# Continuous Interaction with a Virtual Human

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

***Abstract*—Attentive Speaking and Active Listening require that a Virtual Human be capable of *simultaneous perception/interpretation and production* 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 perceives from its partner. This report presents the results of a four week summer project that was part of eNTERFACE'10. The project resulted in progress on several aspects of continuous interaction such as scheduling and interrupting multimodal behavior, automatic classification of listener responses, generation of response eliciting behavior, and models for appropriate reactions to listener responses. A pilot user study was conducted with ten participants. In addition, the project yielded a number of deliverables that are released for public access.**

***Index Terms*—Virtual Humans, Attentive Speaking, Listener Responses, Continuous Interaction**

## I. 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 production* 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 perceives from its partner. This report presents the results of a four week summer project that was part of eNTERFACE'10. The project resulted in progress on several aspects of continuous interaction such as flexible and adaptive scheduling and planning including graceful interuption, automatic classification of listener responses, generation of response eliciting behavior, and models for appropriate reactions to listener responses. We made a start on evaluating the results in classification experiments as well as in a pilot user study. In addition, the project yielded a number of deliverables that are released for public access, among which a public release of Elckerlyc, 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.

## II. Background and Motivation

The design of VHs often focuses on the combination of speech with gestures in conversational settings. They tend to be developed using a turn-based interaction paradigm in which the user and the system take turns to talk. 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 turn-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 [1], [2]. This requires the realizer to be capable of immediate adaptation – in content and in timing – to the dynamics of the environment and the user.

The main objective of this project is to explore this kind of coordination behavior in ECAs, modeling and implementing the sensing, interaction and generation for what we call continuous interaction. A continuous interactive ECA will be able to perceive the user and generate conversational behavior fully in parallel, and can coordinate behavior to perception continuously – a capability which is not yet present in most state-of-the-art ECAs.

One of the major sources of overlap in conversation, and therefore a very good domain for addressing continuous interaction capabilities in Virtual Humans, are Listener Responses [3]. We will take a first step towards the global goal by making a VH that is capable of actively dealing with Listener Responses from the user, while the VH is speaking.

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

### A. Structure of this Report

This report is structured as follows. Section III gives an introduction to the theoretical background of Responses and Attentive Speaking on which we based our approach. Section IV presents the overall system setup of an interactive Virtual Human system as we used it in our development and experiments. Sections V and VI introduce the corpora that we used, and analyse them with respect to the characteristics of Responses that we find in them. Section VII is dedicated to the automatic classification of Responses. Sections VIII and IX concern behavior scheduling and planning for continuous interaction for Virtual Humans: they describe the already existing possibilities as well as the new developments achieved in this project. Sections X and XI discuss our work on the Response Elicitation pilot user study. The paper ends with a discussion of what we have achieved, and where we need to go next.

## III. Listener Responses and Attentive Speaking

An active listener shows his or her interest, attention and/or attitude with respect to the speakers utterances in many ways: through gaze direction and eye contact, using face expressions, using short utterances like "yeah", "okay", and "hm-m", etcetera. An attentive speaker 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. In this section, we discuss (listener) responses and attentive speaking in more detail.

### A. Responses and Listener Responses

The conversational context is that of a VH is explaining a certain route on a map to the user. This conversational context implies that the VH is mostly speaking (is a Speaker), and the user is listening (is a Listener). At some point, the user starts to talk. This may be to give feedback or it may be a question, answer, statement, or other full contribution to the conversation. The user's utterance may overlap an utterance of the VH, or it may be at a moment that the VH was silent.

We refer to as everything the Listener says as "Responses", which implies the role in the conversation.

The Listener commonly utter responses such as "yeah", "mhm", "uhu". Fujimoto [3] propose to call these short utterances Listener Responses. These are short utterances or vocalizations which are interjected into the Speakers' account without causing an interruption, or being perceived as competitive of the floor. They serve many functions, were the most important is to signal that the Listener

hears that the Speaker is talking and nothing more than this neutral function. This function is sometime called back-channeling and is not mandatory. Another common function is signaling understanding to what the speaker is saying. This function is commonly referred to as Acknowledgment. In addition, they may be used as carriers of more subtle information, conveyed by intonation, voice quality, face expression, rhythm, content of the words, etcetera.

From a more generalized point of view, a Response may convey information regarding Understanding (whether the Listener understands the utterance of the Speaker), Attentiveness (whether the Listener is attentive to the speech of the Speaker), Attitude [4] and Affect [5], and may be described as being competitive (interruptive) or cooperative (non-interruptive) [6].

### B. Attentive Speaking

A good speaker pays attention to the listener. He moderates his speech and tailors it to the responses from the listener. Listeners are not merely listening, but are co-narrating along with the speaker [7]. A good virtual human should be able to do this as well.

This interaction between speaker and listener works in various ways. To illustrate this we will give a few examples from literature. Clark and Krych [8] identify several strategies in dialogue that depend on opportunities that arise, intentionally or not, mid-sentence. They claim that speakers make the alterations instantly, typically initiating them within half a second of the opportunities becoming available.

One of the strategies the speakers apply to coordinate their speech is self-interruption. If the listener provides a response 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 of his utterance and starts a new one (see Example 1).

**Interaction Example 1** Self-interruption.

Speaker: So starting from the square, you go...
Listener: euhm?
Speaker: I mean the square with the obelisk on it.

The observations from Goodwin [9] work on a lower level. In his observation, the speaker does not change what he says based on the responses from the listener, but the timing is coordinated with the listener. He makes a distinction between continuers and assessments. Continuers simply acknowledge the receipt of the talk just heard and signals the speaker to continue his talk. Assessments are the result of an analysis of the speakers' talk by the listener based on which, the listener has produced an action that is responsive to the particulars of the talk. Continuers are usually placed between two subsequent speech units, while assessments are placed in the midst of a unit and completed before a new unit starts. This is actually facilitated by the speaker. So, 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.

This coordination does not only facilitate vocal responses from the listener, also nonverbal signals are dealt with by the speaker. Goodwin [10] 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 look back.

This is merely a selection of situations and strategies in which the speaker moderates his speech to the responses of the listener. There are many more, which we did not cover, but they illustrate the type of coordination we are aiming to achieve with our system. It is our aim to create a system which is technically able to achieve the same level of continuous interaction with the user as illustrated by these examples.

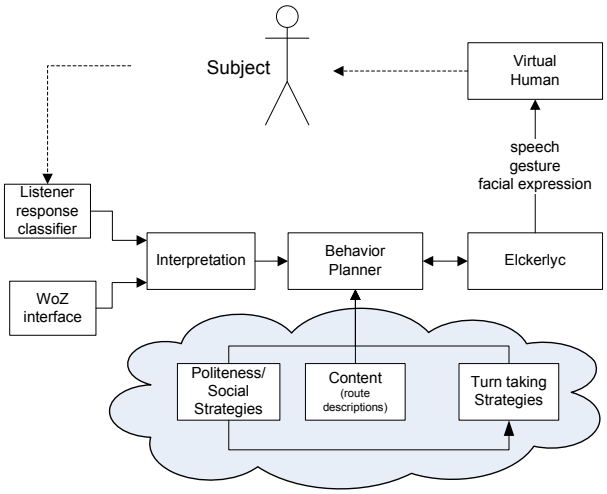



Fig. 1: System architecture

## IV. System Overview

Fig. 1 gives an overview of the architecture of the interactive virtual human system that we have developed. The virtual human explains a route through a city, in such a way as to elicit Responses form the user. We detect the occurrence of Responses(e.g., "uh-huh", "mmm") using both non-verbal vocalization analysis and a Wizard of Oz interface. The behavior planner specifies the behavior to be realized on the basis of politeness and social strategies and conversation content (a specification of the route to explain). The behavior is constructed using speech, gestures, gaze, and face expressions.

If Responses occur, Elckerlyc is instructed to gracefully interrupt the currently running behavior or to retime or re-parameterize (speak louder, increase the amplitude of gestures etc.) its behavior. New behavior can be constructed by selecting and inserting new BML fragments in order to react to interruption. The exact method of feedback handling is influenced by turn-taking strategies and politeness/social strategies. The different components are connected using the SEMAINE framework [11].

## V. Corpora

We used two corpora in this project, namely the MapTask corpus [12] and the MultiLis corpus [13]. These corpora were used for two purposes: (1) to find out more about the content and timing of listener responses, and (2) as training and testing material for our classifiers.

### A. The MapTask Corpus

The HCRC Map Task Corpus is a set of 128 dialogues. The task is for one subject to explain a route to another subject. The one who explains the route is denoted as the "giver" and the one who receives 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 dialogues from the face-to-face condition since it is closer to our scenario of an interaction with a Virtual Human. The two conversations labeled as `q3ec1` and `q3ec5` where discarded due to a buzz in the speech signal.

The segmentation of the dialog in the MapTask corpus is based on manual annotations. For the analyses and experiments discussed in this report, we chose to use instead segmentations based on an ideal voice activity detector, because that will more closely reflect the conditions that we will encounter in the application of a conversation with a Virtual Human. We segment the corpus into *talk spurts* [14], defined as a minimum voice activity duration of 50ms separated by a

TABLE I: Confusion matrix for the annotation of overlapping talk spurts on Competitiveness. Cohen's $\kappa$=0.60; Krippendorff's $\alpha$ (nominal) = 0.45.

| | COMPETITIVE | COOPERATIVE |
|---|---|---|
| COMPETITIVE | 88 | 77 |
| COOPERATIVE | 40 | 319 |

minimum inter-pause of 200 ms. These talk spurts are referred to as "ideal VA Detector talk spurts". If a talk spurt is comprised of more than one MapTask segment, the talk spurt is labeled with the label from the first MapTask segment included in the talk spurt. This gives a consistent segmentation strategy, uses all relevant speech, and the results will better resemble the condition when a real voice activity detector is used.

To simulate real-world conditions even closer, we additionally created a second set of talk spurts using the OpenSmile voice activity detector. For each ideal VA Detector talk spurt, 3 seconds of silence is added in front, and 10 seconds of the original audio following the ideal VA Detector talk spurt is added to the end. Then the first talk spurt detected by the OpenSmile voice activity detector, configured with minimum voice activity duration of 100 ms and a minimum inter-pause of 200 ms, is assigned the same label as the "ideal VA Detector talk spurt" and saved for further experiments. If no talk spurt is detected, then the corresponding label is thrown away. We refer to these segments as "OpenSmile VADetector talk spurts".

We used the official MapTask annotations concerning the distinction between Acknowledgement Moves (MTACK) and other talk spurts (NONMTACK). The precise definition of an Acknowledgment Move is found in [15], but they closely resemble the term Listener Response [3] and thus serve our purpose. According to Carletta et al. [15], these MapTask annotations are good ($\kappa = .83$), although one of the largest confusions did involve the Acknowledgement Moves (confusion with Ready and Reply-Y).

In addition, we annotated part of the data with information whether the talk spurt intends to take the floor (COMPETITIVE) or not (COOPERATIVE).

The following talk spurts were annotated:

- We only annotated NONMTACKs, as MTACKs are supposed to be COOPERATIVE by definition.
- We annotated only Responses in overlap (Listener's talk spurt starts between the start and the end of the Speaker's talk spurt) because the COOPERATIVE/COMPETETIVE dimension only makes sense for overlapping talk spurts.
- We only annotated NONMTACKs, which does not have any MTACKs within the local overlap. For example, a NONMTACK which is intercepted in overlap by MTACK is excluded.

In the data that we used, there are 1232 candidate talk spurts to be annotated. Of these, 524 talk spurts (quad 1-4) were labelled by two annotators. The confusion table and relibaility values are given in Table I. The level of agreement for this annotation is in the range of highly subjective annotations [16]; the annotators agree on a certain amount of talk spurts being COOPERATIVE, but have difficulty agreeing on which talk spurts are COMPETITIVE.

### B. The MultiLis Corpus

Because the mapTask corpus does not contain video recordings, it could not provide us information about nonverbal responses and nonverbal respose elicitation behavior such as gaze, nods, and face expressions. For this, we used the MultiLis corpus.

TABLE II: Top 20 most frequently occurring Acknowledgement talk spurts in the MapTask corpus (MTACK talk spurts), accounting for 7313 out of 9823 of these talk spurts.

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

The MultiLis corpus [13] is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totalling 131 minutes. Participants were assigned the role of either speaker or listener during an interaction. The speakers summarized a video they have just seen or reproduced a recipe they have just studied for 10 minutes. Listeners were instructed to memorize as much as possible about what the speaker was telling. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener. What is unique about this corpus is the fact that it contains recordings of three individual listeners to the same speaker in parallel, while each of the listeners believed to be the sole listener. The speakers saw one of the listeners, believing that they had a one-on-one conversation. The aim of the corpus was to collect responses from different individuals to the same speaker context. The corpus illustrates the individual differences in listening behavior, but also includes differences in the amount of responses that individual speakers were able to elicit.

## VI. ANALYSIS OF RESPONSES IN HUMAN-HUMAN INTERACTION

This section provides an analysis of properties of Responses from the MapTask corpus. Rather than providing a complete analysis, we only adress the parts which are crucial for the design of the system. Table II shows the most frequently occurring word content for MTACK talk spurts, accounting for 7313 out of 9823 of these talk spurts.

### A. MTACK *Content*

Figure 2 shows the duration of MTACKs vs. the other dialog moves. It is clear that MTACKs have a short duration and may (partially) be detected by duration alone. Concerning overlapping speech, we can observe the following: The proportion of overlapped speech in the MapTask corpus is 9.1%, the proportion of MTACKs is 7.3% and the proportion of MTACKs in overlapped speech is 34.9%. Thus, MTACKs are more common in overlap than in non-overlapped speech.

### B. *Gaps Following* MTACK *talk spurts*

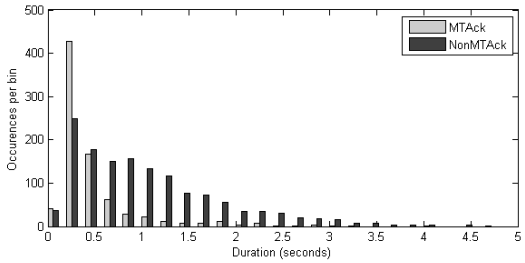



Fig. 2: Duration of MTACKs vs. duration of other dialog moves, using bins of 200 msec.

Since we are trying to build a Virtual Human that can deal with Responses in a continuous interactive way, we also investigated the continuation talk spurt of the Speaker following the onset of a MTACK Response. For all MTACK Responses that do *not* interrupt the Speaker (i.e. the Speaker continues speaking afther the onset of the Response) we calculated the gap between the *end* of the Response and the *beginning* of the continuation talk spurt of the speaker. This gap has a negative value if the Speaker continues speaking before the end of the Response. Figure 3 shows the distribution of the gap for all Speaker continuation talk spurts. The figure shows that the Speaker commonly continues to speak after roughly 0-400 ms. It also shows that negative gap – that is, overlap – is not uncommon. This means that for a responsive dialog with a Virtual Human, Responses from the user need to be classified before they are finished. This might 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 listener responses cannot be accounted for. It is also an open question how well a speech recognizer will perform in detecting grunt-like nature of some listener responses. This is because Responses such as "mmhmm" are tokens which are shown to be unstable in their allophonic surface realizationsm, and there is no standardized annotation scheme for these [17].

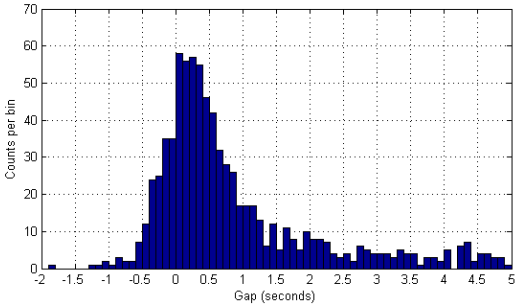



Fig. 3: The gap or overlap (negative gap) between a MTACK Response and the incterlocutors' continuation using bins of 100 ms.

### C. Duration of COMPETITIVE and COOPERATIVE Responses

Figures 4 and 5 give the distribution of the duration of COMPETITIVE and COOPERATIVE Responses, and of the durations of the *overlap* for both types of Responses.

We notice that these distributions are different. Short overlaps around 100 ms are more likely for cooperative speech rather than for competitive speech. The most likely overlap duration for cooperative speech is around 100ms, and it wears off around 2100 ms. The most likely overlap duration for competitive speech is around 300ms, and it wears off around 1100 ms. This means that a detector should give a decision as early as possible after the onset of the Response: preferably at 300ms, but no later than 1100ms.

Secondly, we observe that cooperative talk spurt tend to be shorter in durations than talk spurt for competitive speech. 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 talk spurt and overlap.

## VII. CLASSIFICATION OF LISTENER RESPONSES

This section deals with the classification of Responses based on audio input. Being an attentive speaker includes giving attention to

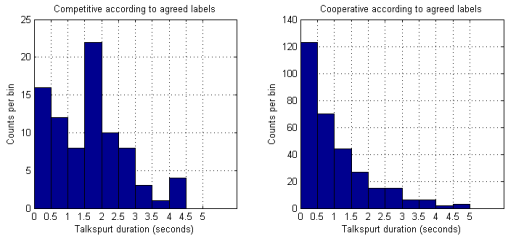



Fig. 4: Durations of talkspurts in overlap with no MTACK context (within the overlap). To the left are COMPETITIVE and to the right COOPERATIVE Responses.

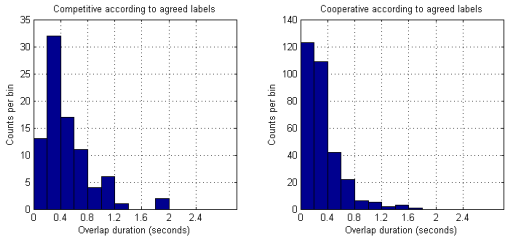



Fig. 5: Durations of the overlap with no MTACK context (within the overlap). To the left are COMPETITIVE and to the right COOPERATIVE Responses.

what the listener says and taking appropriate action. First of all, this involves recognizing Responses and the information they convey. We approach this by classification of incoming voice activity in the audio channel. As mentioned before (Sections IV and VI), it is important to classify incoming talk-spurts before they end, preferably within 300-700 msec of the onset of the speech.

The classifiers are needed for the system to determine, given incoming speech from the user, what the reaction of the Virtual Human should be. If the incoming speech overlaps speech from the Virtual Human, the decision may be to stop speaking, or to continue speaking in overlap. The latter makes sense when the incoming speech is a COOPERATIVE Response. If the incoming speech does not overlap, the reaction of the Virtual Human should very much be determined by the information conveyed by the Response. For example, an MTACK Response probably requires no change of the dialog plan; a Response expressing non-understanding or disagreement may require elaboration, initiation of a clarification dialog, or other more drastic revisions of the dialog plan. The last type of Responses are not dealt with by the classifiers presented here.

We classify Responses using the cascade shown in Figure 6. The first classifier in the cascade is trained on the MapTask corpus to distinguish MTACK talk spurts from other talk spurts. MTACK talk spurts are, among other things, by definition COOPERATIVE Responses. Talk spurts *not* classified as MTACK may be COOPERATIVE or COMPETITIVE (see Section V-A). Concerning these NONMTACK talk spurts we focus on talk spurts produced by the user in overlap as they more urgently require a decision from the Virtual Human (namely, to continue speaking even while the user is speaking too, or not). We tried two different approaches to classify those talk spurts. The first approach was based on classifying them according to the theoretical distinction between COOPERATIVE and COMPETITIVE Responses. The second approach was pragmatically oriented, based on *predicting the outcome of the overlap*, that is, predict whether the Speaker or the Listener is the one who continues speaking after the overlap. The third approach is a hybrid approach, and attempts to exploit a possible relation between the pragmatic "outcome of overlap" rules and the theoretical distinction from the first approach.

All classification experiments were performed using openSMILE [18] for automatic feature extraction and *libsvm* [19] for classification.

In summary, this leads to four main classification tasks.

- **Classifier I** Classification of all Responses into MTACK / NONMTACK
- **Classifier IIa** Classification of NONMTACK, produced in overlap, into COOPERATIVE / COMPETITIVE, based on our manual annotations (the theoretical approach)
- **Classifier IIb** Prediction of the outcome of the overlap for all NONMTACK produced in overlap (the pragmatic approach)
- **Classifier IIc** Classification of NONMTACK, produced in overlap, into COOPERATIVE / COMPETITIVE, based on the hybrid approach

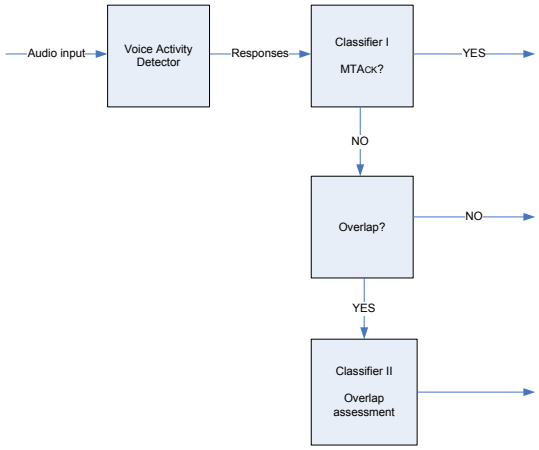



Fig. 6: Cascade used to classify incoming Responses from the user.

### A. Maximum latency classification

The analysis of the gap after a listener response in Figure 3, showed the presence of a negative gap, i.e. an overlap. This means a decision whether incoming speech is a listener response or not has to be made before the the listener response ends. Thus, we consider a maximum latency design for the detector. It is implemented as a voice activity detector which sends an end message after the talk-spurt ends, or at a predefined duration threshold, denoted as the maximum latency. If the duration reaches the threshold, it continues to work as normal voice activity detector internally, otherwise it might trigger again. Note that the detector may trigger before the maximum latency if the talkspurt is shorter than the threshold subtracted by the minimum inter pause threshold. For online detection, this maximum latency design was implemented in openSMILE [18].

### B. Feature trajectories as length-invariant Discrete Cosine Coefficients

To parameterize the trajectories of each feature through out a talkspurt, we use DCT coefficients invariant to segment length:

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}(n+\frac{1}{2})k\right) \qquad k = 0,\ldots,N \qquad (1)$$

where $N$ is the segment length, $x_n$ is the feature value at time $n$ and $X_k$ is the k'th coefficient.

These DCT coefficients are much faster to compute than polynomial regression coefficients, since polynomial regression require matrix inversion. This makes length invariant DCT coefficients more suitable for online systems. 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 which has an highly speaker dependent additive bias, such as F0. These DCT coefficients has been used to visualize a single average trajectory of multiple speech segments [20]. When a DCT is applied on MFCCs, one obtain the cepstrum modulation spectrum. The usage of length-invariant cepstrum modulation spectrum was first introduced by [21], although no specific term was used at the time. The cepstrum modulation spectrum has been use for speech recognition [22] and in its length invariant version for affective detection [23]. By omitting the 0th DCT coefficient for MFCCs in the time dimension, then any channel mismatch which appear as an additive bias in the quefrency will not cause any problem. Our experiments will determine whether omitting the 0'th coefficient still gives a decent classifier. Unless anything else is stated, the 0'th DCT coefficient in time dimension is always omitted.

| | TRAINING | DEVELOPMENT | EVALUATION |
|---|---|---|---|
| MTACK | 775 | 482 | 537 |
| NONMTACK | 1315 | 677 | 1138 |

TABLE III: Number of talkspurts used for training, developing and testing Classifier I.

### C. Support Vector Machine classification

All classifiers use Support Vector Machines (SVM) with Radial Base Function Kernel as implemented in *libsvm* [19]. In a few cases, we consider a minor but pragmatic modification to the standard SVM scheme, which is here denoted as *rescaling*. When feature sets of different nature are evaluated on the development set, quite different optimal $\gamma$ values are found for each feature set. The $\gamma$ parameter in a radial base kernel is proportional to the inverse of the variance in a Gaussian. This means that if each feature set would have different $\gamma$, then a more optimal decision hyperplane may be found. One solution to this problem uses multiple kernels [24]. Here we offer a simple and pragmatic solution for this problem. After each feature set $f$ has been evaluated separately, the optimal $\gamma^f_{optimal}$ is saved. When the combined feature set is created, a rescaling procedure is applied, after the regular scaling to $[-1,1]$ or N(0,1) . The original scaled feature set $x^f$ for each feature set is then rescaled by

$$\hat{x}^f = \frac{\gamma^f_{optimal}}{\min_{i=1\ldots I}\gamma_i} \qquad (2)$$

where $i$ denotes the indexes for all $\gamma$ in the grid search. This rescaling procedure can be applied to most standard SVM implementations with only minor modifications.

### D. Experimental setup

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. The number of talkspurts used in the classification experiments can be found in Table III. The SVM regularization parameters are optimized on the development set, and the best parameters are then used for test on the evaluation set.

As explained in Section V-A, the first series of experiments explores features and combinations thereof under the assumption that an ideal voice activity detector is available (referred to as the "ideal VAD talk spurt" situation). In the second series of experiments, the ideal segmentation is replaced by an actual voice activity detector

based on energy thresholds (referred to as the 'OpenSmile VAD talk spurt' situation). This is done to ensure that the classification results reflects real life conditions as closely as possible. Since a parametrization of the trajectory of each feature is used, the resulting models are expected to be sensitive to mismatch in segmentation. Thus, the same segmentation should be used for training and on-line recognition. Still we considered safe to extrapolate the results from the first series, and use the best found combinations of features for the experiments using the 'OpenSmile VAD talk spurts'.

### *E. Classifier I:* MTACK *vs. Other*

*1) Features:* For the task of classifying incoming speech as a MTACK or not, a set of acoustic features are considered.

- F0: Back-channels has been shown to have a rise or drop in F0 [25][20].
- Intensity: Back-channels has been shown to have distinct intensity contours [25]
- MFCC: Similar lexical content, see Table II, may be captured by MFCCs.
- Duration: As seen in Figure 2, MTACKs have shorter duration then other type of speech. For training, the full talk-spurt 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.

All features are parametrized using DCT-coefficients 1-6 or 0-6, as described in Section VII-B. As classification method, we used a $\nu$-SVM. The parameters $g$ and $c$ were optimized on the DEV set (on F_avg) through a simple gridsearch with growing sequences of the $\nu$ (sequences growing linearly) and $g$ (sequences growing exponentially) parameters within ranges of [0.025, 0.6] and [0.0156, 4] respectively.

For this classifier, a maximum latency of 300ms or 500ms was chosen.

| Feature(s) | 300 ms | 500 ms |
|---|---|---|
| F0 | 55 | 59 |
| Intensity | 60 | 62 |
| MFCC with 0th | 72 | 75 |
| MFCC without 0th | 74 | 75 |
| Duration | 55 | 71 |
| Spectral flux | 66 | 67 |
| Intensity, Sp. flux, MFCC with 0th | 73 | 76 |
| Intensity, Sp. flux, MFCC with 0th, Dur. | 75 | 76 |
| Intensity, Sp. flux, MFCC without 0th | 74 | 76 |
| Intensity, Sp. flux, MFCC without 0th, Dur. | 73 | 76 |

TABLE IV: Average F-scores in percent for "MTACK vs other" classification for all the "ideal VA Detector talk spurts" in the development set.

| max latency (ms) | Features | Avg. F-score |
|---|---|---|
| 300 | Intensity, flux, mfcc without 0th | 73 |
| 500 | Intensity, flux, mfcc without 0th, dur | 76 |

TABLE V: Average F-scores in percent for "MTACK vs other" classification for all the "ideal VAD talk spurts" in the evaluation set.

*2) Results And Discussion:* As expected, we observe in Table IV that MFCCs and duration, at least in the 500ms case, are the main contributors to the distinction between MTACK vs. NONMTACK. The combination of features did not always yield better results. However, note that we only tried a combination of features on feature-level, and that a decision-level fusion might yield better results (which will be investigated in future work). We observe that omitting the 0th DCT for MFCCs, does not hurt performance. Table V shows results for the proposed feature combinations on the evaluation set. Surprisingly little gain is achieved by using the longer maximum latency of 500 ms as compared to 300 ms. Table VI shows the results for the more realistic 'OpenSmile VAD talk spurts'. A small performance drop is observed. Furthermore, the confusion matrix in Table VII shows that it is easier to miss a LR than to miss a NON-LR.

| max latency (ms) | Features | Avg. F-score |
|---|---|---|
| 300 | Intensity, flux, mfcc without 0th | 68 |
| 500 | Intensity, flux, mfcc without 0th, dur | 69 |

TABLE VI: Average F-scores in percent for "MTACK vs other" classification for the 'OpenSmile VAD talk spurts' in the evaluation set.

| | | Classified as | |
|---|---|---|---|
| | | MTACK | NONMTACK |
| True | MTACK | 279 | 258 |
| True | NONMTACK | 171 | 967 |

TABLE VII: Confusion matrix of 500-Intensity-flux-mfcc-without-0th-dur, evaluated on evaluation set

### *F. Classifier IIa:* COMPETITIVE *vs.* COOPERATIVE

This task is based on the theoretical distinction between COMPETITIVE vs. COOPERATIVE incoming speech. The classifier was trained on agreed annotations made by two human annotators who labelled a part of the HCRC Map Task Corpus on perceived COMPETITIVENESS and COOPERATIVENESS of the incoming overlapping speech (as explained in Section V-A).

*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 ([6], [26]). Speaking rate has been studied in [27] where it was noted that competitive overlappers make use of higher speaking rates. However, [28] found speaking rate to be a weak cue for competitive speech. 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 [29]. While average F0 (high) has shown to be a cue for interruption (e.g., [6]), it requires adaptive estimation of F0 range and is not considered here. As shown in the analysis in Section VII-G, talkspurt duration is a good feature. Based on the experience from annotation, we noted a tension in the voice for some interruptions and competitive speech. Thus, voice quality correlates may be useful for this task. Voice quality was measured by spectral centroid, spectral kurtosis, and spectral skewness. The final set of acoustic features was comprised of:

- F0: DCT 1-6
- Intensity: DCT 1-6
- Duration: For training, the full talk spurt duration was used. For testing, the duration up to the maximum latency threshold was used.
- Spectral Flux: 0th DCT
- Voice quality: 0th DCTs of spectral centroid, spectral kurtosis and spectral skewness

TABLE VIII: Average F-scores for predicting Comp vs Coop on development set using "ideal VAD talk spurts" from the corpus

| Max lat.(ms) | 300 | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|---|
| F0 | 54 | 57 | 58 | 57 | 57 |
| Int. | 56 | 53 | 59 | 56 | 55 |
| Sp. Flux | 63 | 61 | 60 | 60 | 58 |
| V.Q. | 53 | 51 | 53 | 51 | 52 |
| dur. | 46 | 47 | 48 | 51 | 51 |
| Comb1 | 57 | 52 | 54 | 55 | 58 |
| Comb2 | 58 | 52 | 54 | 55 | 57 |

TABLE IX: Average F-scores for predicting Comp vs Coop on evaluation set using "ideal VAD talk spurts" from the corpus

| Max lat.(ms) | 300 | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|---|
| Sp. Flux | 61 | 63 | 59 | 63 | 55 |
| Comb1 | 58 | 54 | 53 | 54 | 58 |
| Comb2 | 57 | 53 | 54 | 53 | 58 |

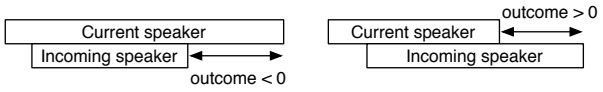



Fig. 7: The outcome of overlap that is to be predicted.

- Comb1: F0, Intensity, Spectral Flux, and Voice Quality, as specified above
- Comb2: As Comb1 with duration added, as specified above

*2) Experimental setup:* For training and testing the classifier, we used the COMPETITIVE and COOPERATIVE annotations that were obtained with two human annotators (see Section V-A). Only those talk spurts that were agreed upon by both annotators were included which yielded 88 and 319 talk spurts 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 (contrary to what was done for the other classifiers). There were 4 quads available. To ensure strict separation of training, development and testing sets, in each fold, 2 quads were held out for development or testing. The models trained for optimization of the SVM parameters were trained with the other 2 quads. All possible combinations of quads with strict separation of training, development, and testing sets were made which yielded 12 folds for the optimization phase. For final testing, the quad initially used for development was added to the training set, which yielded 4 final folds for testing.

*3) Results And Discussion:* Table VIII shows the results for the development data and Table IX for the evaluation data. It is clear that only spectral flux is the only feature which gives anything above chance level. It is hard to speculate on the reason for this, but it should be pointed out that data sparseness, i.e. very few competitive samples, may have contributed to this.

### G. Classifier IIb: Outcome of Overlap

The observed outcome from overlap is defined by a contextual timing feature. This feature is the end-time of the talk-spurt for the speaker who intercept in the overlap subtracted by the end-time of the talk-spurt of the interlocutor, which is the speaker who talked before the overlap. Thus, this feature measures the outcome of the overlap, i.e the winner of the floor, and is hence denoted as the outcome. This is illustrated in Figure 7. Based on the outcome, the following labeling scheme is applied:

```
If outcome < 0 then
label as incoming speaker stops;
else
label as incoming speaker continues.
```

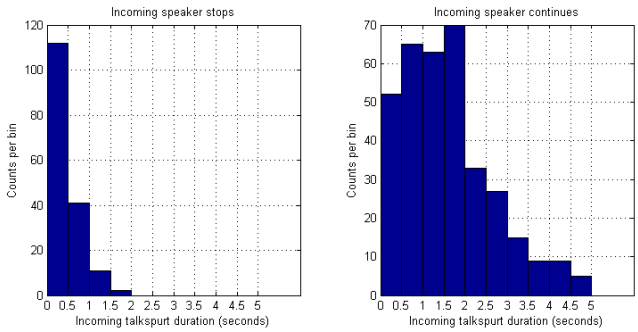



Fig. 8: Durations of talkspurts in overlap with no MTACK context (within the overlap). To the left is when the incoming speaker stops, and to the right is when the incoming speaker continues.

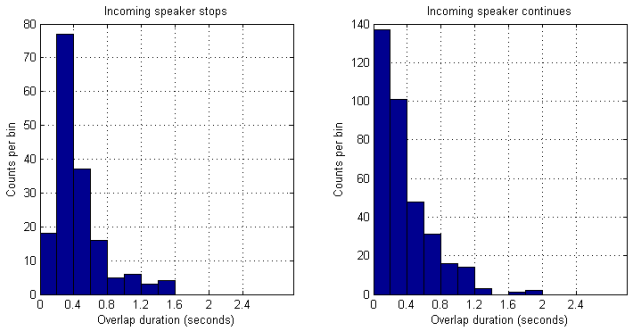



Fig. 9: Durations of overlaps with no MTACK context (within the overlap). To the left is when the incoming speaker stops, and to the right is when the incoming speaker continues.

By using this rule, instead of human annotations of interruptions, or competitive and cooperative speech, the resulting labels are always consistent and objective. If the labels generated by the rule may be predicted using acoustic cues, then the predicted outcome from the overlap can be forwarded to the dialog manager, which in turn can make a decision. In this way, we can think of the rule as an observed habit which may be predicted. However, the labels produced by the rule has no correspondence with the labels derived from annotation, the average F-score is 41.8.

Theoretically, one would expect a relation between the outcome of the overlap described here, on the one hand, and the concept of COMPETITIVE vs COOPERATIVE described earlier, on the other hand: the Speaker will probably more often stop speaking due to incoming COMPETITIVE Responses than due to COOPERATIVE Responses. Figures 8 and 9 show the histograms of the talk spurt durations and the overlap durations for the two possible outcomes of overlap. Compare these with Figures 4 and 5 to see that at least in this respect, there is a relation between observed outcome of overlap, and the manual annotation of COMPETITIVE vs. COOPERATIVE.

*1) Acoustic Features:* The final acoustic feature set is:

- F0: DCT 1-6
- Intensity: DCT 0-6 or 1-6
- Duration: For training, the full talk-spurt duration was used, for testing, the duration up to the maximum latency threshold was used.
- Spectral flux: 0'th DCT

TABLE X: Development set Average F-scores for predicting outcome of overlap given the "ideal VA talk spurts"

| Max lat.(ms) | 300 | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|---|
| F0 | 56 | 66 | 65 | 67 | 69 |
| Int. | 58 | 63 | 61 | 59 | 63 |
| Int. + 0th | 63 | 63 | 61 | 64 | 66 |
| sp. Flux | 61 | 62 | 62 | 62 | 64 |
| v.q. | 58 | 62 | 65 | 65 | 64 |
| dur. | 70 | 75 | 76 | 76 | 76 |
| comb1 | 57 | 62 | 66 | 66 | 66 |
| comb2 | 54 | 66 | 71 | 73 | 74 |
| comb1 rs | 58 | 63 | 63 | 67 | 65 |
| comb2 rs | 60 | 63 | 69 | 77 | 71 |

TABLE XI: Evaluation set Average F-scores for predicting outcome of overlap given the "ideal VA talk spurts"

| Max lat.(ms) | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|
| dur. | 77 | 79 | 79 | 79 |
| comb1 | 52 | 54 | 55 | 58 |
| comb2 | 62 | 67 | 70 | 63 |
| comb1 rs | 53 | 60 | 56 | 61 |
| comb2 rs | 58 | 71 | 77 | 74 |

- Voice Quality: 0'th DCTs of spectral centroid, spectral kurtosis and spectral skewness.
- Comb1: F0,Intensity, Spectral flux and Voice Quality, as specified above
- Comb2: As Comb 1 with duration added, as specified above
- Comb1: Comb 1 with rescaling
- Comb2: Comb 2 with rescaling

The DCT coefficients are computed as described in Section VII-B.

*2) Results And Discussion:* The results, measured by average F-scores, for optimal parameters on the development set given the "ideal VA talk spurts" are shown in Table X. It is clear that performance increases with the maximum latency duration threshold. Adding the 0'th DCT coefficient to Intensity gives some benefit, but it is not included in the combined feature set since it might be sensitive to recording conditions. Duration is the most salient feature overall while the other features gives similar contributions. Rescaling does not show any clear advantage. Eventually, we decided to evaluate the combined feature sets, with and without rescaling, and, finally, duration alone.

The results for the evaluation set are given in Table XI. These results verify that classifier performance increases with the maximum latency duration threshold. Rescaling gives a clear advantage, but the comb2 feature set does not beat duration alone. Especially, the results the comb1 feature set (acoustic features only), are not very strong but clearly above chance for longer maximum latency thresholds.

Then we made the evaluation using the "OpenSmile VA talk spurts", the performance dropped significantly. The cause was hypothesized to be inconsistent segmentation by the energy based voice activity detector. Since he trajectory parametrization by DCT coefficients is likely to be sensitive to segmentation inconsistencies, we decided to only use the 0th DCT coefficient (i.e. corresponds to the arithmetic average). However, this ruled out using F0 and Intensity as features since the arithmetic average of these are dependent on the speaker and the distance between the speaker and the microphone. Consequently, we ended up using the 0th coefficients of Spectral Flux and Voice Quality. The results are shown in Table XII. It is clear that the acoustic features does not perform above chance, leaving only duration as a reliable feature.

TABLE XII: Evaluation set Average F-scores for predicting outcome of overlap given the "OpenSmile talk spurts"

| Max lat.(ms) | 500 | 700 | 900 |
|---|---|---|---|
| dur. | 66 | 71 | 69 |
| comb1 | N/A | 48 | 46 |
| comb2 | 54 | 54 | 49 |

### *H. Classifier IIc: Hybrid approach*

The pragmatic approach in Section VII-G doesn't produce automatic labels that relate to the labels from the annotation. This section describes an attempt to derive a low complexity rule which shows agreement with the labels derived from the human annotations.

Similar to the pragmatic approach in Section VII-G, two types of contextual timing features are defined first. The first one is the duration of the overlap. The second is the end-time of the talk-spurt for the speaker who intercept in the overlap subtracted by the end-time of the talk-spurt of the interlocutor, which is the speaker who talked before the overlap. Thus, this feature measures the outcome of the overlap, i.e the winner of the floor, and is hence denoted as outcome.

To derive a rule from the features a decision tree was used, where the priors for the agreed labels were set to a uniform distribution. The first two rules, at the top of the tree, was:

```
If overlap > 0.15 and outcome < −0.40 then
label as competitive;
else
label as cooperative.
```

This label scheme achieved an average F-score with our agreed labels of $0.67$. The value is above chance and should be compared to the kappa which is decent but not high. Rules with higher complexity may be derived by looking further down into the tree, but these high complexity rules are difficult to explain and understand.

The part of the rule which concerns the amount of overlap, i.e. a minimum overlap of 150 ms, may be interpreted as the minimum duration of a perceivable overlap. Thus, if the overlap is below 150ms it is not perceivable and hence no interruption is perceived either. It should be noted, that despite no listener responses, as defined by the acknowledgments moves, were included for annotation, more than a few listener responses where observed during annotation. These also tend to come immediately before the end of the talk-spurt, possibly within the last 150 ms. Since listener responses are considered as cooperative, the occurrences of these just toward the end of a talk-spurt may be another explanation for this criterion. In any case, a speaker change that is within 150ms before the end of the talk-spurt may simply be considered as a smooth speaker shift. Also, this criterion seem to be non-negligible, since if this part of the rule was removed, the average F-score dropped below chance level. The second part of the rule; outcome $< -0.40$; simply states that the speaker who intercepts in overlap has to speak for 400 ms after the overlap in order to consider it to be competitive. Finally, we notice that the rule implies a minimum talk-spurt duration of $150 + 400 = 550$ ms. We further notice from Figure 2, that listener responses are more likely to be shorter than 500ms compared to non-listener responses. This confirms the findings by [30], where duration was found to be a highly reliable feature for back-channels.

Figures 10 and 11 show the histograms of the talk spurt durations and the overlap durations for the labels generated by the rule. We notice a greater similarity between these histograms and the histograms for the manual annotations (Figures 4 and 5) compared

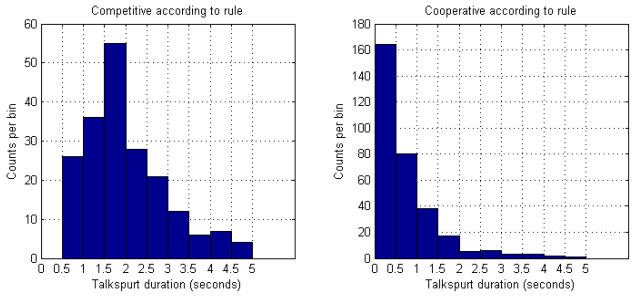



Fig. 10: Durations of talkspurts in overlap with no MTACK context (within the overlap). To the left are COMPETITIVE and to the right COOPERATIVE Responses, both according to the rule.

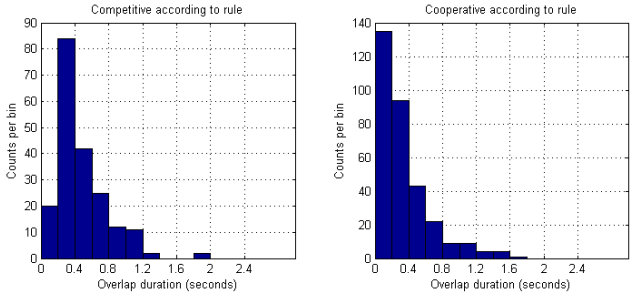



Fig. 11: Durations overlaps with no MTACK context (within the overlap). To the left are COMPETITIVE and to the right COOPERATIVE Responses, both according to the rule.

to the histograms which results from the pragmatic approach. This is especially true for the overlap durations of competitive speech.

To summarize, the motivations for using the hybrid rule are:

1) The rule extracts labels which have some consistency with our human annotations.
2) The rule generates labels which have overlap and duration distributions similar to the human annotations.
3) We can generate labels for more data than what is provided by the annotations.
4) The rule is always consistent and objective.

If the labels generated by the rule may be predicted using acoustic cues, then the predicted labels can be forwarded to the dialog manager, which in turn can make a decision. In this way, we can think of the rule as an observed habit which is also related to cooperative and competitive speech, which may be predicted.

*1) Results And Discussion:* For this classifier, we use exactly the same feature set as for the pragmatic approach (Section VII-G).

TABLE XIII: Development set Average F-scores for predictingCOMPETITIVE speech based on the hybrid approach given the "ideal VA talk spurts"

| Max lat.(ms) | 300 | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|---|
| F0 | 57 | 64 | 64 | 66 | 67 |
| Int. | 61 | 67 | 63 | 64 | 67 |
| Int. + 0th | 61 | 64 | 63 | 69 | 72 |
| sp.flux | 63 | 66 | 64 | 62 | 62 |
| v.q. | 62 | 64 | 67 | 68 | 69 |
| dur. | 41 | 71 | 79 | 79 | 82 |
| comb1 | 61 | 67 | 66 | 64 | 67 |
| comb2 | 50 | 58 | 62 | 65 | 67 |
| comb1 rs | 60 | 64 | 66 | 72 | 70 |
| comb2 rs | 55 | 58 | 69 | 75 | 76 |

TABLE XIV: Evaluation set Average F-scores for predicting COMPETITIVE speech based on the hybrid approach given the "ideal VA talk spurts"

| Max lat.(ms) | 500 | 700 | 900 | 1100 |
|---|---|---|---|---|
| dur. | 67 | 74 | 70 | 81 |
| comb1 | 57 | 57 | 60 | 60 |
| comb2 | 55 | 61 | 55 | 62 |
| comb1 rs | 58 | 57 | 58 | 62 |
| comb2 rs | 56 | 63 | 61 | 67 |

TABLE XV: Evaluation set Average F-scores for predicting COMPETITIVE speech based on the hybrid approach given the "OpenSmile talk spurts"

| Max lat.(ms) | 500 | 700 | 900 |
|---|---|---|---|
| dur. | 57 | 63 | 67 |
| comb1 | 52 | 53 | 49 |
| comb2 | 48 | 42 | 47 |

The results, measured by Average F-scores, for optimal parameters on the development given the "ideal VA talk spurts" are shown in Table XIII. The F-scores pretty much follows the same pattern as for the pragmatic approach (Section VII-G), but the observations are rephrased where for clarity with few but some differences. It is clear that classifier performance increases with the maximum latency duration threshold. Adding the 0'th DCT coefficient to Intensity gives some benefit, but it is not included in the combined feature set since it might be sensitive to recording conditions. Duration is the most salient feature overall while the other features gives similar contributions. Rescaling does show an advantage for maximum latency threshold of 700 ms and above. Eventually, we decided to evaluate the combined feature sets, with and without rescaling and finally duration alone.

The results for the evaluation set are given in Table XIV. These results verify that classifier performance increases with the maximum latency duration threshold. Rescaling gives a clear advantage, but the comb2 feature set does not beat duration alone. Especially, the results the comb1 feature set (acoustic features only), are not very strong but clearly above chance for longer maximum latency thresholds.

For the evaluation using the "OpenSmile VA talk spurts", we adopted the same procedure as for the pragmatic approach. Thus, we ended up using the 0th coefficients of Spectral Flux and Voice Quality along with duration. The results are shown in Table XII. It is clear that the acoustic features does not perform much above chance, leaving only duration as a reliable feature.

### *I. Conclusions from Classification Experiments*

These series of experiments has shown successes and failures. First of all, **Classifier I** (Classification of all Responses into MTACK / NONMTACK) has a clear potential in a fielded system. For the **Classifier II b/c** versions, we have shown some success for acoustic features by using "ideal VA talk spurts". However, under the more realistic condition where "OpenSmile talk spurts" are used, only duration showed to be a reliable feature. It is not obvious to chose between **Classifier IIb** and **Classifier IIc**, mainly because the actual performance is similar, but the more more pragmatic **Classifier IIb** may be the choice since it does not rely on human judgments. 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, which usually are the ones which has a higher maximum latency.

## VIII. Existing Models for Behavior Generation and Specification

Here we describe Elckerlyc, the BML Realizer used to generate virtual human behavior. It is based on the SAIBA Framework [31] (see Fig 12, which describes a generic architecture for virtual human applications. It contains a three-stage process: *communicative intent planning*, multimodal *behavior planning*, resulting in a BML stream, and *behavior realization* of this stream. The Elckerlyc framework used in this project encompasses the realization stage. It takes a specification of the intended behavior of a virtual human written in the Behavior Markup Language (BML) [31] and executes this behavior through the virtual human.


Fig. 12: The SAIBA framework.

The BML stream contains BML requests with behaviors (such as speech, gesture, head movement etc.) and specifies how these behaviors are synchronized (see also Fig. 13). 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. 14) to synchronization points in another behavior. BML can be used to append or merge new behaviors into a running BML stream. Some extension have been proposed to allow the specification of instant removal of a running BML request[1].

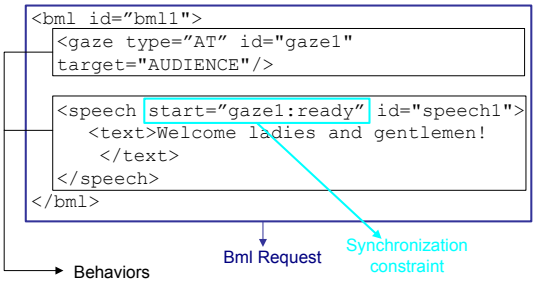



Fig. 13: An example of a BML request containing a gaze and a speech behavior. A synchronization constraint ensures that the speech starts after the gaze is aimed at the audience.

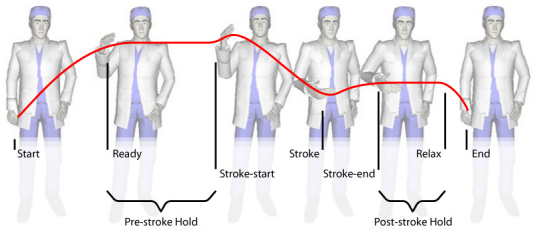



Fig. 14: Standard BML synchronization points (picture from http://wiki.mindmakers.org/projects:bml:main)

## IX. Scheduling and Planning for Continuous Interaction

Currently, BML does not contain mechanisms to slightly modify behavior that is already running, or to interrupt behavior in a more graceful manner. Such mechanisms are crucial to achieve continuous interaction [32]. Some desired changes to planned behavior are only on their timing or parameter values (speak louder, increase gesture amplitude) and should not lead to completely rebuilding the animation or speech plan. Such small adaptations of the timing or shape of planned behavior occur in conversations and other interactions [2]. Elsewhere, we discuss the specification and Elckerlyc's implementation mechanisms that allow such small behavior plan changes to occur instantly [32]. In this paper we focus on graceful interruption and preplanning of behavior that were developed during the eNTERFACE workshop.

We have defined a custom BML extension BMLT [2] to allow the expression of behaviors and the scheduling and interruption mechanisms discussed above that cannot be expressed in BML (yet).

### A. Preplanning

Planning a BML request typically takes a non-neglectable 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 like the one described in this paper. Elckerlyc explicitly models the scheduling stage of BML requests and makes it transparent to the Behavior Planner by providing it with feedback on when the scheduling of a BML request is started and when it it done. BMLT 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 request, using the BMLT preplan attribute in that request. Preplanned BML requests can be activated using another BML request with an onStart attribute. The preplanned behavior is activated as soon as the scheduler finishes planning the behavior with the `onStart` that activates it. Example 1 illustrates the BML used for preplanning.

**BML Example 1** Several BML requests illustrating the preplanning and activation of pre-planned behavior.

```
<bml xmlns:bmlt="http://hmi.ewi.utwente.nl/bmlt"
  id="bml1" scheduling="merge" bmlt:preplan="true">
...
</bml>
```

**(a)** Preplan bml1.

```
<bml xmlns:bmlt="http://hmi.ewi.utwente.nl/bmlt"
  id="bmlX"
  bmlt:onStart="bml1"/>
```

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

```
<bml id="bml3"
xmlns:bmlt="http://hmi.ewi.utwente.nl/bmlt"
  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.

### B. Graceful interruption

The interrupt behavior, first proposed and implemented in the SmartBody BML realizer [33], is used to interrupt a running BML request. This can be used to schedule the interrupt of a BML request relative to some other behavior (e.g. VH looks at the interlocutor

[1] See http://wiki.mindmakers.org/projects:bml:multipleblockissue

[2] See http://wiki.mindmakers.org/projects:bml:bmlt

before it stops to speak). In both BMLT and the SmartBody BML, the interrupt behavior by default immediately interrupts all behaviors in the BML request it targets at the start of the interrupt behavior.

In its simplest form (See Example 2), the BMLT interrupt behavior acts the same as the SmartBody interrupt behavior. The syntax is also very similar.

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

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

We have extended the interrupt behavior to allow a more fine-grained interrupt specification, using the `interruptspec` element inside an `interrupt` behavior. Using the `interruptspec` we can define exactly when certain behaviors inside the target BML request are to be interrupted. All behaviors in the target BML request that are not described in an `interruptspec` are interrupted instantly. The `interruptspec` also allows us to specify preplanned BML requests that are to be activated as soon as a certain behavior is interrupted using the `onStart` attribute. This combination of the interruption behavoir and preplanning allows us to specify the graceful interruption of behavior in other BML blocks, with alternative continuations after the interruption (See Example 3).

**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`).

```
<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>
```

## X. Listener Response Elicitation

Before going into monitoring and handling Responses it is important that the system is able to elicit these Responses. In human-human conversation the speaker often elicits such responses. The speaker creates Response opportunities through vocal and non-vocal cues, such as pausing between statements, modifying the prosody of the speech, and using gaze and face expressions. This section discusses the literature in order to find possibilities for response elicititation cues that can be used in our pilot experiment.

Prosodic elicitation cues for responses are quite well described in literature. Gravano and Hirschberg [34] observe that the final intonation of the interpausal unit (IPU) preceding a response rises in 81% of the cases. Furthermore the mean intensity and pitch level of the preceding IPUs which are followed by a response are higher than IPUs not followed by a response. Furthermore Ward and Tsukahara [35] use in their handcrafted rule based model an area of 110ms of low pitch to predict a response 700ms 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. [36] conclude that 83% of listener responses in their corpus occur during mutual gaze, confirming earlier intuitions of Kendon [37] and Duncan Jr. [38]. Furthermore, head movements have been associated with eliciting responses [39], but there are, to our knowledge, no concrete findings directly applicable to virtual humans.

We performed an observatory study on the MultiLis corpus, where we analyzed the speakers who elicited the most responses from the listeners, with special attention to their nonverbal behaviors. Some speakers were very expressive in their nonverbal behavior, while others were not. For one of the speakers his blinking behavior really stood out. In general his blinking rate was high, but at the end of statement, where he expected a response from the listener, he stopped blinking and stared at the listener. He started blinking again as soon as the listener provided a response.

### A. Enhancing MARY TTS to realize vocal elicitation cues

The MARY TTS platform is an open-source, modular architecture for building text-to-speech systems, including unit selection and statistical parametric waveform synthesis technologies. It has been described in detail elsewhere [40], [41]. The present paper only describes the aspects relevant in the current context. One of those aspects is how to realize vocal elicitation cues using MARY TTS. Prosody modification techniques are the key to realize vocal elicitation cues. Traditionally in MARY, the applications that require control over prosody were using MBROLA diphone synthetic voices, though the voices are unnatural. Nowadays HMM-based voices are reaching high quality synthetic speech.

In HMM-based speech synthesis, trained statistical models (context-dependent HMMs) are used to predict duration and generate parameters like mel-cepstral coefficients, log F0 values, and bandpass voicing strengths using the maximum likelihood parameter generation algorithm including global variance [42]. In the later stages, F0 parameters, bandpass voicing strengths, and the five bandpass filters are used to generate a mixed excitation signal. Finally, speech is synthesized from the mel-cepstral coefficients and the mixed excitation signal using the MLSA filter [43].

Although MARY already supports realization of predicted prosody parameters using HMM synthesis, it did not support explicit prosody specification. This project requires support for prosody modifications specified in MARYXML requests. So, as part of this project, we implemented support for 'prosody' element as described in W3C Speech Synthesis Markup Language (SSML) recommendations; and the different attributes in 'prosody' element like 'rate', 'pitch' and 'contour' are used as specifications to modify predicted phone durations and pitch contour before passing them to the HMM synthesizer. Once the modifications are done according to given specifications, they are realized as normal with HMM-based synthesis strategies.

**MARYXML Example 1** An example which supports explicit prosody specifications

```
<?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>
```

## XI. Pilot Experiment

As a setting for our experiments we chose the route description domain. This domain was chosen since in this domain the fact whether the information given by the agent has reached the user,

and is understood by the user or not, is crucial to the success of the interaction. In this setting, continuous monitoring of the user and reacting appropiately to their responses is very relevant. You may want to repeat certain elements of the explanation to get your point across or skip a part depending on the actions of the user.

Before going into monitoring and handling the responses it is important that your system is able to elicit these responses. In human-human conversation the speaker often elicits such 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. In this experiment we aim to recreate these signals based on literature and corpus analysis and evaluate them in our agent to see which elicitation strategy elicits the most responses. Furthermore we assess each version of our agent on subjective measures related to conversational skill, rapport, personality etc.

### A. Task

During the experiment our route giving agent explains a route to the participants. Afterwards the participant needs to draw the route on a map, which is presented before the interaction begins.

### B. Stimuli

The map contains the layout of a fictional city. 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 agent to identify the landmarks. The current position of the participant is also shown on the map.

There are three different starting points, for three different routes. Each route consists of $n$ steps[3] that take the user to their final destination. Each step is realized by specifying a BML block. The BML block specifies the speech and the behavior the agent performs. The speech is synthesized using Mary TTS [41]. The speech is manually cleaned up, using the prosody tags described in Section X. We removed, where necessary, peculiarities in the synthesized speech, added some extra pause moments and changed the speech rate, to make the agent sound more natural. Aligned with the speech, gestures 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 1.5s, 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 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 [34]. 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 300ms 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 nonverbal inviting behavior found in the MultiLis Corpus [13]. 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 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 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 pretty high blinking rate, so this actually stands out) and stares at the listener. As soon as a 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.

### C. Methodology

We invited 9 participants (8 male, 1 female, aged between 25 and 54, all non-native English speakers) to interact with our route agent. Participants are told that the agent is able to perceive and react to short vocal and nonverbal 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 is taken away before the interaction starts. During the interaction the route agent 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 agent. During each interaction the agent explained a different route. Each route description was given with a different elicitation strategy. Every participant interacted with the *Default* and *Combined* agent and either the *Vocal* or the *Nonverbal* agent. Permutations of routes and elicitation strategies were varied among participants.

### D. 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 agent and the participant, based on the questionnaire used in De Kok and Heylen [13]. Furthermore we measured the perceived impression of the agent by having the participants rate the agent on 26 bipolar semantic differential adjective scales taken from the study of Ter Maat et al. [44]. All questions are on a 7-point Likert scale.

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

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

### E. Results and Discussion

We succesfully elicited responses from the subjects (see Table XVI). The amount of response given seems highly subject dependent (see Table XVI). Over half of the subjects gave a 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

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

| subject | default | combined | vocal | nonverbal | average |
|---|---|---|---|---|---|
| 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 |

TABLE XVI: Response ratio (Responses given/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.

| | Default | Combined | Vocal | Nonverbal |
|---|---|---|---|---|
| Like best: | 5 (56%) | 3 (33%) | 0 (0%) | 2 (50%) |
| In between: | 2 (22%) | 4 (44%) | 1 (20%) | 1 (25%) |
| Like least: | 2 (22%) | 2 (22%) | 4 (80%) | 1 (25%) |
| Most natural: | 5 (56%) | 2 (22%) | 1 (20%) | 1 (25%) |
| In between: | 2 (22%) | 3 (33%) | 1 (20%) | 3 (75%) |
| Least natural: | 2 (22%) | 4 (44%) | 3 (60%) | 0 (0%) |
| Most social: | 5 (56%) | 3 (33%) | 1 (20%) | 0 (0%) |
| In between: | 2 (22%) | 4 (44%) | 1 (20%) | 3 (75%) |
| Least social: | 2 (22%) | 2 (22%) | 3 (60%) | 1 (25%) |
| Most attentive: | 5 (56%) | 3 (33%) | 0 (0%) | 1 (25%) |
| In between: | 2 (22%) | 5 (56%) | 1 (20%) | 1 (25%) |
| Least attentive: | 2 (22%) | 1 (11%) | 4 (80%) | 2 (50%) |

TABLE XVII: Results of the post-questionnaire in which the participants ranked the agents on likeability, naturalness, social ability and attentiveness. Especially the agent with the *Vocal* elicitation strategy performs bad on these scales. The *Default* agent seems best.

elicitation cue. Only 6 out of 237 responses were non-verbal only. 137 were both verbal and nonverbal.

We observed the use of one or more repetions in the responses of five of the subjects (cf. Interaction Example 2).

**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 XVII we notice the bad performance of the agent with the *Vocal* elicitation strategy. Most of the five participant that interacted with this agent rated it the lowest on all scales. The prosodic modifications to the speech to elicit responses should thus be improved. Now they are 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 questionnaire after each session did not yield any insightful results.

### F. Lessons learned

From the results of the pilot we learned that several improvements can be made to the setup. First we want to expand the experiment with a fourth route. This was always our intention, in order to let every participant interact with every elicitation strategy, but due to time constraints we decided to drop one of the routes for the pilot.

Furthermore the vocal elicitation strategy needs some work. On the postquestionnaire it was consistently rated as the least likable, natural, social and attentive of the four strategies. Since the vocal elicitation strategy is also included in the combined strategy, it probably had a negative impact on that condition as well.

Finally, we want to vary the pause between two sentences, since pause in itself is also a response elicitation cue [45], [46]. At this moment this pause is 1.5 seconds, based on the average pause in the MultiLis Corpus. We see in our data that in almost every response opportunity we explicitly created, we get a response. We suspect that the length of the pause is such a strong cue that this dominates our four different strategies and is the cause for this.

## XII. Discussion and Conclusions

In this Enterface workshop, we have developed a virtual human that is able to interact with a 'real' subject in an continuous manner. That is: being capable of interaction in which all partners perceive each other, express themselves, and coordinate their behavior to each other, continually and in parallel. The project 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 interuption, automatic real-time classification of listener responses and models for appropriate reactions to listener responses. We have set up a pilot experiment in which a virtual human interacts with a subject. The aim of the experiment was to elicit Response behavior, to provide us with more information on what user responses occur, and to serve as inspiration for further interaction models.

In this experiment, we have observed that some Responses given by our subjects are much shorter than the waiting time between steps; other Responses are much longer. Furthermore, Responses are not given at every Response Oppurtunity. Starting to speak through a repetition or waiting for a Response that is already finished confused some of our subjects. In a responsive version of the virtual human, we should add dynamic pauses: if no Response comes, continue speaking after a smaller wait. If feedback comes, the virtual human can wait until Response is finished. If a Response is cooperative it often makes sense to immediately continue speaking in overlap.

We have observed several repetitions from the listener, related to speech from the speaker. Detecting such repetitions is still an open issue. Since the repetitions often repeat the landmarks used in the route, perhaps the occurence of landmarks (as detected by a keyword spotter) could be used as one of the cues for the identification of repetions. Assumed that we can automaticly asses whether a response is a repetition, the preplanning mechanisms we have developed during the workshop 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 route description after the acknowledgements is already

planned, Elckerlycs retiming mechanisms (see [47]) can be used to shift it in time so that a full replan of the route description is avoided.

Interruptions are detected as Competitive Responses by our classifier. If the subjects interrupts the Virtual Human (as in Interaction Example 4), his ongoing route description can be gracefully interrupted using mechanisms discussed in Section IX-B. 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 to handle 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 a insecure virtual human could explicitly wait for feedback after each of its utterances. Some of this strategies can potentially already be realized with the existing system (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 [32] for a more elaborate discussion on this topic).

## XIII. Deliverables

The project has resulted in several software components, corpora and annotations, that will be made available to the public:

1) Automatic, real-time classifiers for Responses, implemented as openSMILE components [4]
2) The addition of cooperative/competative annotations in the MapTask corpus [12]
3) A motion capture corpus containing over 100 gestures related to route-giving [5]
4) Extensions that allow prosody modification in HMM voices in the open source speech synthesis system Mary, these will be included in its new release [6]
5) Several extensions and tools for the open source virtual human platform Elckerlyc [7], which will be included in its next release:
   a) A generic WoZ interface framework that allows the set up of Wizard of Oz experiments with Elckerlyc in an easy and flexible manner
   b) Implementation of preplanning and scheduling algorithms that allow gracious interruption of ongoing behavior
   c) Integration of Elckerlyc with the open source SEMAINE api [11], an open source middleware framework that allows easy connection of different modules in emotion-oriented systems.
6) An annotated video corpus of user-interactions with our virtual human during the pilot experiment

[4]http://sourceforge.net/projects/opensmile/
[5]Freely available at http://hmi.ewi.utwente.nl/mocapdb
[6]Available at http://mary.dfki.de/
[7]http://hmi.ewi.utwente.nl/showcases/Elckerlyc

## Acknowledgment

The authors would like to thank the project advisors, Anton Nijholt, Dirk Heylen, and Stefan Kopp, for their support, Marc Schröder for enjoyable discussions and useful tutorials, Albert Ali Salah for organizing eNTERFACE'10, Ronald Poppe and Mark ter Maat for conceptual and practical support, and the GATE project and the NoE SSPNet for sponsoring this project,.

## References

[1] K. R. Thórisson, *Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action*, ser. Multimodality in Language and Speech Systems. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002, pp. 173–207.
[2] A. Nijholt, D. Reidsma, H. van Welbergen, H. op den Akker, and Z. M. Ruttkay, "Mutually coordinated anticipatory multimodal interaction," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, ser. Lecture Notes in Computer Science, A. Esposito, N. G. Bourbakis, N. Avouris, and I. Hatzilygeroudis, Eds., vol. 5042. Berlin: Springer Verlag, 2008, pp. 70–89.
[3] D. T. Fujimoto, "Listener responses in interaction: A case for abandoning the term, backchannel," *Journal of Osaka Jogakuin 2year College*, vol. 37, pp. 35–54, 2007.
[4] D. Neiberg and J. Gustafson, "The prosody of Swedish conversational grunts," in *Proc. of Interspeech*, Sep. 2010.
[5] V. Manusov and A. R. Trees, ""are you kidding me?": The role of nonverbal cues in the verbal accounting process," *Journal of Communication*, vol. 52, no. 3, pp. 640–656, Sep. 2002.
[6] P. French and J. Local, "Turn-competitive incomings," *Journal of Pragmatics*, vol. 7, pp. 17–38, 1983.
[7] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, no. 6, pp. 941–952, 2000.
[8] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, no. 1, pp. 62–81, 2004.
[9] C. Goodwin, "Between and within: Alternative sequential treatments of continuers and assessments," *Human Studies*, vol. 9, no. 2-3, pp. 205–217, 1986.
[10] ——, *Conversational Organization: interaction between speakers and hearers*. Academic Press, 1981.
[11] M. Schröder, "The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems," *Advances in Human-Computer Interaction*, vol. 2010, 2010.
[12] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty-Sneddon, S. Garrod, S. Isard, J. C. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC Map Task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.
[13] I. de Kok and D. Heylen, "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*, 2010.
[14] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *The Bell System Technical Journal*, vol. 47, pp. 73–91, 1968.
[15] J. C. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.
[16] D. Reidsma, "Annotations and subjective machines — of annotators, embodied agents, users, and other humans," Ph.D. dissertation, University of Twente, Oct. 2008.
[17] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics and Cognition*, vol. 14, no. 1, pp. 129–182, 2006.
[18] F. Eyben, M. Woellmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, to appear.
[19] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[20] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in Swedish," in *DiSS-LPSS Joint Workshop 2010*, Sep. 2010.

[21] Y. Ariki, S. Mizuta, M. Nagata, and T. Sakai, "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum," *Communications, Speech and Vision*, vol. 136, no. 2, pp. 133–140, Apr. 1989.

[22] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2003.

[23] D. Neiberg, P. Laukka, and G. Ananthakrishnan, "Classification of affective speech using normalized time-frequency cepstra," in *Prosody 2010*, May 2010.

[24] F. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the twenty-first international conference on Machine learning*, Banff, Canada, 2004.

[25] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in american english," in *Proceedings of the 16th International Congress of Phonetic Sciences 2007*, 2007, pp. 1065–1068.

[26] C. C. Lee, S. Lee, and S. S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *Proceedings of Interspeech*, 2008, pp. 1678–1681.

[27] E. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 1–63, 2000.

[28] E. Kurtic, G. J. Brown, and B. Wells, "Resources for turn competition in overlap in multi-party conversations: Speech rate, pausing and duration," in *Proceedings of Interspeech*, 2010, to appear.

[29] M. F. McKinneya, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *Journal of New Music Research*, vol. 36, no. 1, pp. 1 – 16, Mar. 2007.

[30] J. Edlund, M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg, "Very short utterances in conversation," in *Proceedings of Fonetik*, 2010.

[31] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Intelligent Virtual Agents, 6th International Conference*, ser. Lecture Notes in Computer Science, J. Gratch, M. R. Young, R. Aylett, D. Ballin, and P. Olivier, Eds., vol. 4133. Springer, 2006, pp. 205–217.

[32] H. van Welbergen, D. Reidsma, and J. Zwiers, "A demonstration of continuous interaction with Elckerlyc," in *Multimodal Output Generation*, 2010.

[33] M. Thiebaux, A. N. Marshall, S. Marsella, and M. Kallmann, "Smartbody: Behavior realization for embodied conversational agents," in *Proc. 7th International Conference on Autonomous Agents and Multiagent Systems*, 2008, pp. 151–158.

[34] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proceedings of Interspeech*, Brighton, 2009, pp. 1019–1022.

[35] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[36] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *Journal of Communication*, vol. 52, no. 3, pp. 566–580, 2002.

[37] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[38] S. Duncan, Jr., "On the structure of speaker-auditor interaction during speaking turns," *Language in society*, vol. 3, no. 2, pp. 161–180, Dec. 1974.

[39] D. Heylen, "Head gestures, gaze and the principles of conversational structure," *International Journal of Humanoid Robotics*, vol. 3, no. 3, pp. 241–267, 2006.

[40] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard Challenge 2008," in *Proc. of the Blizzard Challenge*, 2008.

[41] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[42] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[43] A. W. Black, K. Tokuda, and H. Zen, "An HMM-based speech synthesis system applied to English," in *Proc. of 2002 IEEE SSW*, Santa Monica, CA, USA, Sep. 2002.

[44] M. ter Maat, K. P. Truong, and D. Heylen, "How turn-taking strategies influence users impressions of an agent," in *Proceedings of Intelligent Virtual Agents*, Philadelphia, Pennsylvania, USA, Sep. 2010.

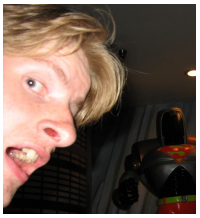
[45] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 51–58.

[46] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, May 2010.

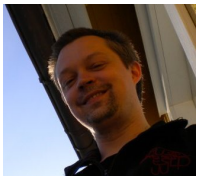
[47] H. van Welbergen, D. Reidsma, Z. M. Ruttkay, and J. Zwiers, "Elckerlyc: A BML realizer for continuous, multimodal interaction with a virtual human," *Journal on Multimodal User Interfaces*, 2010, To appear.

**Dennis Reidsma** Dennis Reidsma is a postdoctoral researcher at the Human Media Interaction group. For his PhD he worked on different aspects of natural interaction systems. He worked, among other things, on problems of annotation and reliability in large multimodal annotated corpora, in the context of the EU FP6 AMI and AMIDA projects. In addition, he worked on research and development of new interactive systems with virtual humans. The *interactive virtual dancer* attempts to invite a human to engage with her, using computer vision, music analysis, and patterns of leading and following behavior. The *interactive virtual orchestra conductor* leads an ensemble of human musicians through a musical performance using advanced interactive graphics developed at HMI and advanced music processing algorithms. His current interests are in exploring continuous interaction with virtual humans in conversational settings. He is one of the Elckerlyc developers.

**Khiet Truong** Khiet Truong is a postdoctoral researcher at the University of Twente in the Human Media Interaction group. She has a background in computational linguistics and speech technology. As a Master student, she carried out research on automatic pronunciation error detection in speech of second-language learners. In 2009, she successfully defended her PhD thesis on automatic emotion recognition in speech based on work carried out at TNO. Currently, she is working on social signal processing in the SSPNet-project.



**Herwin van Welbergen** Herwin van Welbergen received his MSc in Human Media Interaction from the University of Twente's Department of Computer Science. Currently, he is a PhD candidate at the Human Media Interaction group. His research activities focus on real-time multimodal behavior generation for virtual humans, using real-time procedural animation, real-time physical simulation and speech, especially for applications that allow continuous interaction with a virtual human. Herwin is the main developer of the Elckerlyc framework.



**Daniel Neiberg** Daniel Neiberg received a Master of Science degree in electrical engineering in 2003 from KTH (Royal Institute of Technology), Sweden. He is currently a Ph.D. student at the department TMH at KTH. His fields of interest covers automatic affective recognition, conversational interaction, prosody and acoustic-to-articulatory inversion.

**Sathish Chandra Pammi** Sathish Pammi is a Researcher and PhD student in the DFKI LT lab. He has obtained his Masters degree in Computer Science Engineering from International Institute of Information Technology (IIIT, Hyderabad, India). He has joined the DFKI Speech Group in 2007. Since 2008, he has been working in the development of Text-To-Speech (TTS) systems, including synthesis of vocal listener behavior, for Sensitive Artificial Listeners (SAL) in EU FP7 SEMAINE project. He is one of the core developers of the MARY TTS system. His current research interests are interactive speech synthesis, conversational agents and talking robots.

**Iwan de Kok** Iwan de Kok studied Computer Science at the University of Twente, receiving his MSc. in 2009 for his thesis on the influence of videoconferencing and an emotional feedback support system on polyadic negotations. During his studies he did a 3 month internship at the USC Institute for Creative Technologies, working on the multimodal prediction of listener responses. Currently he is a PhD student in the Human Media Interaction Group of the University of Twente. His goal is to create a listener response generation model for virtual humans, with a focus on the nonverbal aspect.

**Bart van Straalen** Bart van Straalen is currently pursuing a PhD at the Human Media Interaction group at the University of Twente. The main focus of his PhD research is on the role of social and emotional capabilities on the selection and generation of communicative behavior in ECAs. As part of his research he works on contextual behavior analysis, modeling of cognitive processes and dialogue systems. His research interests lie in the area of cognitive modeling, emotion appraisal, coping strategies, FML-realization, artificial intelligence and human to ECA communication.