses. We also present our work on multimodal generation capabilities: flexible and adaptive scheduling and planning including graceful interruption, generation of response eliciting behavior, and models for appropriate reactions to listener responses. Finally, we worked on a task-based setup in which a Virtual Human explains a route to a user, combining the above mentioned capabilities with a Wizard of Oz in order to have the Virtual Human act as an Attentive Speaker. Using this setup, two user studies were carried out,

This paper is base upon a project report of the eNTERFACE'10 Summer Workshop on Multimodal Interfaces [42].

D. Reidsma (✉) · I. de Kok · B. van Straalen · K. Truong · H. van Welbergen
Human Media Interaction, University of Twente, Postbus 217, 7500AE, Enschede, Netherlands
e-mail: d.reidsma@utwente.nl

I. de Kok
e-mail: i.a.dekok@utwente.nl

B. van Straalen
e-mail: b.vanstraalen@utwente.nl

K. Truong
e-mail: k.p.truong@utwente.nl

H. van Welbergen
e-mail: h.vanwelbergen@utwente.nl

D. Neiberg
Dept. of Speech, Music and Hearing, KTH Royal Institute of Technology, Lindstedtsv. 24, 100 44 Stockholm, Sweden
e-mail: neiberg@speech.kth.se

S.C. Pammi
Language Technology Lab, German Research Center for Artificial Intelligence DFKI, Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
e-mail: Sathish.Pammi@dfki.de

the results of which are presented and discussed at the end of this paper.

In addition to the results presented in this paper, the project yielded a number of deliverables that are released for public access, among which a public release of Elckerlyc[1] (a new platform for building Virtual Humans), and a database of motion capture animations containing over 100 direction-giving-task related gestures in the route giving domain.[2]

2 Background and motivation

We work on a VH in a conversational setting that uses speech, face expressions, and gestures to express itself. In general, such VHs tend to be developed using a one-speaks-at-a-time based interaction paradigm in which the processing of input —and preparation of the VHs reaction— start at the end of an utterance of the human interlocutor. Such an interaction paradigm introduces decreased responsiveness and interactiveness. If the interaction capabilities of VHs are to become more human-like and VHs are to function in social settings, their design should shift from this end-of-utterance based paradigm to one of *continuous interaction* in which all partners perceive each other, express themselves, and coordinate their behavior to each other, continually and in parallel [39, 51]. This requires the VH to be capable of immediate adaptation—in content and in timing—to the dynamics of the environment and the user.

VHs that can deal with continuous interaction have more possibilities to support conversational alignment with users, leading to increased rapport [30], and generally will support more flexible dialog processes [51]. This need for continuous interaction is also reflected in the recent developments combining incremental perception and incremental generation into incremental dialog systems [45]. Incremental perception means that processing of the user's utterances starts before the utterance has been completed, allowing for much faster response times. Incremental generation [49] means that the generation of behavior starts before the perception submodules finished processing the user's utterance, which leads to more natural dialogs—sometimes even forcing the speech synthesis to produce fillers, like "eh", in a very human-like way and for similar reasons, simply because the speech synthesis module is being told that this is an appropriate moment to say something, while the required content of the speech is not yet known.

Our long term goal is to explore this kind of coordination behavior in VHs. This involves modeling and implementing the sensing, processing, interaction and generation for what we call continuous interaction. A continuous interactive VH will be able to perceive the user and generate conversational behavior fully in parallel, and can coordinate behavior with perception continuously—a capability which is not yet present in most state-of-the-art VHs.

One of the major sources of overlap in conversation, and therefore a very good domain for addressing continuous interaction capabilities in VHs, are Listener Responses [19], explained in more detail in the next section. We will take a first step towards our goal by making a VH that is capable of actively dealing with Listener Responses from the user, while the VH is speaking. The VH explains a route through a city, in such a way as to elicit Listener Responses (e.g., "uh-huh", "mmm") from the user at various points in the explanation. If Listener Responses occur, the VH is able to adapt its ongoing explanation to deal with the Listener Response. In the experimental setup described later in this paper, this adaptation focuses on adjusting the *timing* of, and pauses in, the utterances of the VH.

3 Listener responses and attentive speaking

In human-human conversations, overwhelmingly one person speaks at a time [43]. At the same time, there are also short but frequent segments of overlapped speech [44]. A listener shows his or her interest, attention and/or understanding in many ways during the speaker's utterances: through gaze direction and eye contact, using face expressions, using short utterances like "yeah", "okay", and "hm-m", etcetera. A speaker often will give the listener opportunities for such responses, but will also actively receive the responses, and adjust his or her utterances to the occurrence and content of these responses. A speaker may also actively elicit responses using, e.g., face expressions or vocal cues. In short, interlocutors continuously and in coordination with one another show *attentive speaking* and *active listening* behavior [3, 11]. In this section we discuss Listener Responses and attentive speaking in more detail.

3.1 Listener responses

Listener Responses [19] are short utterances (for example: "yeah", "mhm", "uhu"), vocalizations and/or (facial) gestures which are interjected into the speakers' account without causing an interruption, or being perceived as competitive of the floor, which allows for them to occur as overlapped speech. They serve many functions, of which the most important is to neutrally signal that the listener hears that the speaker is talking. A Listener Response having this function is often referred to as a back-channel [13]. Other functions are usually added to the list, such as acknowledgments [1, 13], continuers [22] and assessments [22, 27]. As

[1] http://elckerlyc.sourceforge.net.

[2] http://hmi.ewi.utwente.nl/mocapdb.

pointed out by Fujimoto [19], the terminology is not standardized as well as confusing, especially since the name of the entity is sometimes the same as one of the specific functions it may serve (as is the case for, e.g., the term 'backchannel'). In this paper we generally use the umbrella term Listener Response to avoid these ambiguities. Listener Responses may also be used as carriers of other subtle information conveyed by intonation, voice quality, rhythm, syllabification, content of the words, and by accompanying face expressions, head nods, gaze, and/or arm gestures [1, 27, 37, 54]. These cues may convey information regarding Understanding (whether the listener understands the utterance of the speaker) [1, 27], Attentiveness (whether the listener is attentive to the speech of the speaker) [1, 27], and Affect [27, 35]. Listener Responses are generally simultaneously expressed by vocal/verbal and by gestural (including facial expressions) means [1].

Listener Responses are a special case of *Cooperative* (multimodal) utterances—i.e., they are not intended to cause an interruption. In contrast, many of the functions mentioned above may also be served through *Competitive* (interruptive) utterances. The distinction between Cooperative and Competitive may be expressed in acoustic cues carried by the speech signal [18], or by the gestural/facial aspect of the utterance. The distinction whether incoming speech from the listener is Cooperative or Competitive is very important for determining how the speaker should deal with this incoming multimodal utterance.

3.2 Listener response elicitation

Speakers may also explicitly encourage the listener to provide Listener Responses. The speaker creates opportunities for Listener Reponses through vocal and non-vocal cues, such as pausing between statements, modifying the prosody of the speech, using gaze and face expressions and syntactic information [14, 21]. *Prosodic elicitation cues* for Listener Responses are quite well described in literature. Gravano et al. [23] observe that the final intonation of the interpausal unit (IPU) preceding a Listener Response rises in 81% of the cases, and the mean intensity and pitch level are higher than for IPUs not followed by a Listener Response. Ward et al. [55] use, in their handcrafted rule based model, a period of 110 ms of low pitch to predict a Listener Response 700 ms after this cue. *Nonverbal cues* are far less concretely described in literature. Such work mostly concerns gaze behavior. In a detailed study, Bavelas et al. [4] conclude that 83% of Listener Responses in their corpus occur during mutual gaze, confirming earlier intuitions of Kendon [28] and Duncan Jr. [15]. Head movements also have been associated with eliciting Listener Responses [26]. According to Duncan [14], using multiple elicitation cues increases the probability of a Listener Response occurring. In the experiments discussed at the end of this paper, we will use both prosodic and nonverbal elicitation cues in order to encourage the user to provide Listener Responses to the VH.

3.3 The attentive speaker

An attentive speaker pays attention to the listener. He moderates his speech and tailors it to reactions from the listener. Active listeners are not merely listening, but are co-narrating along with the speaker [3]. An attentive VH should be able to do both as well.

Clark and Krych [12] identify several ways in which speakers adapt their speech based on opportunities that arise, intentionally or not, mid-sentence. They claim that speakers make the adaptations almost instantly, typically initiating them within half a second of the opportunity arising. Self-interruption (see Example 1) is an example of such coordination with the listener. If the listener provides a reaction in mid-utterance which makes another utterance more relevant at the time (for instance, because the listener signals non-understanding and an elaboration is needed), the speaker cuts off his utterance and starts a new one.

Interaction Example 1 Self-interruption

Speaker: So starting from the square, you go…
Listener: euhm? *(indicates non-understanding)*
Speaker: I mean the square with the obelisk on it.

There are many ways for a speaker to deal with Listener Responses and other incoming multimodal utterances, dependent on the characteristics of the incoming utterance.

In Goodwin's observations, a speaker does not change the *content* of what he says based on the responses from the listener, but rather coordinates the *timing* of his speech influenced by the listener's responses [22]. Listener Responses are frequently found in complete overlap but also occur in partial overlap and silence. Goodwin states that the overlap strategy employed by the speaker depends on whether the listener feedback was a continuer or an assessment. Continuers simply acknowledge the receipt of the talk just heard and signals the speaker to continue speaking. Assessments are the result of an analysis of the speaker's talk by the listener based on which the listener has produced an action that is responsive to the particulars of the talk. If the speaker recognizes an assessment and is about to start a new unit, he *delays* this unit (e.g. by an inhalation or production of a filler) until the listener has completed his assessment. However, the speaker may deal with continuers by resuming speech *before the listener response is actually finished*, in effect letting continuers occur in partial overlap with the speech resumption. This is corroborated by recent research which has shown that only 41–45% of all turn-shifts occur after a

"minimally perceivable pause"; the remainder exhibit a certain amount of overlap [25]. Thus, interlocutors commonly continue to speak or resume their speech even before the listener has finished his/her response. The importance of this is suggested by Goodwin as follows:

> *... moving to a new turn-constructional unit while the recipient's "uhhuh" is still in progress is a proper and appropriate thing for a speaker to do. Indeed this is perhaps the clearest structural way for a speaker to demonstrate that recipient's action has been understood precisely as a continuer, and to act upon that understanding* [22].

The above is merely a selection of situations and strategies in which the speaker moderates his speech to the responses of the listener. For example, also nonverbal signals are dealt with by the speaker. Goodwin [21] showed that speakers are highly sensitive to listeners gaze. If they start a sentence and discover the listener is not looking at them, they restart (and often rephrase) when the listener looks back. There are many more situations, which we did not cover, but they illustrate the type of coordination we are ultimately aiming to achieve with our system. It is our long term aim to build a VH that is technically capable of achieving the same level of continuous interaction with the user as illustrated by these examples. The ability to deal with responses as illustrated above would allow for a VH to be highly responsive and manage the fragmentary nature of spontaneous dialog—a prerequisite for continuous interaction.

4 Analysis of listener responses in human-human interaction

To obtain more information about the exact content and timing of Listener Responses, we have analyzed a corpus of recorded human-human interactions. We are interested in the discriminating features of Listener Responses, other Cooperative utterances, and Competitive utterances. The results of the analysis are to be used in the design of classifiers distinguishing between the various types of utterances (see Sect. 5).

4.1 The HCRC map task corpus

The HCRC Map Task Corpus [2] is a well-known speech corpus consisting of 128 dialogues. The task of the participants in the dialogues was for one subject to explain a route on a map to another subject. Both subjects had their own copy of the map. The one who explained the route is denoted as the "giver" and the one who received the explanation as the "follower". Half of the dialogs were recorded under a face-to-face condition and the other half under a non-visible condition. We used the dialogs from the face-to-face condition since it is closer to our scenario of an interaction with a Virtual Human.[3]

4.1.1 Segmentation

The official (manual) segmentation of the Map Task corpus is based on the dialogue annotation. Annotators were asked to identify *dialogue moves*[4] in the transcripts and label them with the type of contribution. Each dialogue move leads to exactly one Map Task segment. The segmentation, thus, results from interpretation of the speech content. The classifiers that will be developed on the basis of the corpus analysis (see Sect. 5) are intended to discriminate between Listener Responses, other Cooperative utterances, and Competitive utterances. This distinction should be made before an interpretation of the speech content is available. The segmentation that will in practice be accessible to these classifiers will more likely be based upon an on-line voice activity detector. Therefore, we need a corpus segmentation that better resembles the results of an on-line voice activity detector.

For our analyses and experiments we derived—from the Map Task segments—a segmentation into perceptually relevant *talkspurt* segments. The operationalized procedure closely follows [7] who used the term *talkspurt* for the resulting segments (also referred to as Inter Pausal Units in later literature). By treating the Map Task segments as on-off or speech-silence patterns (extra-linguistic sounds are treated as silence), any speech segment shorter than a minimum voice activity duration threshold $\alpha = 50$ ms are set to silence, and any silence segment shorter than an inter-pause duration threshold $\beta = 200$ ms are set to speech. The latter threshold β is approximately equal to the minimum perceptible pause duration [53] for humans. Thus, the talkspurt segmentation gives perceptually relevant segments and the results will better resemble the conditions when an on-line voice activity detector is used, which is typically energy-based with the same duration thresholds. When the derived talkspurt is comprised of more than one Map Task segment, the talkspurt is labeled with the label from the first dialog move included in the talkspurt. In 3.16% of the cases, the merging procedure created talkspurts which started as a ACK and ended as a NONACK (see next subsection). The occurrence of these latter talkspurts are considered to be negligible.

[3] The two dialogs labeled as `q3ec1` and `q3ec5` were discarded due to a buzz in the speech signal.

[4] Anderson et al. [2] structure a dialogue into three levels: *transactions*, that accomplish a major subtask in the dialogue such as getting from one waypoint to the next; *conversational games* that fulfill a purpose within the transactions such as getting a question answered, getting something clarified, consisting of initiations followed by responses; and *dialogue moves*, which are the various types of initiations and responses that make up a conversational game.

To summarize: our talkspurt segmentation—derived from the official corpus annotations—offers several advantages: (1) The resulting segments are perceptually relevant; (2) The dialog move annotations can be reused; (3) Since the segmentation assumes an *ideal* Voice Activity Detector (VAD), the evaluation of proposed technology can be made independent of the efficiency of VAD which allows for separation of this factor and subsequent experiments can then be made to evaluate the integrated system, given different VAD implementations; (4) Talkspurts segmentation allows for a highly reproducible analysis of conversational phenomena without relying on interpretative definitions of a phrase or a turn which are subject to discussion.

4.1.2 Acknowledgement annotations

The first distinction that we want to get from the official Map Task annotations is between talkspurts that are Listener Responses versus other talkspurts from the listener ('follower'). In the Map Task annotations this is best captured by the distinction between Acknowledgment Moves (ACK) and other dialog moves (NONACK). The precise definition of an Acknowledgment Move is found in [8], which closely resembles the term Listener Response and thus serves our purpose. It is described as '*a verbal response that minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and accepted*'. The reliability of these annotations was considered good, with an inter-annotator agreement of $\kappa = 0.83$.

4.1.3 Cooperative/competitive annotations

The second distinction for which we need annotations, is between talkspurts that intend to take the floor (COMPETITIVE) or not (COOPERATIVE). As this information was not yet available in the Map Task corpus, we annotated part of the data with these labels. The following talkspurts were annotated:

- We only annotated NONACKs, as ACKs are supposed to be COOPERATIVE by definition.
- We annotated only talkspurts in overlap (Listener's talkspurt starts between the start and the end of the Speaker's talkspurt) because the COOPERATIVE/COMPETITIVE dimension only makes sense for overlapping talkspurts.
- We only annotated NONACKs, which do not have any ACKs within the local overlap. For example, a NONACK which is intercepted in overlap by ACK is excluded.

In the data that we used, there are 1232 candidate talkspurts to be annotated. Of these, the 524 talkspurts belonging to the first 32 dialogues were labelled by two annotators. The confusion table and reliability values are given in Table 1. The level of agreement for this annotation is in the range of highly subjective annotations [41]; the annotators agree on a certain amount of talkspurts being COOPERATIVE, but have difficulty agreeing on which talkspurts are COMPETITIVE.

Table 1 Contingency matrix for the annotator A1 and A2 of overlapping talkspurts on Competitiveness. Cohen's $\kappa = 0.45$ ($p < 0.01$), maximum $\kappa = 0.83$, proportion of maximum $\kappa = 0.54$; Krippendorff's $\alpha = 0.45$

	A1 COMPETITIVE	A1 COOPERATIVE
A2 COMPETITIVE	88	77
A2 COOPERATIVE	40	319

Table 2 Top 20 most frequently occurring tokens of the Acknowledgment Moves (ACK) found in the Map Task corpus, accounting for 7313 out of 9823 of these tokens

Count	Word
2773	right
1459	okay
525	mmhmm
521	uh-huh
380	yeah
264	oh
227	the
153	that's
145	no
133	i
93	got
89	it
86	you
82	that
73	mm
66	a
65	to
63	fine
58	i've
58	aye

4.2 Analysis of listener responses in the map task corpus

4.2.1 ACK content and overlap

In previous studies, cooperative Listener Responses have been shown to be short, and it is suggested they may be detected by duration alone [16]. This also holds for the ACKs in the Map Task corpus: Fig. 1 shows the duration of ACKs vs. the other dialog moves; Table 2 shows that the word content for ACK talkspurts typically consists of a single short word.

Listener Responses have also been frequently found in overlapped speech [22, 44]. Given a 10 ms frame discretiza-

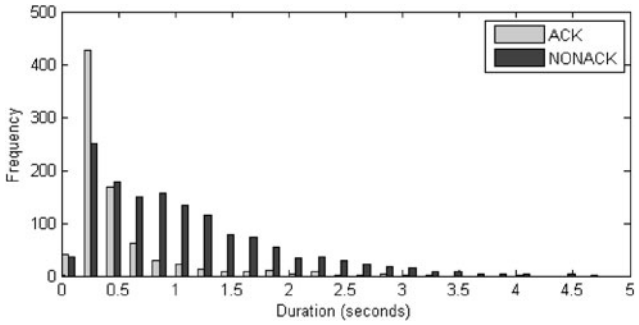


Fig. 1 Duration of ACKs vs. duration of other dialog moves, using bins of 200 msec

tion of the Map Task talkspurts, the following can be observed:

- Given a speech frame in overlap, there is a 34.9% probability that it is an ACK.
- Given a speech frame in non-overlap, there is a 5.2% probability that it is an ACK.

Thus, ACKs are relatively more common in overlap than in non-overlapped speech.

4.2.2 *Between speaker intervals following* ACK *talkspurts*

The listener may produce an ACK talkspurt in complete overlap (i.e., the ACK ends before the speaker's talkspurt to which it is a reaction ends), or the ACK may extend beyond the speaker's talkspurt. In the latter case, the speaker may resume speech *before* the ACK is finished (leading to partial overlap), or the speaker may wait (leading to a gap). Figure 2 shows these three situations.

In this section, we look at *between speaker intervals following* ACK *talkspurts*, defined as the duration between the end of the ACK talkspurt and the beginning of the talkspurt with which the speaker resumes speech. The between speaker interval can be positive (gap) or negative (partial overlap).

First, we consider the two cases of overlap: ACK in complete overlap and the ACK in partial overlap (see Fig. 2; for clarity, the gap is also illustrated). In the case of complete overlap, the attentive speaker must be able to detect the incoming ACK talkspurt as being Cooperative. This detection must preferably happen before or slightly after the peak (or mode) of the overlap duration for Competitive speech, which is given later. The case of partial overlap following a ACK (or: negative between speaker intervals), is what Goodwin suggests to be "a proper and appropriate thing for a speaker to do" [22]. We hereby ask the question: to what extent do speakers actually do this? In other words, is the partial overlap case is more common in ACK context than for no particular context?

Thus, we computed the between speaker intervals for the partial overlap case and the no overlap case, both for all

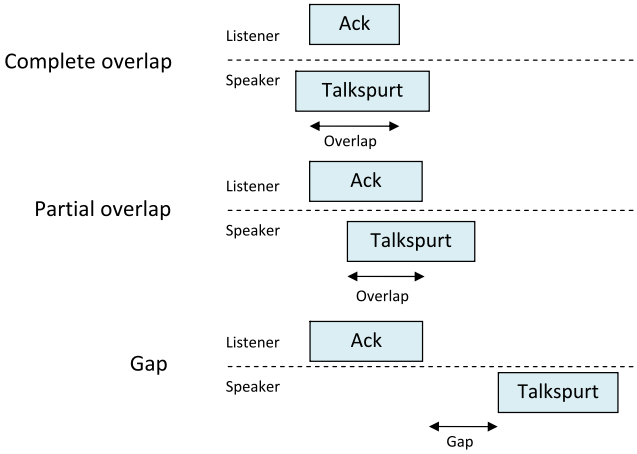


Fig. 2 Complete overlap, partial overlap and no overlap in the context of ACK

speaker changes and for only those that occur in the vicinity of ACK (the latter case includes all gaps before and after ACK, and all partial overlaps with ACK). In addition, the gaps and overlaps following an ACK interjection into silence are computed. This measures the degree of overlap after the speaker resumes his/her speech after an ACK. To facilitate comparison with other work two issues are considered. First, the tails are cut at 2000 ms. Secondly, while our default segmentation of talkspurts in Sect. 4.1 excludes extralinguistic sounds, we add computation of between speaker intervals for all speaker changes with extralinguistic sounds included. These measurements extend and correct the measurements in [38].

The distributions are shown in Fig. 3. The figure shows that the mode, the actual peak of the distribution, is at 100-200 ms for all distributions. First, it is observed that speaker shifts for talkspurts including extralinguistic sounds show a higher degree of overlap compared to the talkspurts that exclude these. The speaker changes in the vicinity of ACK have a higher proportion of smooth shifts, i.e. between 0–400 ms. The cumulative distributions are given in Fig. 4. It show that the proportion of speaker changes up to 200 ms for talkspurts including extralinguistic sounds are 54%. This latter proportion is close to the 57% which is reported for the same corpus in [25] where the VAD used to obtain segmentation is likely to include extralinguistic sounds. However, when extralinguistic sound are excluded the proportion up to a 200 ms gap is 37%, which is lower but close to the 35% reported by [40]. For the same case, the proportion of all speaker shifts in overlap is 20% while the same proportion in the vicinity of ACK is 19%, increasing to 37% by including a 200 ms gap. Our main measure of interest are the gaps and overlaps following an ACK interjection into silence. The proportion of resumptions in overlap are 24% while the proportion of up to a 200 ms gap is 41%.

The most striking difference found is the lower proportion of speaker shifts up a 200 ms gap when extra-linguistic sounds are excluded. This is not too surprising, since these sounds are often found in overlap. However, it also means that a much lower proportion of speaker changes than expected can be due to projection, i.e. end of utterance prediction in overlap from syntax and prosody carried by lexical items. The similar proportion of overlap without any particular context and in the vicinity of ACK has one direct implication. It means that the over-representation of ACK in overlap, as found in the previous section, is mostly due to ACK interjection into complete overlap, rather than partial overlap. Another implication concerns the theory of the different functions Listener Responses may fulfill. A typical distinction is made between back-channels as a type of Listener Responses which the interlocutor does not wait for [13], as opposed to acknowledgments and assessments which the interlocutor waits for, since they incorporate evaluation of what the speaker has said. Since the proportions of speaker changes without any particular context in the vicinity of ACK are the same in overlap and up to a 200 ms gap, it suggests that turn-taking is not different for Listener Responses except for situations of complete overlap. On the other hand, the proportion of interlocutor resumptions up to a 200 ms gap after ACK interjection into silence is 41%. Since ACKs are short, there is not much time to grasp the signaled meaning and to react by resuming one's speech before a perceptible pause. Thus, it is reasonable to assume that around 41% of ACK interjected into silence are back-channels, rather than acknowledgments or assessments. The proposed method may offer a computational but possibly crude way of distinguish between Listener Responses that carry meaning as opposed to the ones that do not, assuming a reasonable collaborative interaction between the listener and the speaker. However, the efficiency of the method remains to be evaluated. The same reasoning also leads to design implications for a VH. Since most ACKs have a duration up to 500 ms, incoming speech has to be detected as ACK or not before these are finished, i.e. preferably before 500 ms. Such a design allows the VH to resume its speech while the listener is still uttering a listener response, as humans do 41% of the time.

4.2.3 *Duration of* COMPETITIVE *and* COOPERATIVE *responses*

Figures 5 and 6 shows the distribution of the duration of COMPETITIVE and COOPERATIVE Responses, and of the durations of the *overlap* for both types of Responses.

Firstly, we notice that the two overlap distributions are different. Short overlaps around 100 ms are more likely for COOPERATIVE Responses rather than for COMPETITIVE Responses. The most likely overlap duration for COOPERATIVE Responses is around 100 ms, and around 95% of these talkspurts have an overlap duration up to 700 ms. The most likely overlap duration for COMPETITIVE Responses is around 300 ms, and around 95% of these talkspurts have an overlap duration up to 1100 ms.

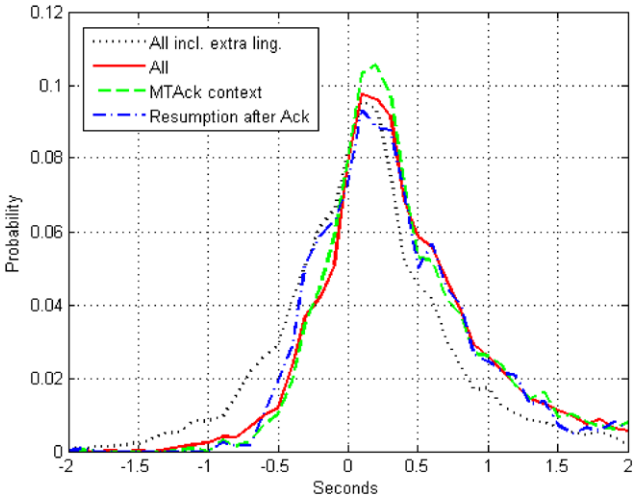


Fig. 3 Probability mass functions for between speaker intervals under different constraints using bins of 100 ms

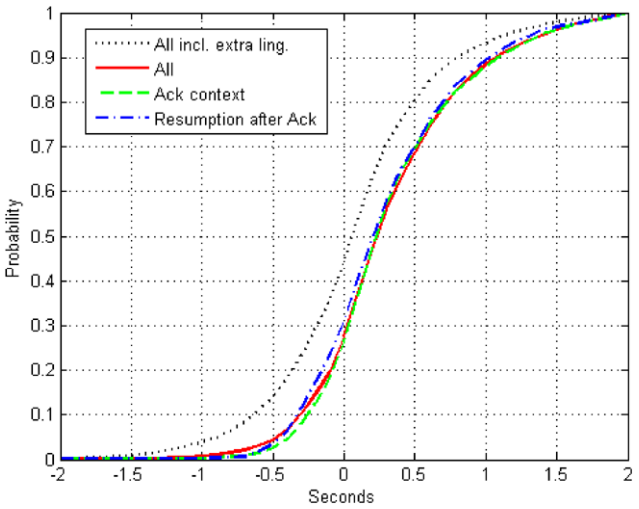


Fig. 4 The cumulative distribution for between speaker intervals under different constraints ACK Response using bins of 100 ms

Secondly, we notice that the two talkspurt duration distributions are different. We observe that COOPERATIVE talkspurts tend to be shorter, peaking in 250 ms, than talkspurts for COMPETITIVE speech which peak at 1750 ms. This means that duration may be used as a feature for competitiveness, but still the decision to stop talking when incoming speech are observed in overlap, is constrained by the observed durations of overlap explained in the previous paragraph. Thus, there is a trade-off between these two constraints, the different durations of talkspurt and overlap: the earlier you want to respond the harder it is to use duration as a feature.

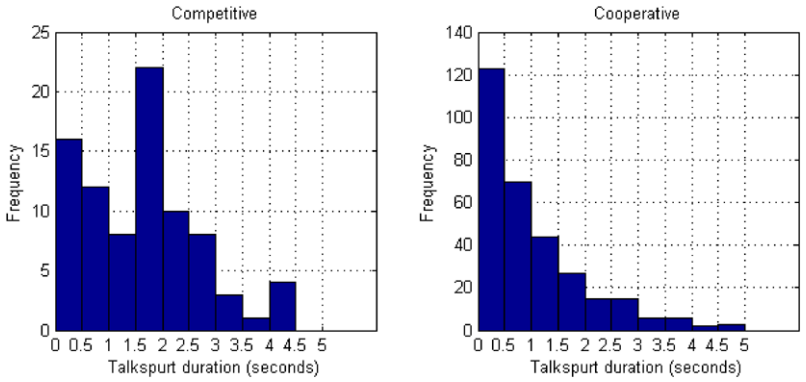


Fig. 5 Durations of talkspurts in overlap with no ACK context (within the overlap) using bins of 500 ms. To the *left* are COMPETITIVE and to the *right* COOPERATIVE Responses

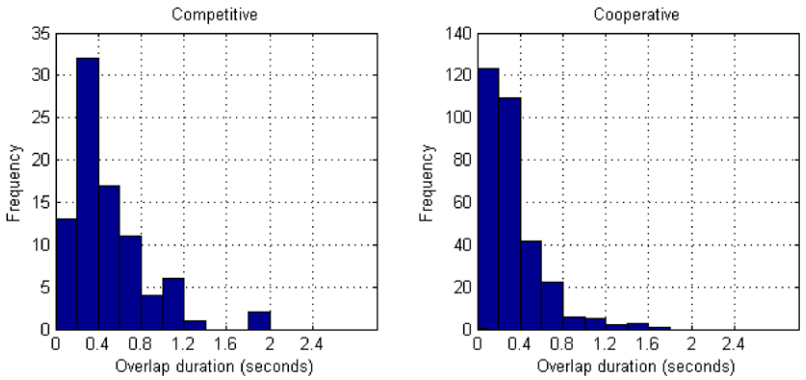


Fig. 6 Durations of overlaps with no ACK context (within the overlap) using bins of 200 ms. To the *left* are COMPETITIVE and to the *right* COOPERATIVE Responses

4.3 Design implications for an attentive speaking VH

For a responsive dialog with a VH, multimodal talkspurts from the user need to be classified into COMPETITIVE/COOPERATIVE before they are finished. The analysis presented here provides us with the following timing constraints on a classifier: (1) (Cooperative) Listener Responses have be detected within 100–300 ms of the onset in overlap (within the minimally perceivable pause duration), and (2) within 100–500 ms of the onset in silence; furthermore, (3) incoming Competitive talkspurts have to be detected within 300–1100 ms of the onset in overlap. The overall duration of a utterance from the listener can potentially be used to as a feature for a COMPETITIVE/COOPERATIVE classifier. We have designed a classifier that adheres to the constraints posed here (see Sect. 5) and that uses (among others) the duration feature proposed above. The annotated Map Task corpus is used as a training and testing set for these classifiers.

5 Classification of Listener Responses

To allow for continuous interaction to occur between humans and VHs, we require detectors that are capable of aiding turn-taking for the turn-shifts that occur before the *minimally perceptible pause* is over. This is achieved by classifying the listener's talkspurts in overlap as being COOPERATIVE or COMPETITIVE before the listener has finished speaking. Since Listener Responses are COOPERATIVE (though not all COOPERATIVE talkspurts are Listener Responses), the first step is to be able to detect these. This sub-task of detecting Listener Responses is carried out regardless of whether incoming talkspurts are in overlap or not. The design must also follow the constraints provided by the analysis in Sect. 4.3 in terms of guaranteeing decisions before certain durations thresholds. This could be done using a speech recognizer running in incremental mode or by using a specialized detector. Since a speech recognizer will only detect lexical content, the special prosodic characteristics of vocal listener responses cannot be accounted for. In addition, automatic speech recognizers (ASR) frequently miss Listener Responses in spontaneous speech [20]. Hence, we developed a specialized detector, the overall cascaded design of which is shown in Fig. 7.

In summary, this leads to two classification tasks.

- **Classifier I** Classification of all Responses into ACK/NONACK, within 100–500 ms of the onset of speech, for which we here give an outline as more details are available in [38].
- **Classifier II** Classification of NONACK, produced in overlap, into COOPERATIVE/COMPETITIVE within 300–1100 ms of the onset of speech.

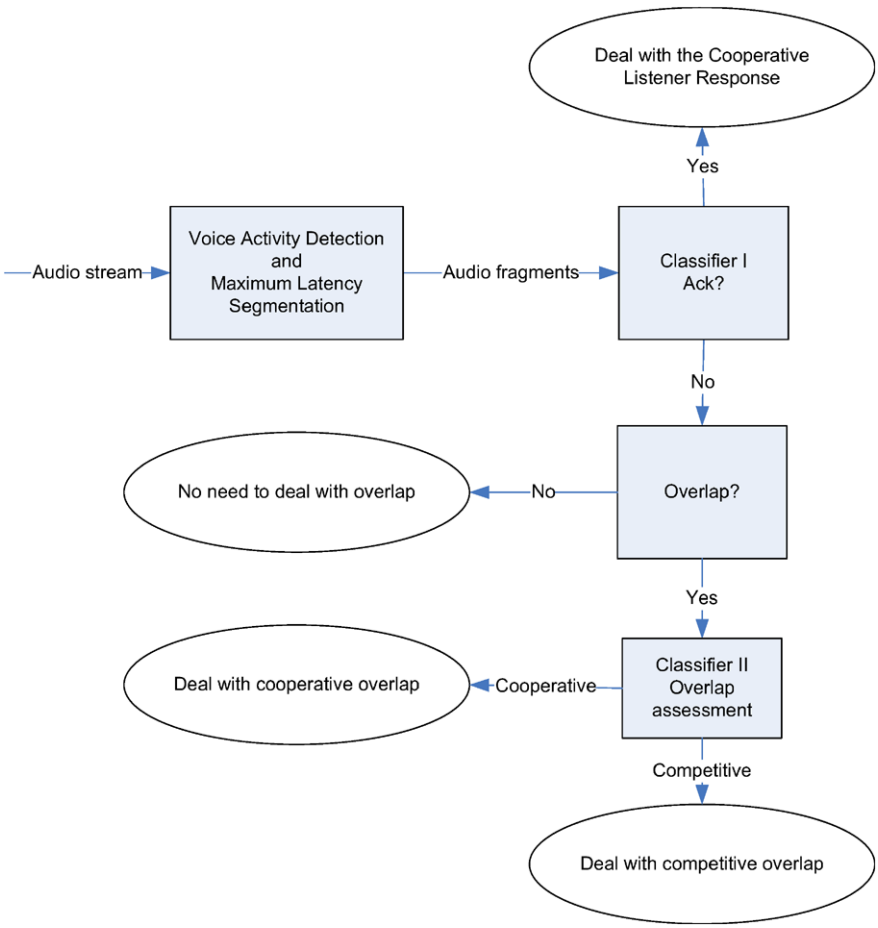


Fig. 7 Cascade used to classify incoming Responses from the user

5.1 Maximum latency classification

The duration constraints for making a decision needs to be incorporated at the fundamental level of the design. Thus, we propose a maximum latency implementation, which is illustrated in Fig. 8. It is implemented as a voice activity detector which sends an end message after the talkspurt ends, or at a predefined duration threshold τ. If the duration reaches the threshold, it continues to work as a normal voice activity detector internally, otherwise it might trigger again. Note that the detector may trigger before the maximum latency threshold is reached which happens when the talkspurt is shorter than the threshold subtracted by the minimum inter pause threshold β. For on-line detection, this maximum latency design was implemented in openSMILE [17].

5.2 General design of detectors

All classifiers use Support Vector Machines (SVM) with Radial Basis Function Kernel as implemented in *LIBSVM* [9]. The SVM regularization parameters are optimized on the development set, and the best parameters are then used for test on the evaluation set. The acoustic features were extracted at a 10 ms frame rate by using openSMILE [17].

To parameterize the trajectories throughout a talkspurt of each feature, we use Discrete Cosine Transform (DCT) coefficients invariant to segment length. These are calculated as a type II DCT divided by the number of frames. This means that coefficients are not affected by stretching in time. This property is useful for the maximum latency segmentation, which creates varying length talkspurts, only limited by a maximum duration. The three most important advantages for using this time-varying parameterization are:

- The DCT basis functions are periodic, which allows good interpolation of syllabic rhythm in speech.
- The length-invariance gives a normalization for duration or speaking rate. If duration or speaking rate is added to the final feature vector, then the machine learning algorithm can determine whether it is a salient cue or just speaker variation.
- The 0'th coefficient is equal to the arithmetic average, which means if it is omitted, then only the relative shape of a trajectory is parametrized. This property is useful for parameterizing features such as F0 (which has a speaker dependent additive bias) or MFCCs (which has an additive channel bias). Although MFCCs has been found to contain speaker dependent elements, speaker normaliza-

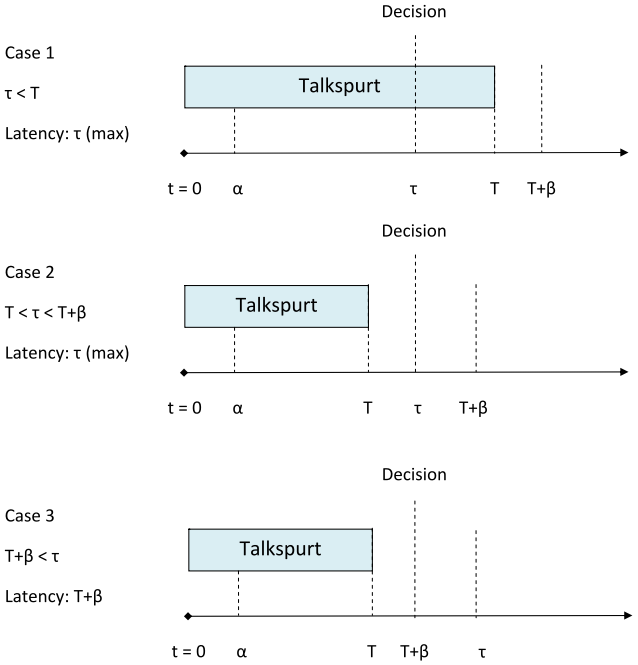


Fig. 8 The figure illustrates maximum latency talkspurt segmentation. T is the talkspurt duration, α is the minimum speech activity threshold, β is the intra-pause duration threshold and τ is the maximum latency duration threshold

tion is usually achieved by affine transformation which is computationally and conceptually more complicated. The affine transformation includes an additive bias, so the proposed parametrization offers a crude speaker normalization by omitting the 0'th coefficient.

To ensure independence of priors and application, the performance is measured as Equal Error Rate (EER) calculated using the SVM decision values. This prior independence allows for comparing results across corpora. At a later stage, when a classifier would be fielded for particular task, then the decision threshold might be adjusted according to the priors or design specifications for the application.

5.3 Classifier I: ACK vs. other dialog moves

5.3.1 Features

For the task of classifying Responses as ACK or not, the combined set of acoustic features was comprised of:

- F0 Envelopes: Back-channels have been shown to have a rise or drop in F0 [5, 24].
- Intensity: Back-channels have been shown to have distinct intensity contours [5].
- MFCC: Distinct lexical content, see Table 2, can be captured by Mel Frequency Cepstral Coefficients (MFCC) which also measures spectral shape and formant trajectories.

Table 3 "ACK vs. other dialog moves" classification task: EER in percent for the evaluation set

Max. lat. τ (ms)	100	300	500
EER	31.7	29.5	26.2

- Duration: As seen in Fig. 1, ACKs have shorter duration then other types of dialog moves. For training, the full talkspurt duration was used, for testing, the duration up to the maximum latency threshold was used.
- Spectral Flux: Common listener responses such as "mmhmm" and "uh-huh" are relatively homogeneous throughout their realization, and spectral flux should capture this property. The spectral flux is computed as the L2-norm of energy normalized FTT-bin difference between two adjacent frames.

All features are parametrized using length invariant DCT-coefficients 1–6 except Spectral Flux for which we use coefficients 0–5, since it is already a delta-type of feature, and for duration the arithmetic average (0th coefficient) is used.

5.3.2 Experimental setup

The set of 64 face-to-face dialogs from the HCRC MapTask are officially divided into 8 subsets called quads. For all experiments, the training set consists of so-called quads 1–4, the development set holds quads 5–6 and the evaluation set holds quads 7–8. Based on the analysis of overlap durations in Sect. 4.3, a maximum latency threshold, τ, of 100 ms, 300 ms or 500 ms is desirable for this task.

5.3.3 Results and discussion

The experiments on the development set showed that MFCCs and duration, at least in the 500 ms case, are the main contributors to the distinction between ACK vs. NONACK, while F0 is the weakest feature. This led us to omit the F0 feature in the combined feature set. The results for unseen data in the evaluation set given the feature combination are shown in Table 3. We observe that a higher maximum latency threshold yields better performance, but the trend is not as strong as expected. It should be noted that when the talkspurt segmentation is obtained by a energy based voice activity detector, a drop of approximately 4% should be expected [38].

5.4 Classifier II: COMPETITIVE vs. COOPERATIVE

This task is based on the distinction between COMPETITIVE and COOPERATIVE Responses in overlap. The classifier was trained on agreed annotations made by two human annotators who labeled a part of the HCRC Map Task Corpus on perceived COMPETITIVENESS and COOPERATIVENESS for a subset of overlapped talkspurts (as explained in Sect. 4.1).

5.4.1 Features

Choosing a good acoustic feature set for this task is not easy since only a few studies are available. Intensity is the most widely studied cue for interruption [18, 33]. Speaking rate has been studied in [44] where it was noted that COMPETITIVE overlappers make use of higher speaking rates. However, Kurtic et al. [32] found speaking rate to be a weak cue for COMPETITIVE Responses. Speaking rate is very difficult to estimate for segments lasting less than 1000 ms. Instead, we try spectral flux which has been used for estimating tempo in music [36]. While average F0 (high) has shown to be a cue for interruption (e.g., [18]), it requires adaptive estimation of F0 range and is not considered here. Instead we rely on the relative shape of the F0 trajectory. As shown in the analysis in Sect. 4.2.3, talkspurt duration is a good feature. However, given the proposed framework, only durations shorter than the maximum latency threshold subtracted by the minimum pause duration threshold will hold information. Based on the experience from annotation, we noted a tension in the voice for some COMPETITIVE Responses. Thus, voice quality correlates may be useful for this task. Voice quality was measured by spectral centroid, spectral kurtosis, and spectral skewness. The combined acoustic feature set was comprised of:

- F0 Envelopes.
- Intensity.
- Duration: For training, the full talkspurt duration was used. For testing, the duration up to the maximum latency threshold was used.
- Spectral Flux.
- Voice quality As measured by spectral centroid, spectral kurtosis and spectral skewness.

Thus, the feature set is identical to the set described in Sect. 5.3.1 except for the lack of MFCCs which was hard to justify for this task, and the addition of voice quality correlates. All features are parametrized using length invariant DCT-coefficients 1–6 except Spectral Flux, spectral centroid, spectral kurtosis, spectral skewness and duration for which we use the arithmetic average (0th coefficient).

5.4.2 Experimental setup

For this experiment, the set-up diverges from the set-up described in Sect. 5.2. For training and testing the classifier, we used the COMPETITIVE and COOPERATIVE annotations that were obtained with two human annotators (see Sect. 4.1). Only those talkspurts which had labels agreed upon by both annotators were used, in total 88 and 319 talkspurts for the COMPETITIVE and COOPERATIVE class respectively. Since we have relatively little data, an N-fold cross-validation scheme was applied for training and testing the classifier. There were 4 quads available. To ensure strict separation of training, development and testing sets, in each fold, 2 quads were used for training, 1 quad for optimization and 1 quad for evaluation. All possible combinations of quads with strict separation of training, development, and testing sets were made which yielded a total of 12 folds. When the optimal parameters were found, the training and optimization set were merged and used for training. This procedure allowed better use of the data, especially the sparse occurrence of the COMPETITIVE class. As pointed out earlier, the desirable choice for the maximum latency thresholds starts at 300 ms, adding 200 ms in steps until 1100 ms.

Table 4 Prediction performance of COMP vs. COOP on the evaluation set

Max. lat. τ (ms)	300	500	700	900	1100
EER	33.6	43.0	38.8	37.2	36.3

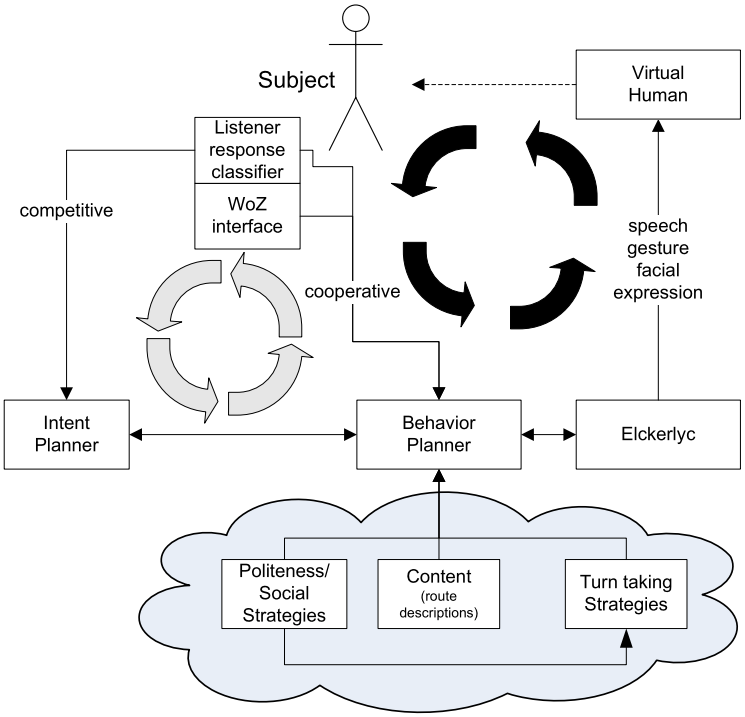
5.4.3 Results and discussion

The results for the classification experiment on the evaluation set are shown in Table 4. Contradictory to expected, the best performance was found at a maximum latency of 300 ms where Duration gives the lowest contribution. However, as found in Sect. 4.2.3 the overlap duration of COMPETITIVE Responses peaks at 200–400 ms while the overlap duration of COOPERATIVE Responses peaks at 0–200 ms. This indicates that the acoustic features are most salient at these maximum latency thresholds, otherwise humans would not be able to react accordingly. The performance is not as strong as for classifier I, but previous studies have shown the difficulty for this task [32, 33] and data sparseness is also an issue.

5.5 Conclusions from Classification Experiments

These experiments have shown that it is actually possible to classify incoming speech from the Listener as COOPERATIVE or COMPETITIVE before the Listener has finished talking. This allows us to mimic observed human-human behavior in terms of the duration of overlap and responsiveness in a VH. Specifically, it is possible to detect Listener Responses (a special case of COOPERATIVE Responses) with EERs of 32–26% guaranteeing a decision before 100–500 ms, by adjusting the maximum latency thresholds. For this task, the trade-off between latency and performance is lower than expected, and the success allowed us to implement an on-line version of this classifier. When Listener Responses are excluded, the task of classifying incoming speech as COOPERATIVE or COMPETITIVE was harder. This task gave EERs of 33-43% guaranteeing a decision before 300—1100 ms. By connecting these classifiers into a

Fig. 9 System architecture



cascade, it is possible to detect incoming speech in overlap as being COOPERATIVE or COMPETITIVE, and incoming speech during silence as being a listener response or not. Finally, it should be noted that all these classifiers may run in parallel for different maximum latency thresholds. Then different decision thresholds may be applied for the more reliable classifiers. Since all these classifiers are binary, the decision threshold can be set by the means of a Receiver Operator Curve which gives the opportunity to trade false alarms to false accepts.

6 Behavior generation and specification for continuous interaction

With the clearer understanding of Listener Responses, how to elicit them, how to detect them, and how to deal with incoming Listener Responses in an appropriate way, we have built an experimental setup of a VH that incorporates elements of Attentive Listening. The task of the VH is to explain a route on a map to the user, eliciting Listener Responses from the user. When the user provides these responses, the VH should, ideally, deal with them by adjusting its utterances on-the-fly (cf. Sect. 3.3). In this section we describe the global setup, which uses the BML Realizer Elckerlyc to generate the VH's behavior, and we introduce the improvements that we had to make to Elckerlyc in order to facilitate the required flexibility.

Figure 9 shows the different components that make up the system architecture of the VH. Communication between

Fig. 10 The SAIBA framework

the components is implemented using the SEMAINE framework [46], a middleware framework for transparent communication between distributed modules. The distinction between *communicative intent planning* for the VH, multimodal *behavior planning*, resulting in a Behavior Markup Language (BML) stream [31], and *behavior realization* of this stream, is based upon the SAIBA framework (see Fig. 10) [31].

In our setup, the Communicative Intent is fixed: the VH needs to explain a route to the user. The Behavior Planner component specifies the behavior that is used to express this Communicative Intent, including Response Elicitation behavior. The behavior is specified as a stream of BML blocks that is sent to the Elckerlyc BML Realizer [56] which executes this behavior through the embodiment of the VH. The Listener Responses that are elicited from the user are detected through the Listener Response classifiers, or, when the performance of the classifiers is not high enough for a robust conversation, through a Wizard of Oz setup. The exact method of handling Listener Responses (explained in more detail later) is influenced by turn-taking strategies and by the conversational content (a specification of the route to explain).

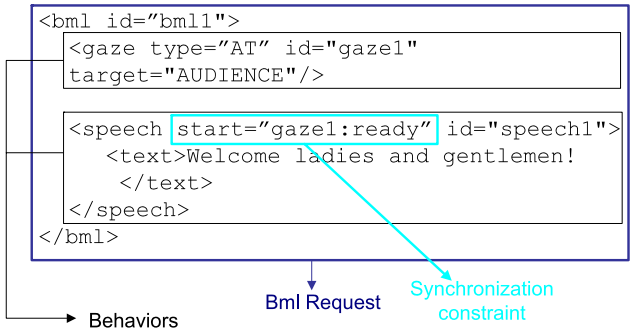


Fig. 11 An example of a BML request containing a gaze and a speech behavior. A synchronization constraint ensures that the speech starts after the gaze is aimed at the audience

6.1 Behavior markup language and Elckerlyc

The BML stream, sent from the Behavior Planner to Elckerlyc, contains BML requests with behaviors (such as speech, gesture, head movement, etc.) and specifies how these behaviors are synchronized to each other (see also Fig. 11). Synchronization of the behaviors to each other is done through BML constraints that link synchronization points in one behavior (start, end, stroke, etc.; see also Fig. 12) to synchronization points in another behavior. BML can be used to append or merge new behaviors into a running BML stream.[5]

In a continuous interaction setting, the behavior planner might require micro-adjustments to timing or to parameter values (speak louder, increase gesture amplitude, slightly delay the stroke of a gesture). Such small adaptations of the timing or shape of planned behavior occur in human conversations and other interactions [39]. Elsewhere, we discuss how Elckerlyc allows such small behavior plan changes to occur instantly [57].

Furthermore, continuous interaction requires mechanisms to allow graceful interruption and to specify an alternative follow-up to an interrupted behavior [57]. Currently, BML does not contain mechanisms to interrupt behavior in a graceful manner. To achieve continuous interaction, we have introduced interrupt behavior and preplanning mechanisms as a part of our BML extension BML^T.[6] The *combination* of this interruption and preplanning allows graceful interruption with an instantly activated follow-up.

In the remainder of this section, we discuss the extensions to BML that we used in our experimental setup to allow modification of the expression, and of the timing, of behaviors and the scheduling and interruption mechanisms discussed above.

[5]Some extensions have been proposed to allow the specification of instant removal of a running BML request (see http://wiki.mindmakers.org/projects:bml:multipleblockissue).

[6]Being developed at the University of Twente, the name of this extension may be read as *BML Twente*.

6.2 Preplanning

Scheduling a BML block typically takes a non-negligible amount of time, especially if the timing of speech is to be obtained through speech synthesis software. This is problematic for developing highly responsive virtual humans. BML^T provides *preplanning* as a mechanism to construct a behavior plan that can be activated later on. In a typical usage scenario of pre-planning, the Behavior Planner already knows what behavior to execute, and wants to execute it (near) instantly later on, for example in reaction to some event such as an incoming response from the user. Preplanning is set up for a BML block, using the BML^T preplan attribute in that block. Preplanned BML blocks can be activated using another BML block with an onStart attribute. The preplanned BML block is activated as soon the BML block containing a matching `onStart` starts its execution. Example 1 illustrates the BML used for preplanning.

BML Example 1 Several BML blocks illustrating the preplanning and activation of pre-planned behavior

```
<bml id="bml1" bmlt:preplan="true">
  ...
</bml>
```

(a) Preplan **`bml1`**.

```
<bml id="bmlX" bmlt:onStart="bml1"/>
```

(b) Activate preplanned behavior **`bml1`**.

```
<bml id="bml3" scheduling="append-after(bml2)"
    bmlt:onStart="bml1,bml5">
    ...
</bml>
```

(c) Schedule **`bml3`** to be appended after **`bml2`**, activate preplanned behaviors **`bml1`** and **`bml5`** as **`bml3`** is started.

6.3 Graceful interruption

The BML^T interrupt behavior provides us with the capability of specifying precisely when behaviors should end and what new behavior should be activated after a behavior is interrupted.

A simple example would be to start a "look-at" behavior by the VH, while it is speaking, and to interrupt the speech behavior as soon as the "look-at" behavior has finished.

In its simplest form (see BML Example 2) the BML^T interrupt behavior, as soon as it executes, interrupts a complete BML block, referred to as the "target". Interrupts are normal BML behaviors, so they have standard BML attributes like an `id` or `start` sync point, and can be synchronized with other behaviors as usual.

In a more refined version, the interrupt behavior is allowed to refer to specific behaviors inside the interrupted target block, to specify the exact moment where these behaviors are to be interrupted. A second refinement is that such

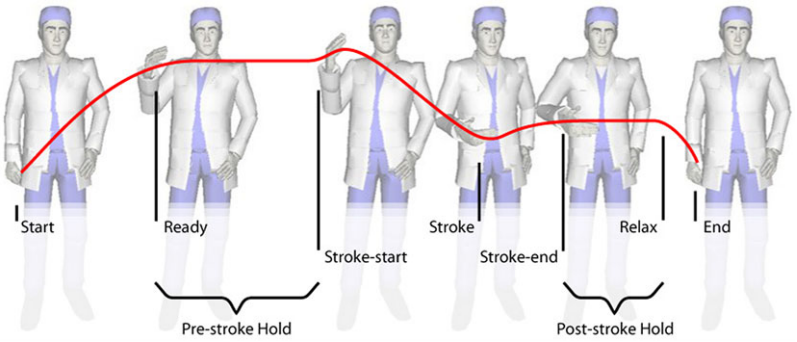


Fig. 12 Standard BML synchronization points (picture from http://wiki.mindmakers.org/projects:bml:main)

BML Example 2 Interrupt `bml1` as soon as `shake1:stroke` is reached

```
<bmlt:interrupt id="interrupt1"
                start="shake1:stroke"
                target="bml1">
</bmlt:interrupt>
```

interrupted behaviors can trigger other preplanned BML blocks, that will effectively replace the interrupted behavior. A typical example is that of a speech behavior that can be interrupted at certain predefined places. Upon being interrupted, the original speech behavior is then replaced by a short fragment of speech that gracefully terminates the interrupted behavior.

The syntactic element that enables this more refined interrupt is the `interruptspec` element inside an `interrupt` behavior, as shown below in BML Example 3. Here, the `interruptSync` attribute specifies the point where to interrupt, whereas the `onStart` attribute specifies the preplanned replacement behavior block. (All behaviors in the target of the interrupt that are not explicitly mentioned within an `interruptspec` element will be interrupted as usual, that is, the interrupt acts immediate, and there is no replacement behavior.)

So within BML Example 3, the `speech1` behavior from block `bml1` will be interrupted at synchronization point `sync1`, and will then be replaced by the behaviors from block `bml3`. The `gesture1` behavior from the same `bml1` block is interrupted at a different point, viz, at the `stroke_end` point, and will then be replaced by the `bml4` behavior.

The Smartbody Realizer [50] provides an interrupt behavior that has similar functionality as the simple (as in BML Example 2) form of our interrupt behavior.

6.4 Anticipators

Anticipators are a mechanism to specify in the BML stream that certain behavior of the VH should be aligned to external events in the real world. An Anticipator instantiates synchronization points that can be used in the BML stream

BML Example 3 The realizer interrupts all behaviors in `bml1`. `speech1` is interrupted at `sync1` and gracefully ended with some trailing speech using `bml3`, `gesture1` is interrupted at its `stroke-end`, and followed by the content of `bml4`. All other behaviors in `bml1` are interrupted at the start of `interrupt1` (that is, at `shake1:stroke`). Note that in many cases the alternative follow-up after an interruption (here specified in block `bml4`) can be derived automatically: a gesture interrupted before its stroke-start should be retracted before the stroke; a gesture interrupted during its stroke phase should complete the stroke before being interrupted, etcetera

```
<bmlt:interrupt id="interrupt1" target="bml1"
  start="shake1:stroke">
  <bmlt:interruptspec behavior="speech1"
    interruptSync="sync1" onStart="bml3"/>
  <bmlt:interruptspec behavior="gesture1"
    interruptSync="stroke_end" onStart="bml4"/>
</bmlt:interrupt>
```

to constrain the timing of behaviors. It uses perceptions of events in the real world to continuously update the actual timing of these synchronization points, by extrapolating the perceptions into *predictions* of the timing of future events. When the timing of the Anticipator synchronization points is updated, the timing of behavior of the VH that was synchronized to these points, is automatically changed as well. BML Example 4 shows how an Anticipator allows an elegant specification of a segment of speech to start immediately after a listener response.

6.5 Realization of vocal response elicitation cues

Elckerlyc allows the use of any Text-To-Speech system for the speech generation. For this project, we used and extended the MARY TTS platform [47]. The MARY TTS platform is an open-source, modular architecture for building text-to-speech systems, including unit-selection and Hidden Markov Model (HMM) based synthesis technologies [47, 48]. In this section, we describe the use of MARY framework to realize vocal response elicitation cues. Prosody

BML Example 4 In Experiment 2 (see also Sect. 7), we aim to elicit listener responses from a listener. If such responses occur, we would like, in that specific experiment, the VH to wait until the listener is finished speaking before continuing the VHs speech. Here we show how this could be expressed in BML. Speech is started once the a `speechStopAnticipator` indicates that the interlocutor has stopped speaking. Such a `speechStopAnticipator` could be an automatic detector, or, as in Experiment 2, it could be hooked up to a key press or release by a wizard in a Wizard of Oz setting. The anticipator allows us (1) to plan the speech beforehand so it is executed without planning delay and (2) to *specify* alignment of the VH's behavior to events outside the world of the VH

```
<bml id="bml1">
  <speech id="speech1"
  start="anticipators:speechStopAnticipator:stop">
    <text>Bla bla</text>
  </speech>
</bml>
```

MARYXML Example 1 An example of prosody specifications using MARY TTS. The speech text surrounded by the prosody tag is first generated using default prosody parameters (predicted from text). Subsequently, the prosody tag is applied: the first 10% of the speech is changed to a lower pitch; at the 80% mark the pitch should be 10% above default; and the fragment should end at a 5 semi-tones higher pitch than the default expectation

```
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
 xmlns:xsi="http://www.w3.org/2001/
            XMLSchema-instance"
 xmlns="http://mary.dfki.de/2002/MaryXML"
 xml:lang="en-US">
 <p>
  <prosody rate="fast"
     pitch="+10%"
     contour="(10%,low)(80%,+10%)(100%,+5st)">
   Welcome to the world of speech synthesis!
  </prosody>
  </p>
</maryxml>
```

modification techniques are the key to realize such cues. Traditionally, applications that required control over prosody used MBROLA diphone synthetic voices, even though these voices sound unnatural. Nowadays HMM-based voices, which can support prosody modification, are reaching high quality synthetic speech.

The most recent versions of the MARY TTS framework support reliable prosody modification using the 'prosody' element (see MARYXML Example 1). The 'prosody' element is well described in the W3C Speech Synthesis Markup Language (SSML) recommendations.[7] The different attributes in 'prosody' element such as 'rate', 'pitch' and 'contour' are used as specifications to modify predicted phone durations and pitch contour before passing them to the HMM synthesizer. Once such modifications are done according to given specifications, they are realized as normal with HMM-based synthesis strategies [6, 52].

In addition to XML based prosody tuning support, as part of this research, we also implemented a new parameter which can enable high intonational rise at the final part of the speech utterance. Whereas the 'prosody' element support is useful for manual tuning of the overall quality of the speech through prosody parameters, the feature that supports the final intonation rise serves as a prominent, vocal response elicitation cue.

7 Experiments with an attentively speaking virtual human

As a setting for our experiments we chose the route description domain. This domain was chosen because the well-structured nature of the message content affords clear opportunities for eliciting Listener Responses. This makes it easy to manipulate the behavior of the VH to display various response elicitation strategies. Also, it is fairly easy to define a few simple strategies for reacting appropriately to Listener Responses. For example, the VH could repeat certain elements of the explanation to get a point across, or skip a part depending on the reactions from the user.

The two experiments described in this section are a first step towards testing the complete setup of an Attentively Speaking VH. Before we can go deeper into monitoring and handling the listener responses it is important that our system is able to elicit these responses. The experiments in this section are aimed at exploring ways in which we can elicit listener responses and at collecting data of behavior displayed by users interacting with the system.

7.1 Nonverbal response elicitation behavior

One of the elements in the experiments described in this section are the response elicitation cues displayed by the VH. In Sect. 3.2 we described possible vocal cues. For the nonverbal cues, the literature offers little information, so we turned to the MultiLis corpus, in which a speaker explains a recipe or an animation movie. Details on the corpus, its setup, content, and purpose, can be found elsewhere [29]. Here we only remark that the speakers in the corpus often exhibit nonverbal response elicitation cues and that there are marked differences in the amount of responses that individual speakers were able to elicit.

[7] http://www.w3.org/TR/speech-synthesis/.

Fig. 13 The map used in the two experiments

7.2 Experiment 1

In human-human conversation the speaker often elicits listener responses. The speaker creates response opportunities by providing eliciting cues to the listener, such as pausing between statements, modifying the prosody of the speech and displaying various nonverbal behaviors, as discussed in Sect. 3.2. In this experiment we aim to recreate such signals on our VH, and to evaluate them to see which elicitation strategy elicits the most listener responses. Furthermore we assess each version of our VH on subjective measures related to conversational skill, rapport, and personality.

7.2.1 Task

The participants sit at a desk, facing a large screen on which the VH is displayed. During the experiment, the VH explains a route through a fictional city to the participant. The participant needs to listen to, and remember, the route. Afterwards, the participant is asked to draw the route on the map that was shortly presented to him before the start of the interaction.

7.2.2 Stimuli

The map contains the layout of a fictional city (see Fig. 13). Landmarks are highlighted on the map, such as a cathedral, a stadium, and bridges. With the map comes a legend explaining the terminology used by the VH to identify the landmarks. The current position of the participant is also shown on the map.

There are three different starting points, one for each of three different routes. Each route consists of n steps[8] that take the user to their final destination (e.g., "take the first street on the right" or "go past the cathedral to the end of the street"). For each step, a BML block specifies the verbal and non-verbal behavior that the VH uses to explain the step. The BML block specifies the speech, gestures and facial expressions to be performed by the VH, as explained in the previous section. The speech is synthesized using MARY TTS [47]. The speech is manually cleaned up using the prosody tags described earlier. We removed, where necessary, peculiarities in the synthesized speech, added some extra pauses, and changed the speech rate at a few places, to make the VH sound more natural. Aligned with the speech, motion captured gestures[9] are added to accompany the explanation of the route (e.g. pointing to the left or making an iconic gesture representing a landmark). The pause between the blocks is 1.5 s, which is based on the mean pause between statements in the MultiLis corpus.

These pauses between the blocks are the response opportunities where we explicitly elicit listener responses. For each route we created four versions, each with different response elicitation behavior. These four different behavior are:

- **Default**: No explicit elicitation behavior.
- **Vocal**: Rising pitch at the end of the step.
- **Nonverbal**: Emphasis head and face gestures, interruption of blinking and gaze away as conformation behavior.
- **Combined**: Combination of the Vocal and Nonverbal behavior.

In the *Default* version no explicit elicitation behavior is employed. This version was our baseline from which we created the three following versions, by changing the pitch contours, or adding extra behaviors according to strict rules.

In the *Vocal* version we modified the pitch of the speech. The modification were inspired by Gravano and Hirschberg [23]. In their analysis of the Columbia Games Corpus, which is a task-oriented corpus, comparable to our setup (as opposed to spontaneous dialogues), they concluded that, among other features, the rising of the pitch in the final 200 to 300 ms of speech is a response eliciting cue. We applied this finding to our synthesized speech in this version, by giving the last word of a step in the route a rising pitch contour.

In the *Nonverbal* version we added the nonverbal elicitation behavior found in the MultiLis Corpus [29] described in Sect. 7.1. More specifically, we choose one of the speakers and recreated his nonverbal response eliciting behavior. This speaker was chosen by looking at the top 5 speakers with the highest rate of elicited listener responses per minute and selecting the speaker where nonverbal cues were most prominently present (according to our perception). His eliciting behavior was the following. He emphasizes the last word in

[8] For Route 1 and 3, $n = 8$, for Route 2, $n = 7$.

[9] This motion capture data is publicly available through http://hmi.ewi.utwente.nl/mocapdb.

a sentence by accompanying it with a subtle head nod and short eyebrow raise. At the same time he stops blinking (he generally has a relatively high blinking rate, so this actually stands out and tries to establish mutual gaze with the listener. As soon as a listener response is given, he starts blinking again and averts his gaze to formulate his next sentence. This behavior is recreated in the nonverbal version.

In the *Combined* version we combine both the vocal and nonverbal behavior changes to the default version.

7.2.3 Methodology

We invited 9 participants (8 male, 1 female, aged between 25 and 54, all non-native English speakers) to interact with our route giving VH. Participants were told that the VH is able to perceive and react to short vocal and nonverbal listener responses (like nodding, saying *"Uh-huh"*, or *"Yes"*).

Before each interaction the user was presented with the map with the starting point of the route. This map was taken away before the interaction started. During the interaction, the route giving VH gave a route description to the user. It was the task of the user to remember the route and reproduce it on the map afterwards.

Each participant interacted three times with the route giving VH. During each interaction the VH explained a different route. Each route description was given with a different elicitation strategy. Every participant interacted with the *Default* and *Combined* VH and either the *Vocal* or the *Nonverbal* VH. Permutations of routes and elicitation strategies were varied among participants.

7.2.4 Measures

Before the experiment the participants filled in a prequestionnaire measuring their age, gender, native language and highest level of education.

After each route they filled out a questionnaire about the interaction. The questionnaire measures the rapport between the VH and the participant, based on the questionnaire used in [29]. Furthermore we measured the perceived impression of the VH by having the participants rate the VH on 26 aspects on 7-point Likert scales, taken from the study of [34].

In the postquestionnaire after the final route, we asked which version of the VH they liked best, they thought was the most natural, the most social and the most attentive.

Our final measures are obtained from the video recordings of the interaction. In these video recordings we counted the number and the type (nonverbal, vocal or both) of the listener responses they provided to the VH.

7.2.5 Expectations

Our main expectation was that the verbal and nonverbal elicitation strategies would result in more listener responses than the default strategy, and that the combined method would result in yet more listener responses. Furthermore, we expected that not all response opportunities would actually yield a listener response.

Table 5 Listener Response ratio (Listener Responses given/Listener Response opportunities in the route-description) per subject per elicitation strategy. The value '–' means that the specific elicitation strategy was not presented to the subject or that the recording failed

Subject	Default	Combined	Vocal	Nonverbal	Avg
1	1	1	1	–	1
2	0.6	0.9	–	1	0.8
3	1	0.8	–	1	0.9
4	1	1	0.8	–	0.9
5	1	1	1	–	1
6	0.3	–	–	1	0.6
7	0.6	0.2	–	0.3	0.3
8	1	1	0.3	–	0.8
9	0.3	0.5	0.3	–	0.4

7.2.6 Results and discussion

We successfully elicited listener responses from the subjects (see Table 5). The amount of listener responses given seems highly subject dependent (see Table 5). Over half of the subjects gave a listener response on all response elicitation positions in the route explanation, even if no explicit elicitation strategy was used. Perhaps the pauses between segments in the route explanations provide a very strong feedback elicitation cue. Only 6 out of 237 listener responses were nonverbal only. 137 were both verbal and nonverbal.

We observed that five of the subjects used several instances of "teach back": the user would repeat part of the sentence said by the VH by way of listener response (cf. Interaction Example 2). Sometimes, when this happened, the VH would resume its speech (starting to explain the next step of the route), without waiting for the listener to finish. This was experienced as disruptive.

Interaction Example 2 Example of repetition in the recordings

Virtual Human: Take the second street on your right.
Subject: second street on my right.

Non-understanding was expressed in both intrusive (13x, for example: "over the square with the what?") and non intrusive ways (5x, for example: hesitant feedback: "Oh.. Keeeey" or with a puzzled look).

If we look at the result of the post-questionnaire (presented in Table 6) we notice the bad performance of the VH

Table 6 Results of the post-questionnaire in which the participants ranked the VHs on likeability, naturalness, social ability and attentiveness. For each dimension, for each condition, the numbers show how many participants rated that condition best/mid/least on the dimension; between parentheses, the percentage of all participants. Especially the VH with the *Vocal* elicitation strategy performs bad on these scales. The *Default* VH seems best

	Default	Combined	Vocal	Nonverbal
Like				
best:	5 (56%)	3 (33%)	0 (0%)	2 (50%)
mid:	2 (22%)	4 (44%)	1 (20%)	1 (25%)
least:	2 (22%)	2 (22%)	4 (80%)	1 (25%)
Natural				
most:	5 (56%)	2 (22%)	1 (20%)	1 (25%)
mid:	2 (22%)	3 (33%)	1 (20%)	3 (75%)
least:	2 (22%)	4 (44%)	3 (60%)	0 (0%)
Social				
most:	5 (56%)	3 (33%)	1 (20%)	0 (0%)
mid:	2 (22%)	4 (44%)	1 (20%)	3 (75%)
least:	2 (22%)	2 (22%)	3 (60%)	1 (25%)
Attentive				
most:	5 (56%)	3 (33%)	0 (0%)	1 (25%)
mid:	2 (22%)	5 (56%)	1 (20%)	1 (25%)
least:	2 (22%)	1 (11%)	4 (80%)	2 (50%)

with the *Vocal* elicitation strategy. Most of the five participants that interacted with this VH rated it the lowest on all scales. The prosodic modifications to the speech to elicit listener responses should thus be improved; the version used in this experiment is perceived as very unnatural. These modification also have a negative influence on the *Combined* elicitation strategy, since in this condition the same prosodic modifications are used. We think this is the reason why *Default* is generally considered the best condition on these measures.

The rapport questionnaire after each session did not yield any insightful results. Rapport between human is established through subtle interaction of nonverbal behavior, which are in sync (high rapport) or not (low rapport). A lot of these subtle nonverbal behaviors were not simulated. Furthermore, in every condition it was the user that interacted with the VH and not the other way around, making synchronization of these behaviors impossible.

7.3 Experiment 2

We learned from the last experiment that the user responded at almost every response opportunity in our routes. There was no difference between conditions in that regard. Between the blocks there was a pause of 1.5 seconds. Pause is another cue which is associated with listener responses. We think that this factor by itself is such a strong cue that it dominates the other conditions. In this second experiment we vary the pause length to test this hypothesis. In addition, we added a Wizard of Oz version of the COMPETITIVE/COOPERATIVE detector, to avoid the VH resuming its speech while the listener response of the user was not yet finished (something which happened when the user gave the longer "teach back" style responses).

7.3.1 Task

The task is the same as for the first experiment. The VH explains a route to the user, who has to reproduce the route on a map, afterwards.

7.3.2 Stimuli

For this experiment we use two of the routes (1 and 3) from the previous experiment. These routes each consist of 8 steps. For each step we vary the condition on two dimensions, the elicitation strategy and pause length.

For elicitation strategy we have either *elicitation* or *no elicitation*. The elicitation strategy is the combined version of the experiment 1 and the no elicitation strategy is the default version of experiment 1.

The pause length is varied between 0 ms, 500 ms, 1000 ms and 1500 ms. A pause of 0 ms means that there is no additional pause applied after the end of the sentence generated by MARY TTS.

7.3.3 Methodology

We invited 24 participants (14 male, 10 female, aged between 23 and 54; all, except 1, non-native English speakers) to interact with our route giving VH. Participants are told that the VH is able to perceive and react to short vocal and nonverbal listener responses (like nodding, saying *"Uh-huh"*, or *"Yes"*).

Before each interaction the user was presented the map with the starting point of the route. This map was taken away before the interaction started. During the interaction the route giving VH explained a route to the user. It was the task of the user to remember the route and reproduce it on the map, afterwards.

Each participant interacted two times with the route giving VH. The VH presented a different route each time. At the various response opportunities in the explanation of one route (i.e., between the steps) different combinations of elicitation strategy and pause length were offered. Each combination was offered exactly once per route per participant. The order of the conditions was varied between subjects.

In the previous experiment the participants sometimes repeated part of the sentence spoken by the VH as a acknowledgment (see Interaction Example 2). The VH started during

Table 7 Results of the second experiment

Condition	Nonverbal	Verbal	Both	Total
Elicitation	24	71	40	132
No Elicitation	28	59	43	127
0 ms	17	33	18	65
500 ms	11	32	23	65
1000 ms	13	31	20	64
1500 ms	11	34	22	65

this repetition with its next sentence. To prevent this from happening again we build a Wizard-of-Oz setup emulating the COMPETITIVE/COOPERATIVE detector. We did this using the anticipaters presented in Sect. 6.4. The wizard would press a button when the participant was speaking and release it when the participant stopped. The next step would not start as long as the button was pressed.

7.3.4 Measures

For each response opportunity we annotated whether the participant responded to the VH or not. Furthermore we annotated which modalities (verbal and/or nonverbal) were used by the participant in the response.

7.3.5 Results and discussion

In Table 7 the results of the experiment are presented. If we compare *elicatation* versus *no elicitation* we can see that there is no significant difference in the amount of elicited responses. Also if we compare the different pause lengths we can see no difference in amount of elicited responses. In case of the 0 ms pause length there are a little more nonverbal only responses than with the other pause lengths. A reason for that could be the fact that nonverbal feedback is less intrusive than verbal feedback, but again the results are not conclusive in that regard.

Although it does not show in the results, we did observe a few response occassions, where there was a response opportunity with a long pause, where the participant were late with their response. It was as if they were not inclined to give a response, but then were convinced by the long pause to give a response after all.

7.4 Conclusions

So far we have not been able to manipulate when a listener responds using subtle cues. Replicating cues discussed in literature and seen in corpora, like the use of rising pitch near the end of a sentence, a head nod accompanying the rise in pitch, looking for mutual gaze and manipulating pause lengths, we could not measure a significant effect on the amount of responses elicited by the VH.

The fact that participants did respond at most of the response opportunities created by the system (74% in Experiment 1, 67% in Experiment 2) suggests that another cue, present in all variations, is much more important. Most probably this is the syntax of the sentences and the nature of the task. Almost every sentence ended with a response opportunity. Each of these sentences carry a piece of information about the next step in the route. Given the task-oriented nature of the interaction, the user is inclined to acknowledge or express misunderstanding about the piece of information just given, like we do in real life when we are explained a route. The few variations we add with the vocal and nonverbal cues are dominated by this fact.

Another reason why some of our elicitation cues did not elicit more responses, is the fact that the task was difficult for the participants. This is illustrated by the fact that in the second experiment, only 9 out of 48 routes were drawn correctly afterwards. Because of this some of the participants did not always look at the virtual human, but at a neutral point, to reduce cognitive load. Therefore, they would not see the nonverbal elicitation cues we have implemented.

8 Discussion and conclusions

We have worked on a virtual human that, in the long term, should be able to interact with a human subject in an continuous manner. That is: it should be capable of the human-like interaction in which all partners perceive each other, express themselves, and coordinate their behavior to each other, continually and in parallel. The work reported in this paper resulted in progress on several aspects of continuous interaction, such as flexible and adaptive scheduling and planning of multimodal behavior (speech, gestures, facial expressions) including graceful interruption, automatic online classification of listener responses, and models for appropriate reactions to listener responses. We have set up two experiments in which a virtual human interacts with a subject. The aim of the experiment was to elicit Listener Response behavior, to provide us with more information on what user responses occur, and to serve as inspiration for further interaction models.

In the first experiment, we have observed Listener Responses given by our subjects that are much shorter than the waiting time between steps; other Listener Responses are much longer. Furthermore, Listener Responses are not given at every response opportunity. Starting to speak through a repetition or waiting for a Listener Response that is already finished confused some of our subjects. In the second experiment we used dynamic pauses. We used a Wizard-of-Oz setup to put the virtual human on hold while the subject was

speaking. The resulting behavior was that, if no Listener Response was forthcoming, the virtual human would continue after the planned waiting time. If a response was given, the virtual human waited until it finished. None of the subjects in the second experiment experienced confusion because the virtual human reacted too soon. In an autonomous version, the Wizard-of-Oz setup could be replaced by speech detection in combination with the classifiers presented in Sect. 5 to achieve similar results.

We have observed several "teach back" responses from the listener, in which (s)he repeated part of the utterance from the speaker. Detecting such repetitions is still an open issue. Since the repetitions often repeat the landmarks used in the route, perhaps the occurrence of landmarks (as detected by a keyword spotter) could be used as one of the cues for the identification of repetions. Assuming that we can automatically assess whether a response is a "teach back" repetition, the new preplanning mechanisms we have developed can be used to generate an acknowledgment of the repetition (see Interaction Example 3).

Interaction Example 3 Handling repetition

Virtual Human: Turn right before the obelisk.

Subject: right before the obelisk.

Virtual Human: Yes. Then turn left and cross the bridge.

A generic set of such acknowledgements (e.g., "that's correct", "yes", "uhhuh") can be preplanned and activated instantly when needed. If the next step of the route description that follows the acknowledgements has already been planned, Elckerlyc's retiming mechanisms (see [56]) can be used to shift it in time so that a full replan of the route description is avoided. In human-human interaction, such a set of *appositional beginnings* is also frequently used to allow an interlocutor to take the turn without having a plan in hand [10, 43].

Interruptions are detected as Competitive utterances by our classifier. If the subject interrupts the Virtual Human (as in Interaction Example 4), his ongoing route description can be gracefully interrupted using mechanisms discussed in Sect. 6.3. We can either preplan all alternative explanations, or use in-between generic preplanned sentences to cover up the scheduling, like "Ok, let me explain that again".

Interaction Example 4 Graceful interruption

Virtual Human: Turn left at the square with the obelisk. Then take the second ...

Subject: over the square with the what?

Virtual Human: [gracefully interrupts ongoing behavior, selects an alternative for "Turn left at the square with the obelisk"] So you enter the square, there is an obelisk at the center of the square.

In the current implementation we have not yet explored different strategies for the VH to deal with Listener Responses from the user. Depending on the type of behavior that we would like to realize, such strategies are selected in concordance with a politeness strategy and certain personality traits (e.g., dominance or impatience). For example: a rude or dominant virtual human could explicitly ignore interruptive responses by speaking louder and leaning forward to keep the turn, while an insecure virtual human could explicitly wait for feedback after each of its utterances. Some of these strategies can already be implemented with the existing modules (e.g. merge a lean forward behavior, wait for feedback, then continue). Elckerlyc can modify parameter values of ongoing behavior in an adhoc manner, allowing changes to for example gesture amplitude or speech volume. We are currently exploring how such parameter value changes can be *specified* in a formal manner, either through BML or through another channel that communicates with Elckerlyc (see [57] for a more elaborate discussion on this topic).

Acknowledgements This research has kindly been supported by the GATE project, funded by the Dutch Organization for Scientific Research (NWO) and the Dutch ICT Regie, and by the FP7 NoE SSPNet. The authors would like to thank Albert Ali Salah for organizing the eNTERFACE'10 summer event that led to this project.

References

1. Allwood J, Cerrate L (2003) A study of gestural feedback expressions. In: Paggio P, Jokinen K, Jönsson K (eds) 1st Nordic symposium on multimodal communication, pp 7–22
2. Anderson AH, Bader M, Bard EG, Boyle E, Doherty-Sneddon G, Garrod S, Isard S, Kowtko JC, McAllister J, Miller J, Sotillo C, Thompson H, Weinert R (1991) The HCRC Map Task corpus. Lang Speech 34:351–366
3. Bavelas JB, Coates L, Johnson T (2000) Listeners as co-narrators. J Pers Soc Psychol 79(6):941–952
4. Bavelas JB, Coates L, Johnson T (2002) Listener responses as a collaborative process: The role of gaze. J Commun 52(3):566–580
5. Benus S, Gravano A, Hirschberg J (2007) The prosody of backchannels in American English. In: Proceedings of the 16th international congress of phonetic sciences 2007, pp 1065–1068
6. Black AW, Tokuda K, Zen H (2002) An HMM-based speech synthesis system applied to English. In: Proc of 2002 IEEE SSW, Santa Monica, CA, USA
7. Brady PT (1968) A statistical analysis of on-off patterns in 16 conversations. Bell Syst Tech J 47:73–91
8. Carletta JC, Isard S, Doherty-Sneddon G, Isard A, Kowtko JC, AH Anderson (1997) The reliability of a dialogue structure coding scheme. Comput Linguist 23(1):13–31
9. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

10. Clark HH (1996) Using language. Cambridge University Press, Cambridge
11. Clark HH, Brennan SE (1991) Grounding in communication. In: Resnick LB, Levine JM, Teasly SD (eds) Perspectives on socially shared cognition. American Psychological Association, Washington
12. Clark HH, Krych MA (2004) Speaking while monitoring addressees for understanding. J Mem Lang 50(1):62–81. doi:10.1016/j.jml.2003.08.004
13. Dhillon R, Bhagat S, Carvey H, Shriberg E (2004) Meeting recorder project: Dialog act labeling guide. Tech Rep ICSI Technical Report TR-04-002, International Computer Science Institute
14. Duncan S Jr (1972) Some signals and rules for taking speaking turns in conversation. J Pers Soc Psychol 23(2)
15. Duncan S Jr (1974) On the structure of speaker-auditor interaction during speaking turns. Lang Soc 3(2):161–180. doi:10.1017/s0047404500004322
16. Edlund J, Heldner M, Al Moubayed S, Gravano A, Hirschberg J (2010) Very short utterances in conversation. In: Proceedings of fonetik
17. Eyben F, Woellmer M, Schuller B (2010) openSMILE—the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM multimedia, pp 1459–1462
18. French P, Local J (1983) Turn-competitive incomings. J Pragmat 7:17–38
19. Fujimoto DT (2007) Listener responses in interaction: a case for abandoning the term, backchannel. J Osaka Jogakuin 2 Year Coll 37:35–54
20. Goldwater S, Jurafsky D, Manning CD (2010) Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Commun 52:181–200
21. Goodwin C (1981) Conversational organization: interaction between speakers and hearers. Academic Press, San Diego
22. Goodwin C (1986) Between and within: alternative sequential treatments of continuers and assessments. Hum Stud 9(2–3):205–217. doi:10.1007/bf00148127
23. Gravano A, Hirschberg J (2009) Backchannel-inviting cues in task-oriented dialogue. In: Proceedings of interspeech, Brighton, pp 1019–1022
24. Gustafson J, Neiberg D (2010) Prosodic cues to engagement in non-lexical response tokens in Swedish. In: DiSS-LPSS Joint Workshop
25. Heldner M, Edlund J (2010) Pauses, gaps and overlaps in conversations. J Phonetics 38(4):555–568. doi:10.1016/j.wocn.2010.08.002
26. Heylen D (2006) Head gestures gaze and the principles of conversational structure International. Int J Humanoid Robot 3(3):241–267
27. Heylen D, Bevacqua E, Tellier M, Pelachaud C (2007) Searching for prototypical facial feedback signals. In: Pelachaud C, Martin JC, André E, Chollet G, Karpouzis K, Pelé D (eds) Proceedings of the 7th international conference intelligent virtual agents. Lecture notes in computer science, vol 4722. Springer, Berlin, pp 147–153. doi:10.1007/978-3-540-74997-4_14
28. Kendon A (1967) Some functions of gaze direction in social interaction. Acta Psychol 26:22–63
29. de Kok I, Heylen D (2011) The MultiLis corpus—dealing with individual differences of nonverbal listening behavior. In: Proceedings of COST 2102: toward autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues, pp 362–375
30. Kopp S (2010) Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Commun 52(6):587–597. doi:10.1016/j.specom.2010.02.007
31. Kopp S, Krenn B, Marsella SC, AN Marshall, Pelachaud C, Pirker H, Thórisson KR, Vilhjálmsson HH (2006) Towards a common framework for multimodal generation: the behavior markup language. In: Gratch J, Young MR, Aylett RS, Ballin D, Olivier P (eds) Proceedings of the 6th international conference on intelligent virtual agents. Lecture notes in computer science, vol 4133. Springer, Berlin, pp 205–217
32. Kurtic E, Brown GJ, Wells B (2010) Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration. In: Proceedings of interspeech, pp 2550–2553
33. Lee CC, Lee S, Narayanan SS (2008) An analysis of multimodal cues of interruption in dyadic spoken interactions. In: Proceedings of interspeech, pp 1678–1681
34. ter Maat M, Truong KP, Heylen D (2010) How turn-taking strategies influence users' impressions of an agent. In: Allbeck J, Badler NI, Bickmore T, Pelachaud C, Safonova A (eds) Proceedings of the 10th international conference on intelligent virtual agents, Philadelphia, Pennsylvania, USA. Lecture notes in computer science, vol 6356. Springer, Berlin, pp 441–453. doi:10.1007/978-3-642-15892-6_48
35. Manusov V, Trees AR (2002) "Are you kidding me?": The role of nonverbal cues in the verbal accounting process. J Commun 52(3):640–656. doi:10.1111/j.1460-2466.2002.tb02566.x
36. McKinneya MF, Moelants D, Davies MEP, Klapuri A (2007) Evaluation of audio beat tracking and music tempo extraction algorithms. J New Music Res 36(1):1–16
37. Neiberg D, Gustafson J (2010) The prosody of Swedish conversational grunts. In: Proc of Interspeech
38. Neiberg D, Truong KP (2011) Online detection of vocal listener responses with maximum latency constraints. In: Proc of ICASSP, p 2011
39. Nijholt A, Reidsma D, van Welbergen H, op den Akker H, Ruttkay ZM (2008) Mutually coordinated anticipatory multimodal interaction. In: Esposito A, Bourbakis NG, Avouris N, Hatzilygeroudis I (eds) Verbal and nonverbal features of human-human and human-machine interaction. Lecture notes in computer science, vol 5042. Springer, Berlin, pp 70–89
40. Norwine AC, Murphy OJ (1938) Characteristic time intervals in telephonic conversation. Bell Syst Tech J 17:281–291
41. Reidsma D (2008) Annotations and subjective machines—of annotators, embodied agents, users, and other humans. PhD thesis, University of Twente. doi:10.3990/1.9789036527262
42. Reidsma D, Truong K, van Welbergen H, Neiberg D, Pammi S, de Kok I, van Straalen B (2010) Continuous interaction with a virtual human. In: Salah AA, Gevers T (eds) Proceedings of the eNTERFACE'10 summer workshop on multimodal interfaces, pp 24–39
43. Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50:696–735
44. Schegloff E (2000) Overlapping talk and the organization of turn-taking for conversation. Lang Soc 29:1–63
45. Schlangen D, Skantze G (2009) A general, abstract model of incremental dialogue processing. In: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-09)
46. Schröder M (2010) The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. Adv Hum-Comput Interact 2010:319406. doi:10.1155/2010/319406
47. Schröder M, Trouvain J (2003) The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int J Speech Technol 6(4):365–377
48. Schröder M, Charfuelan M, Pammi S, Türk O (2008) The MARY TTS entry in the Blizzard Challenge 2008. In: Proc of the Blizzard Challenge

49. Skantze G, Hjalmarsson A (2010) Towards incremental speech generation in dialogue systems. In: Proceedings of SIGdial
50. Thiebaux M, Marshall AN, Marsella SC, Kallmann M (2008) Smartbody: Behavior realization for embodied conversational agents. In: Proceedings of the 7th international conference on autonomous agents and multiagent systems, pp 151–158
51. Thórisson KR (2002) Natural turn-taking needs no manual: Computational theory and model, from perception to action. In: Granström B, House D, Karlsson I (eds) Multimodality in language and speech systems. Kluwer Academic, Dordrecht, pp 173–207
52. Toda T, Tokuda K (2007) A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans Inf Syst E90-D(5):816–824
53. Walker MB, Trimboli C (1982) Smooth transitions in conversational interactions. J Soc Psychol 117:305–306
54. Ward N (2006) Non-lexical conversational sounds in American English. Pragmat Cogn 14(1):129–182
55. Ward N, Tsukahara W (2000) Prosodic features which cue back-channel responses in English and Japanese. J Pragmat 32(8):1177–1207
56. van Welbergen H, Reidsma D, Ruttkay ZM, Zwiers J (2010a) Elckerlyc: A BML realizer for continuous, multimodal interaction with a virtual human. J Multimodal User Interfaces 3(4):271–284. doi:10.1007/s12193-010-0051-3
57. van Welbergen H, Reidsma D, Zwiers J (2010b) A demonstration of continuous interaction with Elckerlyc. In: Proceedings of the third workshop on multimodal output generation, CTIT Workshop Proceedings. vol WP2010, pp 51–57