Online Backchannel Synthesis Evaluation with the Switching Wizard of Oz

Ronald Poppe, Mark ter Maat, and Dirk Heylen *

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{r.w.poppe,m.termaat,d.k.j.heylen}@utwente.nl

Abstract. In this paper, we evaluate a backchannel synthesis algorithm in an online conversation between a human speaker and a virtual listener. We adopt the Switching Wizard of Oz (SWOZ) approach to assess behavior synthesis algorithms online. A human speaker watches a virtual listener that is either controlled by a human listener or by an algorithm. The source switches at random intervals. 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.

1 Introduction

Our aim is to generate human-like listening behavior for artificial listeners, virtual agents that signal attention, understanding and interest in a face-to-face dialog with a human speaker [1]. The quality of the behavior model should be measured by how the animated behavior is perceived. Humans are particularly sensitive to flaws in the displayed behavior, both in form and timing [2]. This effect also occurs when certain behaviors are not animated (e.g. eye blinks, respiration). Consequently, a virtual agent’s behavior is typically perceived as rather unrealistic, especially in experimental settings where the behavior of the virtual agent is varied systematically among only one or a few modalities (e.g. [3]).

In this paper, we perceptually evaluate a backchannel synthesis algorithm [4] using a recently introduced methodology, the Switching Wizard of Oz (SWOZ, [5]). This approach combines ideas behind the Turing test with those of a Wizard of Oz setup. At the heart of SWOZ is a distributed video-conferencing setting with a human speaker and a human listener. The speaker is observed with a camera and a microphone and algorithms are employed to analyze the verbal and nonverbal behavior in real-time. These observations are fed into a behavior synthesis model. The speaker is shown a virtual representation of the listener, animated based on one of two sources: (1) directly on the observed behavior of the listener, or (2) on the output of the synthesis model. Both sources share the same behavior animation capabilities and limitations. During a conversation, the source of animation of the representation of each subject switches occasionally.

* This publication was supported by the Dutch national program COMMIT.
When the displayed behavior deviates from what is typically regarded as human-like, the speaker should notice and press a button (the yuck button [3]). The ratings can be used to evaluate and improve the behavior synthesis models.

We will discuss learning and evaluation of listening behavior synthesis models in the next section. In Section 3, we describe our SWOZ experiment on backchannel synthesis in speaker-listener dialogs. Results appear in Section 4. We conclude in Section 5 and present directions of further research and application.

2 Related Work on Backchannel Synthesis

Models of nonverbal behavior are pre-dominantly learned from corpora of dialogues between human subjects [6], or based on observations from literature. Corpus annotation involves manual labelling of specific nonverbal behaviors such as nodding and pose shifts. For listening behavior, models for the listener are conditioned on the observed behavior of the speaker [1]. The aim of behavior modelling is to determine when social behaviors occur within the context of the interaction. 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 (e.g. [7]), but can also be specified by hand (e.g. [8, 9]).

Linguistic features such as the end of a grammatical clause [10] have been found to be good predictors of backchannels. However, in an online conversation, such semantic features cannot be obtained in real-time. Therefore, researchers have focussed on low-level features from the speaker’s speech and gaze. A region of high or rising pitch [8], a period of pause [11] and a moment of mutual gaze [7] have been found to cue backchannels.

2.1 Evaluation of behavior models

The quality of behavior synthesis models is typically measured by comparing generated behaviors to those performed in the corpus. Objective measures such as precision and recall do not take into account the optionality of social behavior. We argue that behavior performed differently from that in the corpus can also be appropriate. However, objective measures will discredit such alternative behavior.

Perceptual evaluation, where human observers provide subjective ratings, can be used to determine whether the generated behavior is perceived as human-like. Such an evaluation requires that the behavior is generated in such a way that humans can perceive it in a natural manner. Virtual agents are typically used for this task (e.g. [12, 9]). While such ratings give a general idea of the performance of the behavior synthesis model, they suffer from two 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. This hinders the understanding which aspects of a behavior synthesis algorithm require adaptation.
The first issue was addressed by Poppe et al. [3], who had human observers watch a video of a speaker and an animation of a listener side-by-side. The listener produced specific social signals at predetermined moments. Observers were instructed to press a button when they judged the produced social behavior as inappropriate. This approach gives subjective ratings 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. amount of backchannels, time between two backchannels) are not 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.

2.2 Online synthesis and evaluation

A final step in this process is to animate these sequences on a virtual agent and display this to the human conversational partner. Several systems have been introduced that combine online observation and behavior generation. Huang et al. [12] implemented a virtual agent with the aim of maximizing the feeling of rapport between the agent and a human conversational partner by producing speech utterances, smiles and head nods. Several authors have investigated mediated conversations in which the representation of the conversational partner is (systematically) 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 [13]. Evaluation of these systems is carried out over entire conversations by looking at the amount of speaking [13] or subjective ratings of rapport [12]. In this paper, we will describe a framework based on these previous works, but specifically aimed at online evaluation of generated behavior on a shorter time-scale, involving one or only a few modalities. Our work shares similarities with the work by Bailenson et al. [14], but we repeatedly switch between human and algorithm, and focus on online evaluation.

3 Experimental Setup

We conducted an evaluation of backchannel behavior in speaker-listener dialogs. We use an asymmetric setting (see Figure 1) of the SWOZ framework. Only the behavior of the speaker is observed and used as input to animate the behavior of the virtual listener. Only the speaker makes perceptual judgements about the behavior of the virtual listener. We discuss the components of our experiment in the subsequent sections. A discussion of the results appears in Section 4.

3.1 Switching Wizard of Oz Setup

The Switching Wizard of Oz setup is schematically depicted in Figure 1. Speaker and listener are seated at distributed locations. The setup at the speaker’s side is shown in Figure 2(a). 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, and generates discrete backchannel events by pressing a button. At the same time, an algorithm predicts when to generate a backchannel from audio features extracted from the speaker’s speech.

The source of the virtual listener (i.e. listener or algorithm) is switched at random time intervals. To evaluate the quality of behavior synthesis algorithm, speakers are presented with a yuck button which they press when they believe the displayed behavior does not originate from the human listener. Human and algorithm use the same modalities for communication, which allows us to identify how quantity, type and timing of the behaviors affects the perception.

Subject observation We only use the speaker’s speech, recorded with a microphone. We obtain the first 12 mel-frequency cepstrum coefficients (MFCC) and speech intensity at 30 frames per second using the CoMIRVA toolkit [15]. As different speakers might have very different acoustic speech profiles, we calculate z-scores instead of the raw MFCC and intensity features. Our processing follows that of De Kok et al. [4]. 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 500ms, 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 stops talking, we negate the time difference. We combine all features into a 79-dimensional vector \( (2 \times 3 \times (12 + 1) + 1) \). At the same time, the speaker’s audio and video are recorded and directly played back to the human listener.
Behavior synthesis and model learning  The extracted features are used as input for a Support Vector Machine (SVM), a machine learning classifier that gives a classification score for each feature vector. We apply a threshold on the classifier scores and animate a backchannels for the virtual listener, provided that no backchannel has been performed in the previous second. The SVM is trained on data gathered using a similar setup but without the switching component. We recorded six conversations and extracted feature vectors at moments where backchannels were produced by the human listener. In addition, we sampled the same number of negative samples at moments where no backchannel was produced within a 1 second window. We trained the SVM using LibSVM [16] with default parameters. We then empirically set the threshold on the classification scores. For the animation of the backchannels, we used the Elckerlyc virtual human platform [17], see Figure 2(b). Animations were planned directly on the virtual listener when prompted by the human listener or algorithm. The delay due to network and planning time is estimated at 50ms.

![Fig. 2.](image)

Behavior switching  An important aspect of the SWOZ framework is the switching between the human listener and the backchannel synthesis algorithm at random time intervals, sampled from a normal distribution with mean length 30 seconds and a standard deviation of 10 seconds. Lengths shorter than 10 seconds and longer than 50 seconds were set to 10 and 50 seconds, respectively. As a backchannel might have been generated just before the switch, we again enforced a minimum of one second between subsequently generated backchannels.

3.2 Procedure

Participants were introduced to each other and explained the aim of the study. The listener was seated and instructed to press a button whenever he would give a backchannel to the speaker. The speaker was seated 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. In our experiment, only backchannels (a nod with a “uh-huh” vocalization) were displayed. Speakers were instructed to press the yuck button if they thought the behavior was not performed by the human listener. 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. They were free to discuss any topic for any length of time. To avoid speech disfluencies, we deliberately chose to have conversations in Dutch, the native language of most of our staff and students.

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. Recordings were stopped when the speaker ran out of conversation topics. In other cases, we stopped recording when a topic change occurred after more than 7.5 minutes.

### 3.3 Participants

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

### 4 Results and discussion

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

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 55 yucks. We consider a segment to be the interval 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. It is clear that the backchannel behavior of the algorithm is perceived as less human-like than that of the human listener. Some of the yucks might have been given primarily because of the behavior shown before a switch. Just before the speaker pressed the button, the framework might have switched. This is likely to be the case for both sources.

About one third of the yucks has been performed when the virtual listener was operated by the human listener. Apparently, some of the characteristics of the displayed behavior is perceived as inappropriate. We look at average number of backchannels per minute for those segments that were yucked. This is the number of backchannels between the last switch and the yuck, divided by the time of this interval. The frequency histograms appear in Figure 3.
Fig. 3. Frequency histograms of backchannels per minute, calculated per segment without yucks (left) and with yucks (right).

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 and very high frequencies. Many of these correspond to the algorithm. In several cases, the algorithm produced many nods in sequence while in other cases, it rarely produced a nod. This also 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 due to the fixed threshold on the classification score of the algorithm. This is not the best solution for all participants.

Apart from these findings, we note that many yucks are given directly after the display of a backchannel. Participants of the experiment repeatedly reported that they pressed the yuck button when a backchannel was performed when they did not make a statement or after an “uhm” utterance.

5 Conclusion and Future Work

We evaluated a backchannel synthesis algorithm in the context of speaker-listener dialogs using the Switching Wizard of Oz, a framework to evaluate non-verbal behavior in an online setting. In a distributed setup, a human speaker had a conversation with a virtual listener that was either animated by a human listener, or by an algorithm. The algorithm used acoustic features obtained from the speaker, classified using an SVM. The system switched between the two at random time intervals. The speaker could indicate (by pressing a button) that he perceived the behavior as inappropriate. It was found that the rate of backchannels per minute partly determined whether the backchannel behavior was perceived as human-like. This finding gives rise to adaptation of the way in which the backchannel generation threshold is set for different individuals.

Future work will aim at larger-scale evaluation of behavior synthesis models. In addition, we plan to replace the listener’s button with observation modules that automatically recognize nods and vocalizations. We will incorporate other modalities such as facial expressions and head movement. Also, we plan to analyze how the behavior is perceived by external observators. Eventually, we want to make the behavior synthesis models more person-independent. 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.
References