J Multimodal User Interfaces (2014) 8:109–117
DOI 10.1007/s12193-013-0131-2

ORIGINAL PAPER

Switching Wizard of Oz for the online evaluation of backchannel behavior

Ronald Poppe · Mark ter Maat · Dirk Heylen

Received: 5 April 2013 / Accepted: 29 October 2013 / Published online: 16 November 2013

Abstract The Switching Wizard of Oz (SWOZ) is a setup to evaluate human behavior synthesis algorithms in online face-to-face interactions. Conversational partners are represented to each other as virtual agents, whose animated behavior is either based on a synthesis algorithm, or driven by the actual behavior of the conversational partner. Human and algorithm have the same expression capabilities. The source is switched at random intervals, which means that the algorithm's behavior can only be identified when it deviates from what is regarded as appropriate. The SWOZ approach is especially suitable for the controlled evaluation of synthesis algorithms that consider a limited set of behaviors. We evaluate a backchannel synthesis algorithm for speaker–listener dialogs using an asymmetric version of the framework. Human speakers talk to virtual listeners, that are either controlled by human listeners or by an algorithm. Speakers indicate when they feel they are no longer talking to a human listener. Analysis of these responses reveals patterns of inappropriate behavior in terms of quantity and timing of backchannels. These insights can be used to improve synthesis algorithms.

Keywords Wizard of Oz · Online evaluation · Social behavior synthesis · Listening behavior · Backchannels

1 Introduction

Advances in animation and sensor technology allow humans to engage in face-to-face conversations with virtual agents. One challenge is to generate the agents' appropriate, human-like behavior contingent with that of their human conversational partners. We focus on the nonverbal conversational behavior of a virtual agent, which can be specified by hand using response rules, or by machine learning models learned from annotated corpus data. In either case, there is a need to evaluate the behavior synthesis model. Offline, corpus-based analyses can be performed to determine how the model matches the data it was derived from. Such studies have led to insights in the features and models that can predict the production of behaviors in the context of a conversation. For virtual agents, it also needs to be determined how the behavior generated by these models is *perceived* by human observers. Initial steps have been taken in this direction, but using offline stimuli. The approach introduced in this paper targets the evaluation of behavior synthesis algorithms in online dialogs.

Humans are particularly sensitive to flaws in the displayed behavior, both in form and timing [16], also when virtual agents are used [19]. This effect also occurs when certain behaviors are not animated (e.g. eye gaze or respiration), which makes the behavior of the virtual agent more static. Consequently, a virtual agent's behavior is typically perceived as rather unrealistic. This is especially true in experimental settings where the behavior of the agent is varied systematically along one or a limited number of modalities. We argue that the current lack of good perceptual evaluation methods hampers progress in the design and implementation of behavior synthesis algorithms.

To this end, we propose a methodology that combines ideas behind the Turing test with those of a wizard of Oz (WOZ) setup. A behavior generation algorithm passes the

This publication was supported by the Dutch national program COMMIT. Preliminary versions of this work appeared as [22,23].

R. Poppe (✉) · M. ter Maat · D. Heylen
Human Media Interaction Group, University of Twente, P.O. Box 217,
7500 AE Enschede, The Netherlands
e-mail: r.w.poppe@utwente.nl

D. Heylen
e-mail: d.k.j.heylen@utwente.nl

human Turing test when, given limited expression channels, a human observer is not able to make out whether behavior originates from an algorithm or from a fellow human [27]. In a Wizard of Oz setup, tasks that are to be carried out by a machine are actually performed by a human, without the interacting subjects' awareness [7]. Such a setup is common during the design and evaluation of complex algorithms, for example those that produce human-like behavior.

At the heart of the Switching Wizard of Oz (SWOZ) is a distributed video-conferencing setting with two human subjects. Each subject is observed with a camera and a microphone and algorithms are employed to analyze the nonverbal behavior in real-time. These observations are used as input to a behavior synthesis model. Both subjects are shown a virtual representation of the other, animated based on one of two *sources*: (1) directly on the observed behavior of the other, or (2) on the output of a behavior synthesis model. Both sources share the same expression capabilities and limitations in terms of the type and animation of the behaviors. Subjects are not informed of the source of their partner's virtual representation. However, especially when the synthesis algorithm is less sophisticated, subjects might be aware that the behavior does not originate from their human partner. To prevent that they will not pay attention to the dialog and the behavior of the virtual representation, the system switches between the sources at random times. This forces the subjects to stay focused and to evaluate behavior in a slightly larger context.

When the displayed behavior deviates from what is regarded as human-like, the observer should press a button. The behavior preceeding a button press can be compared with the behavior that is not judged as inappropriate, to reveal patterns of (un)human-like behavior. The ratings can thus be used to evaluate and improve the behavior synthesis models. In principle, one could even learn or adapt these models in an online fashion. The SWOZ methodology lends itself well for the evaluation of synthesis algorithms that focus on a limited number of distinct behaviors. As observations of the subjects are continuously recorded, the framework doubles as a tool for the study into nonverbal behavior.

We will discuss learning and evaluation of behavior synthesis models in the next section. The SWOZ framework is introduced in Sect. 3. In Sect. 4, we demonstrate the framework in the context of backchannel behavior in speaker–listener dialogs, and present results in Sect. 5. We outline directions for further research and application in Sect. 6.

2 Related work

We first discuss the offline learning and evaluation of nonverbal behavior synthesis models, with a focus on those targeting listening behavior. Next, we turn to online evaluation in Sect. 2.3.

2.1 Learning behavior models

Nonverbal behavior models are predominantly learned from annotated corpora of dialogs between human subjects [18], or based on simple observations from literature such as (co)occurrence statistics. The annotation of these corpora typically involves manual labeling of the occurrences of specific nonverbal behaviors such as nodding, pose shifts and smiles. Particularly for listening behavior, such models for the listener are conditioned on the observed behavior of the speaker [12]. The aim of behavior modeling is then to determine when behaviors occur within the context of the interaction, for example at the end of a speaker turn. This results in a (probabilistic) mapping from observed behavior of the speaker to a likeliness of the production of the behavior for the listener. These mappings are commonly learned using machine learning algorithms [21], but can also be specified by hand [24,30].

For the synthesis of listening behavior models, research typically focuses on backchannels, signals from the listener to indicate continued attention, interest and comprehension without the aim of taking the floor [32]. Xudong [31] and Duncan [8] discuss backchannels and their nonverbal forms. Bavelas et al. [2] distinguish between *generic* and *specific* backchannels. The latter are more accurately timed and tailored to the speaker's discourse, e.g. a surprised face or a "wow!" utterance. Generic backchannels are signals of continued attention, and are typically communicated with nods or "uh-huh" utterances. There are differences in the quantity, type and timing of backchannels between cultures and subjects [9,17]. Apart from a few recent studies (e.g. [29]), current work on synthesis models of listening behavior mainly focuses on generic backchannels.

Linguistic features such as the end of a grammatical clause [32] or part-of-speech tags [5] have been found to be good predictors of backchannel production in the listener. In online conversations, such features cannot be obtained robustly in real-time. Therefore, researchers have focussed on low-level features from the speaker's speech and gaze. A region of rising pitch [30], a period of pause [26] and mutual gaze [2,21] have been found to cue backchannels in the listener. These features can be obtained in real-time and therefore are suitable as input to online behavior synthesis algorithms.

2.2 Evaluation of behavior models

The quality of behavior synthesis models is typically measured by comparing generated behaviors to those actually performed in the corpus. Objective measures such as precision and recall are used, but these do not take into account the *optionality* of social behavior. We argue that social behavior performed differently from that in the corpus can also be regarded as appropriate. However, objective measures will discredit such alternative behavior which hinders generalization of behavior synthesis models. Moreover, there is no guarantee that the generated behavior is also perceived as less appropriate should it be performed by a virtual agent. This is due to their limited or adapted animation possibilities.

Perceptual evaluation, where human observers provide subjective ratings, is used to determine whether the generated behavior is perceived as human-like. It requires that humans can perceive the behavior naturally, e.g. using virtual agents [3]. Huang et al. [15] and Poppe et al. [24] use subjective ratings to perceptually evaluate generated sequences of behavior. While such ratings give a general idea of the performance of the model, they suffer from three main drawbacks.

First, it cannot be determined how aspects of the synthesis model (e.g. quantity, type and timing of specific nonverbal behaviors) affect the rating. There is a need for evaluation on a shorter time-scale. Second, the fact that many modalities are not animated has been found to decrease the perceptual ratings as the resulting behavior is more static. As a consequence, systematic variation of generated behavior is also affected by factors that are not controlled. This hinders the understanding which aspects of a behavior synthesis algorithm require adaptation. Third, the evaluations are performed offline. We argue that online evaluation is more ecological valid as the dialogs are then contingent.

The first issue was addressed by Poppe et al. [24], who had human observers watch a video of a speaker and an animation of a listener side-by-side. The listener produced backchannels at predetermined moments. Observers were instructed to press a button when they judged the produced social behavior as inappropriate. With this approach, subjective ratings were obtained at the level of individually generated nonverbal behaviors. While this gives insight in when not to produce a behavior, characteristics of the behavior over time (e.g. number of backchannels, time between two backchannels) are not explicitly taken into account. The work presented in this paper addresses this issue, while at the same time dealing with the limited animation capabilities of a systematic perceptual evaluation approach. Moreover, we focus on online dialogs.

2.3 Online synthesis and evaluation

Engaging in a face-to-face conversation with a virtual agent requires that behavior can be generated in real-time contingent with the observed behavior of a human conversational partner. These observations can be obtained from microphones or cameras and encoded as low-level features. Behavior synthesis models use the features as input to generate sequences of behavior. A final step in this process is to animate these sequences on a virtual agent and display them to the human conversational partner.

Several systems have been introduced that combine online observation and behavior generation. Recently, Huang et al. [15] implemented a virtual agent with the aim of maximizing the feeling of rapport between the agent and a human conversational partner. The agent produces speech, smiles and head nods based on observed speech, smiles, nods and eye gaze of the human subject. Authors have investigated mediated conversations in which the representation of the conversational partner is controlled. MushyPeek is a real-time system where the lip synchronization and head orientation of a virtual agent are generated based on detected voice activity [11]. Different head orientation strategies are evaluated, based on the speech/no-speech state of both conversational partners.

Evaluation of these systems is carried out over entire conversations by looking at the amount of speaking [11] or subjective ratings of rapport [15]. In this paper, we describe a framework specifically aimed at online evaluation of generated behavior on a shorter time-scale, involving one or only a few modalities. The framework is based on the nonverbal Turing test [27], in which human observers have to indicate whether their conversational partner is operated by a human or by an algorithm. Our work shares some similarities with the work of Bailenson et al. [1], who presented a speaker with an animation of a listener based on either the listener's head movements or the speaker's own time-delayed head movements. The work described in this paper is different as we focus on behavior synthesis algorithms in a fully interactive setting. Moreover, we aim at online evaluation on a shorter time scale, in order to gain insight into the (generation of) appropriate listening behavior.

3 Switching Wizard of Oz

The symmetric SWOZ framework is schematically depicted in Fig. 1. Two human subjects A and B, seated at distributed locations, are shown virtual representations of each other. The representation of B displays either the behavior performed by B, or behavior synthesized by an algo-

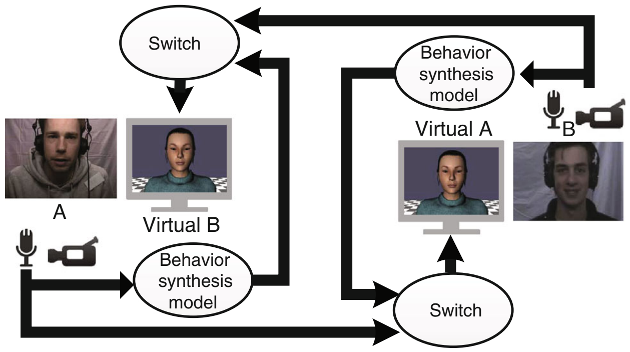


Fig. 1 Schematic representation of the Switching Wizard of Oz framework with two subjects A and B. They are shown a virtual representation of the other either animated directly based on the observed behavior, or based on a behavior synthesis algorithm

rithm, based on observations of A. These observations can be obtained using camera and microphone, or from sensors such as Kinects or gaze trackers. The behaviors displayed by the virtual representations can be discrete (e.g. nods, smiles) or continuous behaviors (e.g. head movement).

To evaluate the quality of behavior synthesis models, both subjects are presented with a yuck button which they press whenever they believe the displayed behavior does not originate from the other subject. The concept of a yuck button was also used as a post-check in [20] and for quantitative analysis in [24]. Given that both human and algorithm use the same modalities for communication, we can compare the behavior and the yucks to identify how quantity, type and timing of the behaviors influences the perception. During a conversation, the source (i.e. human or algorithm) of the virtual agent is switched at random time intervals. This forces the subjects to continuously evaluate the behavior, also on a shorter timescale compared to when presented with stimuli originating from a single source. Consequently, we can evaluate more behavior data in less time.

The three components *subject observation*, *behavior synthesis* and *behavior switching* are discussed subsequently. In Sect. 3.4, we describe how the framework could be used in an experimental setting.

3.1 Subject observation

The conversational partners are observed via sensors, whose outputs are encoded into features in real-time. For cameras, head tracking, body pose estimation or gesture recognition software could be used. For microphones, acoustic analysis software could be used to obtain speech features such as pitch and intensity level. Keyword spotting and (incremental) speech recognition could be used as additional sources. It should also be possible to regenerate the observed behavior on the virtual representation of the subject. For example, the virtual agent's head movement and facial expressions can be animated based on the output of head tracker and facial expression analysis software. In addition, the human's speech can be replayed directly.

3.2 Behavior synthesis

The extracted features are subsequently used in a behavior synthesis algorithm, to determine whether or not certain behaviors should be animated. These algorithms can be manually engineered sets of classification rules or machine learning classifiers trained on previously recorded corpus data. We treat the algorithm as a black box and only assume that its output is a (confidence) score or classification. Based on this output or the observations of the actual conversational partner, the behavior is animated on a virtual agent. Behaviors can be discrete or continuous.

3.3 Behavior switching

An important aspect of the framework is the switching between the two sources at random time intervals. The duration of an interval can differ between research questions. Some behaviors can be analyzed at a smaller time scale than others. It is important that the switching is not noticeable for the subjects. For discrete events, this implies that the currently animated behavior should be finished and a new behavior should not be directly animated. For continuous behaviors, it should also be ensured that the displayed behavior is continuous. As the switching component of the framework is presented with the behavior of both the conversational partner and the algorithm, the switching time can be selected when the two sources are more similar, to allow for interpolation between the two. For example, the source could be switched when the difference in head orientation and movement between both sources is small.

3.4 Implementation of the framework

The framework is general and several adaptations are possible to focus on specific situations. First, we presented the framework as symmetric, with both conversational partners being observed and animated. In the next section, we will present an asymmetric variant that is well suited for speaker–listener dialogs in which only the behavior of the listener is animated and evaluated. In this case, the audio and video of the speaker are directly shown to the listener. Only the speaker provides perceptual ratings. An asymmetric setup is suitable for behavior where the two subjects have different (conversational) roles or modalities of expression.

Another implementation consideration is whether to animate discrete or continuous behavior. In the latter case, the behavior of the conversational partner and the algorithm are required to be continuously measured and generated. In the former case, discrete events need to be recognized from the observations. Instead of recognizing these from sensors, we can also use a keyboard. Conversational partners then have to press a button whenever they would give a certain nonverbal signal. While this seems artificial, the use in Parasocial Concensus Sampling [14] demonstrates that the distributions in the number and timing of backchannels are comparable to actual performance of the behavior. We will use this idea in the next section.

At each moment in time, the framework receives behavior specifications from the human subject and the behavior synthesis algorithm. The behavior of both can be recorded, and used later for offline evaluation or for training the algorithms. For example, one can analyze in which multimodal context the human subject produced a smile, or one can adjust a threshold for the production of nods based on the confidence scores obtained from evaluating the machine learning algo-

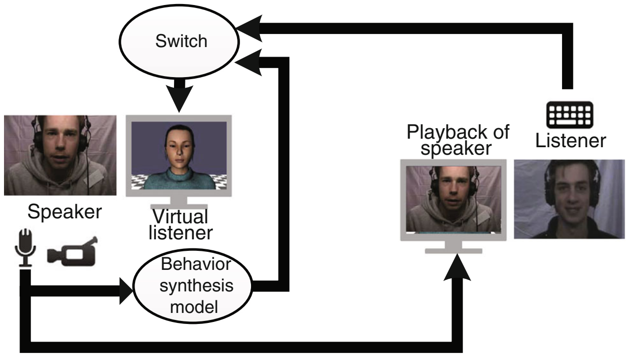


Fig. 2 Asymmetric setting of the Switching Wizard of Oz framework with a speaker and listener. The speaker is shown a virtual representation of the listener, while the listener sees the video of the speaker

rithm. In addition, the recorded data can be used for human perception studies.

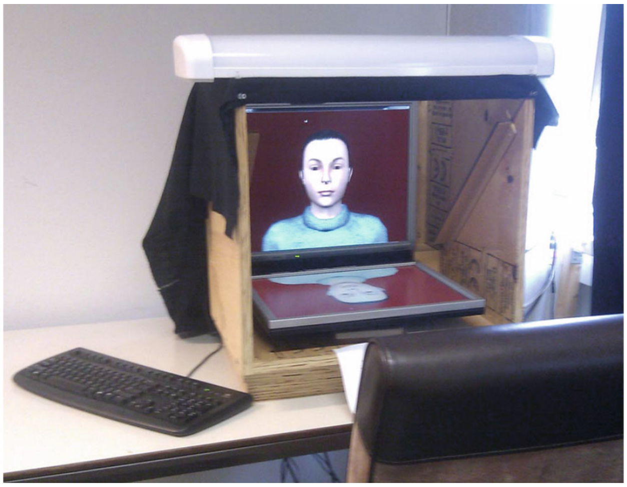


Fig. 3 Experiment setup at speaker side. The setup at the listener's side is similar, but displays the video of the speaker instead of a virtual representation. A camera is placed behind the one-way mirror

4 SWOZ for backchannel synthesis evaluation

To demonstrate the use of the SWOZ framework, we applied it to the evaluation of backchannel behavior in speaker–listener dialogs [12]. Given the nature of such dialogs, we use an asymmetric setting (see Sect. 3.4, and Fig. 2). Only the behavior of the speaker is observed, and used as input to animate the behavior of the virtual listener. Consequently, only the speaker makes perceptual judgements about the behavior of the virtual listener. We discuss the setup, procedure and participants in the subsequent sections, followed by a summary and discussion of the results in Sect. 5.

4.1 Setup

Speaker and listener are seated at distributed locations. The setup at the speaker's side is shown in Fig. 3. A one-way mirror is used to record the speaker through the projection of a virtual listener, to achieve a better sense of eye-contact. The listener sees a video of the speaker on a screen, and generates discrete backchannel events by pressing the space bar on a keyboard. It circumvents the recognition of nods and vocalizations from video and audio, respectively, which would add an additional delay. Also, incorrect recognition of the backchannels would introduce noise into the analyses.

4.1.1 Subject observation

Currently, we do not analyze the video of the speaker and only focus on the speaker's speech, recorded with a microphone. Previous work (e.g. [15,26,30]) has demonstrated that there are differences in the acoustics of the speaker's speech prior to the production of a backchannel in the listener. This makes acoustic features good predictors for backchannels. We obtain the first 12 mel-frequency cepstrum coefficients (MFCC) and speech intensity at 30 frames per second using the CoMIRVA toolkit [25]. As different speakers might have very different acoustic speech profiles, we calculate z-scores instead of the raw MFCC and intensity features. Our processing largely follows that of De Kok et al. [10]. Specifically, when a new measurement is available, we calculate the mean and slope over the past 3, 6 and 15 measurements, which correspond to intervals of approximately 100, 200 and 500 ms, respectively. Additionally, we use one feature to indicate the time since the last change from speaking to pause and vice versa. We calculate the relative offset in milliseconds to the moment where the speaker starts or stops talking, based on thresholded energy values. When the speaker stopped talking, we negate the time difference. We combine all features into a 79-dimensional vector ($2 \times 3 \times (12 + 1) + 1$) per time instant.

4.1.2 Behavior synthesis and model learning

The observed features are used as input for a support vector machine (SVM), a machine learning classifier that gives a score for each individual feature vector. The temporal dimension is not taken into account in the classifier, only in the feature encoding through the use of windows and offsets. We apply a threshold on the classifier scores. If the score is above the threshold, we animate a backchannel for the virtual listener, provided that no backchannel has been performed in the previous second. In this paper, we treat the synthesis algorithm as a black box, i.e., we do not analyze the contribution of individual features, nor do we validate the model that is learned.

The SVM is trained on data gathered using a similar setup but without the switching component, behavior synthesis algorithm and yuck button. Effectively, this renders the setup to a Wizard of Oz setting where the speaker always sees a virtual representation of the human listener. We recorded six conversations and extracted feature vectors at moments where backchannels were produced by the human listener. These are used as positive samples. In addition, we sampled the same number (312) of negative samples at moments where no backchannel was produced within a window of 1 s (500 ms before and after). The first 15 s of each interaction were used to determine the mean and standard deviation of each feature. For the remainder of the recordings, these were used to calculate the z-scores. We trained the SVM using LibSVM [6] with the default parameters. We then empirically determined the value of the threshold on the classification scores. This threshold was fixed for all experiment sessions.

For the animation of the virtual agent, we used the Elckerlyc virtual human platform [28]. Backchannels can have many forms, including nods, short vocalizations, smiles and other facial expressions [8]. For experimental control, and in line with recent research (e.g. [21]), we use head nods together with a "uh-huh" vocalization. These are regarded as discrete generic feedback [2]. Backchannel animations were planned and animated directly on the virtual listener when prompted by the human listener or algorithm. The delay due to network and planning time was estimated at a maximum of 50 ms.

4.1.3 Behavior switching

We switched the source of the virtual listener at random time intervals, sampled from a normal distribution with mean 30 s and a standard deviation of 10 s. Sampled lengths shorter than 10 s and longer than 50 s were set to 10 and 50 s, respectively.

The nods and vocalizations we performed in this experiment are discrete behaviors. To ensure that the moment of the switch was not perceptible to the speakers, we again enforced a minimum of 1 s between subsequently generated backchannels.

4.2 Procedure

Participants were explained the aim of the study. The listener was seated in front of a 21 in. screen and instructed to press the space bar on a keyboard whenever he would give a backchannel to the speaker. The listener was unaware of the source of the virtual listener. The speaker was seated in the adjacent room and told explicitly that the displayed behavior of the listener could originate either from the actual listener or from an algorithm, and that these would switch occasionally. Nothing was revealed about the switching interval. Speakers were instructed to press the yuck button if they thought the displayed behavior was not human-like. They were explained that pressing the button would switch the source of the virtual listener to the actual listener. Speakers were given a list of possible conversation topics, including their favorite dishes and their opinion on societal issues. They were free to discuss any topic for any length of time. To avoid speech disfluencies due to language difficulties, we deliberately chose to have all conversations in Dutch, the native language of most of our staff and students. Recordings were stopped by the experimenter when the speaker ran out of conversation topics or made an indicative remark about this. In other cases, we stopped recording when a topic change occurred after more than 7.5 min of conversation.

Before the start of the conversation, speakers were asked to introduce themselves briefly. We recorded their speech and calculated the average and standard deviation of the MFCC and speech intensity features over this interval. These were used to calculate the z-scores for use in the algorithm.

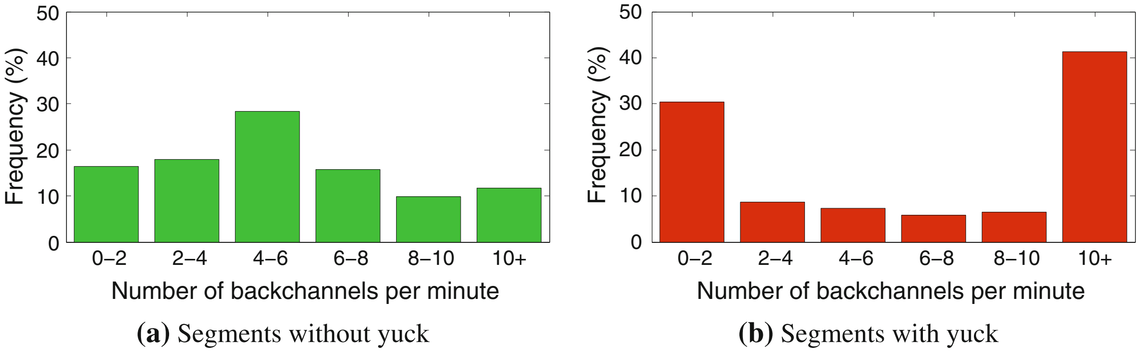
4.3 Participants

In total, we recruited 24 participants (5 female, mean age 27.25) in 12 pairs. For each pair, the first speaker was chosen randomly. The roles were switched after the first conversation.

5 Results and discussion

We recorded 24 conversations with a total duration of 192 min. In 60.22 % of the time, the virtual listener was operated by the human listener. This above-average percentage is partly due to the fact that the source of the virtual listener started with and switched to the human listener when the yuck button was pressed.

First, we take a look at the yuck presses in relation to the source of the virtual listener. Of all 138 yucks, 96 (69.57 %) were given when the virtual listener was operated by the algorithm. Corrected for the unequal distribution of time over the two sources, one would expect only 55 yucks (39.78 % of 138). Apparently, the behavior from the synthesis algorithm is more often regarded as inappropriate. To investigate this further, we analyze segments, intervals between two switches (possibly due to a yuck). In total, 45.93 % of the algorithm segments received a yuck, compared to 15.91 % of the segments originating from the human listener. Of course, some of the yucks might have been given because of the behavior shown before a switch. Just before the speaker pressed the button, the framework might have switched, causing the yuck to be attributed to the other source. This is likely to be the case for both sources. We will not try to compensate for these yucks as it an arbitrary task to determine why the yuck button was pressed. In the following, we will discuss the quantity



(a) Segments without yuck **(b)** Segments with yuck

Fig. 4 Frequency histograms of backchannels per minute, calculated per segment. Yuck presses typically follow quickly the display of a backchannel

and timing of the backchannels in relation to whether a yuck was issued.

5.1 Backchannel quantity

About one third of the yucks has been performed when the virtual listener was operated by the human listener. We investigate whether there are characteristics of the displayed behavior are perceived as inappropriate, independent of the source of the virtual listener. As a first analysis, we looked at the average number of backchannels per minute for those segments that were yucked. We calculated this number by dividing the number of backchannels between the last switch (or yuck) and the yuck, divided by the length of this interval. Initially, we distinguished only between the segments that received a yuck and those that did not, independent of the source of the virtual listener. The frequency histograms appear in Fig. 4.

The backchannel frequency of segments that did not receive a yuck peaks between 4 and 6 backchannels per minute. Yucks have been given especially for segments with very low (including zero) and very high frequencies. In line with [24], the more the frequency deviated from 4 to 6, the more likely it is that the segment received a yuck. This number is informative for the design of backchannel synthesis algorithms. We subsequently analyzed the source of these frequencies, and found that many of these corresponded to the algorithm. This was noted by the participants in the recordings as well. In several cases, the algorithm produced many nods in sequence while in other cases, it rarely produced a nod. This shows in the data, where 74.03 % of the segments without any backchannel corresponded to the setting where the virtual listener was operated by the algorithm.

This effect is partly due to the fixed threshold on the classification score of the algorithm. Another cause could be the normalization of the features, based on the speech recorded prior to the actual conversation. The nature of this speech could be different from that of the remainder of the conversation. Especially when the standard deviation of features was either lower or higher, this typically caused more or less backchannels in the conversation, respectively. Both issues could be solved by applying an adaptive threshold, for example by continuously considering the last 15 or 30 s of speech. While this still depends on the characteristics of the discourse (e.g. topic, involvement), differences between subjects can be mitigated to some extent.

5.2 Backchannel timing

Besides differences in backchannel frequency, the production or omission of a backchannel at a certain moment in time could be regarded as inappropriate. Next, we turn to this timing of individual backchannels. We noticed that many yucks were given directly after the display of a backchannel. Inappropriate timing probably has caused the subject to press the yuck button. To analyze this, we calculated the time between the last backchannel and the end of the segment. We distinguished between segments that ended with a yuck and those that did not. Frequency histograms appear in Fig. 5. Many yuck presses follow shortly after a backchannel has been produced. This is probably a response to an inappropriately timed backchannel. Again, the virtual listener was driven by the backchannel synthesis algorithm in the majority of these cases.

We compare the backchannel timings of the human listener and the backchannel synthesis algorithm. The better the timings match, the more the algorithm approximates the backchannel behavior of a human listener. In the SWOZ setup, we stored the key presses of the actual listener and the predictions of the algorithm, also when they were not animated. As such, we can compare the timings of both sources. To investigate how well the timings matched, we considered them matching if they were produced within a margin of 1 s (500 ms earlier or later). Of all backchannels shown to the

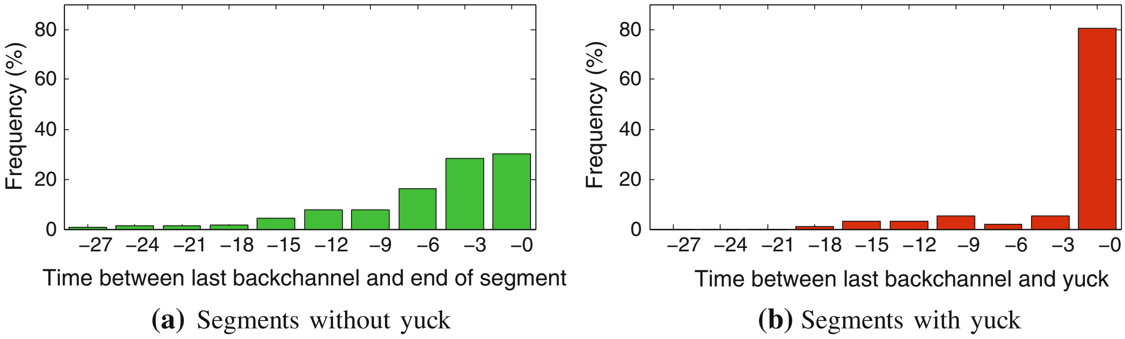


Fig. 5 Frequency histograms of time between the last backchannel and segment end. Segments with very few (0–2) or many (10+) backchannels per minute more often receive a yuck

speaker, 19.61 % matched a backchannel produced by the other source. However, from the backchannels performed in a segment that received a yuck, only 15.65 % matched (compared to 20.74 % when the segment ended without a yuck). Apparently, the backchannel behavior is perceived as more appropriate when the timings produced by the algorithm and actual listener are more similar. It should be noted that these percentages appear rather low. This is mainly due to the optionality of backchannels. They can be produced at various places within the speaker's discourse and still be perceived as human-like.

We noted that both the human listener and the algorithm occasionally produced backchannels when the speaker just started a sentence, halfway a sentence and after filled pauses (e.g. "uhm"). Given that we obtained speech state estimates using the CoMIRVA toolkit [25], we analyzed whether backchannels were performed during the speaker's discourse. It shows that none of the backchannels produced by the algorithm were made while speech was detected. This might be due to the training of the SVM. In comparison, the human listener made 22.25 % of the backchannels while speech was detected. Closer analysis revealed that the speech/non-speech estimates were not always accurate. Still, speech was detected in 46.97 % of the time, which means that backchannels are more often produced in pauses.

6 Conclusion and future work

We introduced the SWOZ, a framework to evaluate and record nonverbal behavior in an online mediated setting. The setup combines ideas from the Turing test with those of a Wizard of Oz setup. In a distributed setup, two conversational partners are shown, on a virtual agent, either the behavior of the other or behavior generated by an algorithm. The system switches between the two at random time intervals. Each conversational partner can indicate (by pressing a button) that he perceives the behavior as inappropriate. Humans and algorithm use the same limited set of modalities, which eliminates any bias in the perceptual judgements due to modalities or behaviors that are not animated. This allows for the quality of the algorithm to be expressed in terms of characteristics such as the quantity, type and timing of the behaviors.

To demonstrate the potential of the SWOZ framework, we conducted a user experiment on backchannel behavior in online speaker-listener dialogs. We used the asymmetric SWOZ framework with only the speaker's voice recorded. Based on the pitch, intensity and speech state, we evaluated a trained SVM and thresholded the output. The virtual listener displayed a nod and short vocalization to the speaker, based either on the algorithm or on button presses of the actual listener. Speakers judged the listener's behavior with too many or too few backchannels per minute as inappropriate. In the majority of the cases, the virtual listener's behavior then originated from the algorithm. The high and low backchannel frequencies were mainly due to the fixed threshold used to decide whether a backchannel should be produced. These findings give rise to the adaptation of the way in which the threshold is set for different individuals. In addition, we found that backchannels were more often found inappropriate when the human listener and the algorithm's generated timings did not match.

The SWOZ methodology has some unique advantages over corpus-based evaluation and offline perceptual evaluation studies. First, evaluation in online dialogs is more ecologically valid, and enables us to obtain more data in less time. Second, the fact that both human listener and algorithm have the same expression capabilities ensures that the omission of certain modalities or behaviors does not lead to a bias in the perception of the animated behavior. Third, the switching ensures that we can evaluate behavior at a suitable time-scale, typically shorter than a conversation and, when needed, longer than an individual behavior.

Future work will consider larger-scale evaluation of behavior synthesis models, initially in a similar asymmetric setting. We plan to replace the listener's button with modules that recognize nods and vocalizations from the listener's head movements and voice. This requires that these events are detected at an early stage, to avoid delays between the detection and generation of the backchannel [13]. In addition, we plan to incorporate other modalities such as facial expressions (e.g. smiles and frowns) and head movement as these can also have a backchannel function [4,8]. A second avenue for future work is to make the behavior synthesis models more flexible. For the type of backchannel behavior that we evaluated in this paper, we will look at ways to automatically adapt the threshold for the generation of backchannels. Also, we consider using machine learning models that take into account the time dimension in the classification.

The SWOZ framework can also be used as a recording tool, and we intend to use the yucks to adapt the behavior synthesis models. We can use the yuck moments as negative samples in the training of machine learning algorithms, for example using the iterative approach by De Kok et al. [10].

References

1. Bailenson JN, Yee N, Patel K, Beall AC (2008) Detecting digital chameleons. Comput Hum Behav 24(1):66–87
2. Bavelas JB, Coates L, Johnson T (2002) Listener responses as a collaborative process: the role of gaze. J Commun 52(3):566–580
3. Bente G, Krämer NC, Petersen A, de Ruiter JP (2001) Computer animated movement and person perception: methodological advances in nonverbal behavior research. J Nonverbal Behav 25(3):151–166
4. Brunner LJ (1979) Smiles can be back channels. J Pers Soc Psychol 37(5):728–734
5. Cathcart N, Carletta J, Klein E (2003) A shallow model of backchannel continuers in spoken dialogue. In: Proceedings of the conference of the European chapter of the association for computational linguistics, Budapest, Hungary, vol 1, pp 51–58
6. Chang CC, Lin CJ (2011) LibSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27
7. Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of Oz studies: why and how. In: Proceedings of the international conference on intelligent user interfaces (IUI), Orlando, FL, pp 193–200
8. Duncan S Jr (1974) On the structure of speaker–auditor interaction during speaking turns. Lang Soc 3(2):161–180
9. de Kok I, Ozkan D, Heylen D, Morency LP (2010) Learning and evaluating response prediction models using parallel listener consensus. In: Proceedings of the international conference on multimodal interfaces (ICMI), Beijing, China
10. de Kok I, Poppe R, Heylen D (2012) Iterative perceptual learning for social behavior synthesis. Technical report, TR-CTIT-12-01, University of Twente
11. Edlund J, Beskow J (2009) Mushypeek: a framework for online investigation of audiovisual dialogue phenomena. Lang Speech 52(2–3):351–367
12. Heylen D, Bevacqua E, Pelachaud C, Poggi I, Gratch J, Schröder M (2011) Generating listening behaviour. In: Cowie R, Pelachaud C, Petta P (eds) Emotion-oriented systems cognitive technologies. Springer, Berlin, pp 321–347
13. Hoai M, la Torre FD (2012) Max-margin early event detectors. In: Proceedings of the conference on computer vision and pattern recognition (CVPR), Providence, RI, pp 2863–2870
14. Huang L, Morency LP, Gratch J (2010) Learning backchannel prediction model from parasocial consensus sampling: a subjective evaluation. In: Proceedings of the international conference on interactive virtual agents (IVA), Philadelphia, PA, pp 159–172
15. Huang L, Morency LP, Gratch J (2011) Virtual rapport 2.0. In: Proceedings of the international conference on interactive virtual agents (IVA), Reykjavik, Iceland, pp 68–79
16. Krauss RM, Garlock CM, Bricker PD, McMahon LE (1977) The role of audible and visible back-channel responses in interpersonal communication. J Pers Soc Psychol 35(7):523–529
17. Li HZ (2006) Backchannel responses as misleading feedback in intercultural discourse. J Intercult Commun Res 35(2):99–116
18. Martin JC, Paggio P, Kuehnlein P, Stiefelhagen R, Pianesi F (2008) Introduction to the special issue on multimodal corpora for modeling human multimodal behavior. Lang Resour Eval 42(2):253–264
19. McDonnell R, Ennis C, Dobbyn S, O'Sullivan C (2009) Talking bodies: sensitivity to desynchronization of conversations. ACM Trans Appl Percept 6(4):A22
20. McKeown G, Valstar M, Cowie R, Pantic M, Schröder M (2012) The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput 3(1):5–17
21. Morency LP, de Kok I, Gratch J (2010) A probabilistic multimodal approach for predicting listener backchannels. Auton Agents Multi-Agent Syst 20(1):80–84
22. Poppe R, ter Maat M, Heylen D (2012) Online backchannel synthesis evaluation with the Switching Wizard of Oz. In: Joint proceedings of the intelligent virtual agents (IVA) 2012 workshops, Santa Cruz, CA, pp 75–82
23. Poppe R, ter Maat M, Heylen D (2012) Online behavior evaluation with the switching wizard of Oz. In: Proceedings of the international conference on interactive virtual agents (IVA), Santa Cruz, CA, pp 486–488
24. Poppe R, Truong KP, Heylen D (2013) Perceptual evaluation of backchannel strategies for artificial listeners. J Auton Agents Multi-Agent Syst 27(2):235–253
25. Schedl M (2006) The CoMIRVA toolkit for visualizing music-related data. Technical report, Department of Computational Perception, Johannes Kepler University Linz
26. Truong KP, Poppe R, de Kok I, Heylen D (2011) A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Proceedings of interspeech, Florence, Italy, pp 2973–2976
27. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460
28. van Welbergen H, Reidsma D, Ruttkay Z, Zwiers J (2010) Elckerlyc—a BML realizer for continuous, multimodal interaction with a virtual human. J Multimodal User Interfaces 3(4):271–284
29. Wang Z, Lee J, Marsella S (2013) Multi-party, multi-role comprehensive listening behavior. J Auton Agents Multi-Agent Syst 27(2):218–234
30. Ward N, Tsukahara W (2000) Prosodic features which cue backchannel responses in English and Japanese. J Pragmat 32(8):1177–1207
31. Xudong D (2009) The pragmatics of interaction. chap. Listener response. John Benjamins Publishing, Amsterdam, pp 104–124
32. Yngve VH (1970) On getting a word in edgewise. In: Papers from the sixth regional meeting of Chicago Linguistic Society. Chicago Linguistic Society, Chicago, pp 567–577