Backchannel Prediction for Mandarin Human-Computer Interaction

Xia MAO†, Member, Yiping PENG†, Yuli XUE†(a), Na LUO†, and Alberto ROVETTA††(b), Nonmembers

SUMMARY In recent years, researchers have tried to create unhindered human-computer interaction by giving virtual agents human-like conversational skills. Predicting backchannel feedback for agent listeners has become a novel research hot-spot. The main goal of this paper is to identify appropriate features and methods for backchannel prediction in Mandarin conversations. Firstly, multimodal Mandarin conversations are recorded for the analysis of backchannel behaviors. In order to eliminate individual difference in the original face-to-face conversations, more backchannels from different listeners are gathered together. These data confirm that backchannels occurring in the speakers’ pauses form a vast majority in Mandarin conversations. Both prosodic and visual features are used in backchannel prediction. Four types of models based on the speakers’ pauses are built by using support vector machine classifiers. An evaluation of the pause-based prediction model has shown relatively high accuracy in consideration of the optional nature of backchannel feedback. Finally, the results of the subjective evaluation validate that the conversations performed by generating appropriate verbal or non-verbal behaviors in backchannel feedback improved the narrator’s performance [11].

1. Introduction

In the process of human-computer interaction, computers should be able to perceive human thoughts and respond accordingly. In order to make such interaction more natural and efficient, virtual agents have been developed. They are computer generated, animated and artificial intelligence virtual characters (usually with human-like appearance [11]), which are capable of responding with adequate verbal or non-verbal behaviors to the users [2], [3]. With the development of virtual agent technologies, researchers are seeking new methods to make the agent perceive and perform more like humans [4]–[6].

In such research, one major issue is to endow virtual agents with the potential to sustain conversations with human speakers. In human-human conversations, when the speaker is speaking, the listener will naturally produce some behaviors or short utterances as feedback to the speaker. This kind of feedback, which is called backchannel, could be defined as visual or audio cues which do not carry certain meanings, and acts only as a sign of the listener’s attentiveness, e.g. head nodding, smiling, shaking one’s head, saying ‘yeah’ and so on, varying from culture to culture. Findings on communication in interaction show that people are not aware of most backchannel feedbacks [7]. Backchannel feedback is pervasive in conversations and is an important kind of feedback that promotes dialogue and shows the listener’s interest by encouraging the speaker to continue.

Dittmann and Llewellyn discussed what they call listener response [8]. They found that the listener tends to produce vocal and non-vocal feedback at the end of the speaker’s utterance. Yngve was the first to take an interest in backchannel feedback [9], where the backchannel communication was found by analyzing English conversations. Later, Duncan termed this kind of listener’s feedback as backchannel behaviors [10]. Earlier studies showed that when people interacted with others, they used backchannel feedback which included speech prosody, gesture, gaze, posture and facial expression to establish a sense of rapport. Backchannel feedback plays an important role in daily dialogues. Bavelas et al. emphasized that application of appropriate backchannel feedback improved the narrator’s performance [11].

In human-computer interaction, on the other hand, backchannels were also proved to help create more natural conversation environment. Nishimura et al. developed a Japanese spoken dialog system which could spontaneously generate chat-like responses including backchannel [12]. The system collected and analyzed the user’s speech and then made use of a decision tree to generate response timing. All the responses were generated by a speech synthesizer. The subjective evaluation results of the backchannel generation method showed a high degree of naturalness. Almost all participants felt a natural familiarity with the backchannels and found the system friendly.

With human-like appearance, virtual agents have the capability of simulating human-human multimodal interaction in a more natural way. In other words, in order to be perceived as natural, virtual agents need to respond to the users by generating appropriate verbal or non-verbal behaviors in conversations. To improve the fluidity and expressiveness of human-computer interaction, researchers have focused on addressing backchannel feedback to virtual agents. It has become a novel research hot-spot in recent years. The Institute for Creative Technologies in the University of Southern California has addressed this issue with a research program named ‘lifelike virtual agent’.

"Manuscript received June 25, 2014.
Manuscript revised December 26, 2014.
Manuscript publicized March 2, 2015.
†The authors are with the School of Electronic and Information Engineering, Beihang University, 37 Xueyuan Road, Haidian District, 100191, Beijing, China.
††The author is with the Department of Mechanics, Polytechnic University of Milan, Via La Masa 34, 20156 Milan, Italy.
a) E-mail: yuli_xue@buaa.edu.cn
b) E-mail: alberto.rovetta@polimi.it
DOI: 10.1587/transinf.2014EDP7214"
California designed a virtual agent called Virtual Rapport 2.0[13]. It can detect human speaker’s silence, nod, gaze aversion and smile then take these features as input to predict backchannel timing, end-of-turn and affective response. The evaluation results indicated that the agent increased the fluency and engagement of the speaker. The Sensitive Artificial Listeners (SAL) project[14] is part of the EU project SEMAINE. It aims at building an autonomous agent which is able to exhibit emotional and nonverbal behaviors. The proposed agent listener is capable of sustaining the conversations by producing appropriate backchannels. Sakai et al. developed a listener agent for elderly people with dementia[15]. The listener agent can generate backchannel feedback, such as head nod and verbal acknowledgement, on the basis of the user’s speech information. The elderly people were found to pay more attention to the agent when it gave backchannel feedback and they were more interested in communicating with agents which could produce social signals. Researchers at the Human Media Interaction Group in the University of Twente have built a Dutch multimodal system called the MultiLis corpus which has been used in backchannel prediction research[16].

However, little research on Chinese backchannel prediction has been done. Ward and McCartney have developed a visualization tool that helps discovering certain prosodic cues which may indicate change of turn, and used this utility in Chinese backchannel modeling[17]. Meanwhile, since Chinese culture introduces distinguishing interaction styles, backchannel prediction in Mandarin proves a challenge. For example, vocal backchannels might be considered as interruptions in Chinese culture, which are impolite. As a result, Chinese people tend to produce less vocal backchannels in conversations than average, and most of them are performed during speakers’ pauses. More over, in Standard Mandarin, four main tones might occur. The tones change the pitch of individual characters, thus requiring an adaptation to pitch-based prediction methods. For this reason, a novel architecture of a backchannel prediction system is put forward in our previous work[18].

Even though the habits of producing backchannels are unique in different languages, earlier research in feature selection and model building are inspiring. In this paper, a model-based Mandarin backchannel prediction method is introduced. We first designed and recorded a face-to-face conversation corpus, then used it as a stimuli to collect different listeners’ backchannel behaviors to the same speaker. Parasocial Consensus Sampling (PCS) analysis was applied to eliminate individual difference to obtain generalized backchannel timings. We extracted various verbal and non-verbal cues from the speakers’ to train a SVM classifier, which was then used to predict backchannel timing for virtual agents. The subjective evaluation showed promising result for our method, showing its potential of encouraging mutual attentiveness in Mandarin human-computer conversation.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces video recording of multimodal Mandarin conversations. Section 4 describes the data collection and the analysis of experiment data. Section 5 shows the multimodal features. Section 6 proposes Mandarin backchannel prediction models. A further subjective evaluation is presented in Sect. 7. Section 8 is the discussion and conclusion.

2. Related Work

Researchers have spent significant effort in predicting backchannels for virtual agent listeners. There are two categories in existing prediction methods: rule-based method and model-based method. Poppe et al. evaluated six widely used rule-based strategies in his study[19]. These include:
- **Copy**: use all the backchannel timing of the actual listener
- **Random**: use a random distribution to generate the backchannel timing
- **Ward and Tsukahara**: use the rule proposed by Ward and Tsukahara[20] (shown in Table 1) to generate the backchannel timing
- **Gaze**: produce backchannel feedback when detect the speaker gazes at the listener after at least 1000ms of no gaze
- **Pitch and Pause**: produce backchannel feedback when detect both a pause and a rising or falling pitch in the speaker’s speech
- **Combination of Gaze, Pitch and Pause**: combine the Gaze strategy and the Pitch and Pause strategy

With the same backchannel timing as the actual human listener, the Copy strategy provided higher scores than the other five strategies in the subjective evaluation results. The average scores of the other five strategies were all below 45 (the full score was 100). This indicates that there is still a long way for current rule-based strategies to improve before making virtual agents perform as well as real human. In fact, rule-based strategies have an inevitable weakness. Since the rule or strategy is made according to a particular corpus, the parameters are closely connected with the corpus. The rule will perform well if it is applied in a similar corpus, but it is not suitable for corpus which is different from the original one.

The model-based method has wider application in predicting backchannels. Researchers collect different features from the speakers and then use them to train a prediction

| Rule | Description |
|------|-------------|
| P1   | a region of pitch less than the 26th-percentile pitch level and |
| P2   | continuing for at least 110ms, |
| P3   | coming after at least 700ms of speech, |
| P4   | provided that no backchannel has been output within the preceding 800ms, |
| P5   | after 700ms wait. |
model. Under theoretical conditions, the prediction model will always perform well if there are sufficient training data. Morency et al. used Hidden Markov Model (HMM) and Conditional Random Fields (CRF) to train their prediction models [21]. Experimental results showed that their method using CRF outperformed the rule-based approach of Ward and Tsukahara. Unfortunately the precision and recall were still quite low (precision = 0.1862, recall = 0.4106, F1 = 0.2562). One of the reasons was not taking individual differences into account.

Backchannel feedback is optional and will be affected by the listener’s personality. Individual differences will have a great influence on the prediction performance. Huang et al. proposed a parasocial consensus sampling (PCS) method to collect more data and improve prediction [22]. This novel data collection method can collect a large amount of behavioral data quickly and eliminate individual difference in backchannel responses. This is achieved by aligning and accumulating individual responses on the same time axis, then use a threshold (response level) to filter out the responses that are relatively low in probability. Evaluation results revealed that virtual agent driven by the PCS data was perceived as more believable than the one driven by pure face-to-face interaction data. For the above reasons, the PCS method is used in the data collection process (Sect. 4).

3. Video Recording of the Mandarin Conversations

As no corpus suitable for analyzing backchannel behaviors in multimodal Mandarin conversations exists, a new corpus with face-to-face Mandarin daily conversation was designed and recorded. The corpus consisted of video clips of different persons talking about a vast variety of topics. It was then used as a stimuli to collect many others’ listening behavior to reduce the individual difference, as mentioned in the next section.

3.1 Recording Procedure

A total number of 22 participants (13 males) were recruited in the recording procedure. All participants were students from Beihang University with ages between 21 and 28. They were divided into 11 pairs, in all of which the two participants were familiar with each other beforehand. Before the recording, the participants were informed about the goal of this study, and were told that their task was to perform face-to-face Mandarin spoken conversations. No topic limitation was applied since the corpus only serves as the stimuli rather than comparison targets. To minimize the awkwardness of talking with certain constraints in a controlled environment, we only applied few basic restrictions to the conversations:

- The participants can act naturally, as long as their actions are recorded by the camcorder.
- This is a turn-based conversation. While one participant is speaking, the other acts as the listener and shall perform valid backchannels to the speaker. Their role can change at will.

A scenario of the conversation recording is shown in Fig. 1. Two digital cameras are put between the speaker and the listener so as to record the facial movements of both participants and not to disturb the interaction. Each participant is wearing a head-mounted microphone to collect speech. The cameras and microphones are connected to the same computer. The two video sources are recorded separately, but synchronized for convenience.

Video collection begun when both participants agreed and commenced talking. The recording continued until one or both participants felt like stopping. The result is a series of raw video materials, with a total length of 410 minutes.

3.2 Video Segmentation and Annotation

Since we had made minimal restrictions to video recording procedure, the raw video materials needed post-processing to meet quality criteria. We truncated parts that any of the two participants felt they performed poorly or unnatural. Then we extracted the speakers’ long spontaneous speeches into separate video fragments. Finally, 84 fragments with an
average length of 47.3s were produced.

The video annotation is accomplished using ELAN [23]. ELAN is a professional tool for the creation of annotations on video and audio resources. Annotations can be created on multiple interconnected layers and time-aligned to the media. It is easy to export the annotations to a text file. We used ELAN to annotate start time and end time of the speaker’s smile, head movements and the listener’s backchannel in the video fragments. Figure 2 shows a screenshot of using ELAN to create annotations.

4. Data Collection and Analysis

Huang et al. suggested that face-to-face interaction corpus could not meet the demand for virtual agent’s backchannel prediction modeling since there are shortcomings in applying the original listener’s behavior because of the optional nature of backchannel [22]. This nature affects backchannel timing in both frequency and occurring occasions. The PCS method allows multiple individuals to experience the same social situation by watching the pre-recorded face-to-face interaction videos. Then reactions of all the individuals are collected together as the consensus view of interaction behaviors. The PCS method can collect a large amount of behavior data from multiple individuals and use the consensus data to reduce individual difference limitations in backchannel responses, compensating both frequency and occasion bias.

4.1 Data Collection with the PCS Method

The PCS method is used to obtain more appropriate backchannel timing from multiple listeners. Thirty-two graduate students (15 females and 17 males, not involved in the video recording procedure) with ages between 20 and 32 participated in the data collection. They were requested to watch the speaker video fragments extracted in Sect. 3 and click a button as soon as they felt like to give a backchannel. Though one cannot express specific backchannel style during this process, since we focus on the timing prediction, this is acceptable. All the clicking time were recorded by the time-stamp.

The participants were asked to assume that the speakers in the videos were having a face-to-face talk with them. Before the collection, they took brief training about the data collection interface. The process began when the participant reported that one was able to fluently press the backchannel button with no hesitation or awkwardness. We repeated each video fragment twice to give participants more time to familiarize with the speakers’ topic, and only collected data in the second time.

Although the participants click the button as soon as possible, there is still a time delay between the real backchannel time and the collected clicking time. The time delay is understandable and inevitable as cognition delay. We compared the starting time of PCS-collected backchannels with the original ones to find out the time delay to be approximately 200ms, which is the same as the delay proposed by Poppe et al. when they analyzed the influence of quantity, type and timing for backchannel generation of virtual listeners [24]. The final PCS data are obtained by eliminating 200ms delay from the collected clicking time.

4.2 The Analysis of the PCS Data

Pragmatics researchers have conducted a lot of studies on backchannel behaviors of human. Clancy et al. investigated the conversational use of backchannel feedback in English, Japanese and Mandarin [25]. They summarized that conversational backchannels in the three languages differed in frequency and placement of occurrence. The frequency of backchannel in Mandarin appeared lower in comparison with English and Japanese. They also found Japanese and English speakers used backchannels more than three times as frequently as Mandarin speakers. The placement of backchannel feedback in Mandarin had its distinguishing characteristics. They found Mandarin speakers would not tend to place backchannels in the middle of the primary speaker’s clause. The interaction styles in Mandarin favor people not breaking in the other’s speaking turn because of the implied and introverted characters of Chinese people.

Table 2 shows the average, highest and lowest frequencies of backchannels occurring in the speakers’ pauses, and the original listeners’ for contrast. The average frequency
of PCS-listeners’ is 92.65%, which is slightly above the original listeners’ (we counted both verbal and non-verbal ones). The above analysis of the PCS data indicates that backchannels occurring in the speaker’s pauses form a vast majority, thus prediction of these backchannels is well worth studying.

When further analyzing backchannels not occurring in the speakers’ pauses (occurring during the speakers’ speech), we notice that most of them are occurring with several key words. For example, the listeners (both the original listeners and PCS listeners) prefer to give a backchannel before key words; such as, ‘ran hou’ (means ‘and then’) or ‘jiu shi’ (means ‘that is’). In other cases, backchannels are reactions to what the speaker is talking about. For instance, most male listeners produce a backchannel when the speaker is talking about a girlfriend. Although key words and content of conversations are helpful to backchannel prediction, they are not included in current investigation and will be discussed in the future.

5. The Analysis of Multimodal Features

Literature indicated there were various features used to predict backchannel feedback in English and Japanese. Existing research results should be fully used in current Mandarin backchannel prediction while a few adjustments of them are required to accommodate Mandarin conversations. This section introduces the analysis of multimodal features.

5.1 Prosodic Features

Prosody is a spoken feature of speech. It refers to the rhythm, stress and intonation. Prosodic features have been widely accepted as important evidence in deciding when to introduce backchannels. Cathcart et al. proposed that backchannels were often produced after a short pause in the speaker’s discourse [26]. According to the discussion in Sect. 4.2, backchannels occurring in the speaker’s pauses show high frequency among all backchannels in Mandarin. Pause duration is considered as one of the prosodic features for backchannel prediction.

Koiso et al. [27] and Ohsuga et al. [28] found that backchannels occurred following particular pitch and power contour patterns in the last mora of an utterance. Kitaoka et al. used first-order regression coefficients of pitch and power contours to describe patterns and generate response timing [29]. Nishimura et al. pointed out that both the last short regions and the longer ones contained information which triggered backchannel responses [30]. Thus, first-order regression coefficients of pitch and power contours in both the last 90ms and the last 500ms of the utterances are adopted to represent the changing trend of prosody. The last 90ms is divided into three parts with 50ms each part and no overlaps. First-order regression coefficients of pitch and power in all parts are calculated.

Ward and Tsukahara proposed five rules for backchannel prediction in English and Japanese conversations [20]. The rules have been generally accepted and applied in predicting backchannels for virtual agents. Rules for English are shown in Table 1 (mentioned in Sect. 2). The exact parameters of these rules were chosen to be consistent with their corpus data as closely as possible. These rules are also found to be effective in the pre-recorded multimodal Mandarin conversation videos. After experimental validation based on the pre-recorded video data, optimal parameters for Mandarin are set to: P1=1.5 \times \text{avg} (a region of pitch under 1.5 \times \text{avg}, where \text{avg} is the average pitch level of the utterance), P2=110ms (continuing for at least 110ms) and P5=330ms (after 330ms wait). These values are obtained by iterative testing with the Mandarin conversation videos. Parameters P3 and P4 are ignored because Ward and Tsukahara mentioned that conditions P1, P2 and P5 mainly expressed the core of the rules [20].

5.2 Visual Features

Visual evidences are as important as prosodic features in the multimodal interactions. Maatman et al. presented a mapping from head movements and posture shifts to the agent’s listening behaviors [31]. In the pre-recorded Mandarin conversation videos, the listeners always smile in response to the speaker’s smile and produce backchannels triggered by the speaker’s head movements. For the above reasons, the speaker’s smile and head movements (including nod and shake head) are used to predict backchannels.

6. Building Backchannel Prediction Models

This section introduces how to built models for Mandarin backchannel prediction. By analyzing the PCS data, the minimum duration of pauses associated with backchannel is measured to be 150ms. Then 26 speaker video fragments are divided into smaller segments based on pauses longer
than 150ms. For prediction, model-based method is chosen instead of rule-based method. Support Vector Machine (SVM) is applied to classify whether it is possible to give a backchannel in the pause of each video segment.

6.1 Training Features

In consideration of the analysis in Sect. 5, the following features for each segment are used in backchannel prediction:

- **Pause Duration**: duration of the speaker’s pause
- **Short Contour**: first-order regression coefficients of pitch and power contours in the last 90ms (divided into three parts)
- **Long Contour**: first-order regression coefficients of pitch and power contours in the last 500ms (divided into five parts)
- **Pitch Start**: duration from the start of the speaker’s utterance to the start of low-pitch region
- **Pitch Duration**: duration of the low-pitch region
- **Pitch End**: duration from the end of low-pitch region to the end of the speaker’s utterance
- **Smile Duration**: duration of the speaker’s smile
- **Head Duration**: duration of the speaker’s head movement

In order to test the performance of the models with optimal features, Pause Duration is manually annotated. In this way, samples used to train and test the SVM model are segments that consist of the annotated pauses and the preceding utterances.

First-order regression coefficients of pitch and power contours are calculated by Matlab scripts.

Feature Pitch Start, Pitch Duration and Pitch End are proposed based on the rules of Ward and Tsukahara which have been discussed in Sect. 5.1. Low-pitch regions are regions which satisfy the Mandarin parameter P1 (1.5 × avg) and P2 (110ms). In other words, low-pitch regions are regions with pitch less than 1.5 times the average pitch level and continuing for at least 110ms. However, not all the low-pitch regions are used to extract feature Pitch Start, Pitch Duration and Pitch End. Parameter P5 (330ms) is applied to sift the low-pitch regions. When preparing features for the samples, only low-pitch regions which end 330ms before the end of utterances are used. According to the rules of Ward and Tsukahara [20], these low-pitch regions provide effective evidences for backchannel prediction.

Duration of the speaker’s smile and head movement are obtained from manual annotation using ELAN.

6.2 Training Labels

As Huang et al. mentioned [22], probabilities of producing backchannels in speakers’ segments were obtained from the PCS data. The probability is determined by the number of participants agreeing to give a backchannel in that segment. When preparing experimental data from PCS, a threshold is set to filter out backchannels with low probabilities. The selected PCS threshold influences the expressiveness of the virtual agent listener; the lower threshold leads to a more expressive agent. However, the agent listener is not expected to be too expressive. While the original listener gives backchannels in a natural situation, we want the virtual agent to respond as naturally as the original speakers do. The number of agent listener’s backchannels should be closest to that of the original listener’s backchannels. Finally, the PCS threshold is set to 30% as shown in Fig. 4, which means the number of backchannels agreed by more than 30% of the participants is closest to that of the face-to-face interaction data.

For training labels, segments which produce backchannels are labeled as positive and ones not giving backchannels as negative. Therefore, segments containing backchannels agreed by more than 30% of the participants are labeled as positive for the SVM training.

6.3 Building the Prediction Models

Four prediction models are trained to compare how different features influence the prediction results. The models are described as follows.

- **Pitch and Power Contours Only** The first model is trained with feature Short Contour and Long Contour mentioned in Sect. 6.1. This model is referred as Contours model in the rest of the paper.
- **Low-pitch Regions Only** The second model is trained with feature Pitch Start, Pitch Duration and Pitch End mentioned in Sect. 6.1. This model is referred as Low-pitch model in the rest of the paper.
- **Prosodic Features** In the third model, in order to evaluate the prediction performance of prosodic features, we use all the prosodic features including Pause Duration, Short Contour, Long Contour, Pitch Start, Pitch Duration and Pitch End. This model is referred as Prosodic model in the rest of the paper.
- **Multimodal Features** The last model is trained with all the features mentioned in Sect. 6.1 to evaluate the prediction performance of multimodal features. This model is referred as Multimodal model in the rest of the paper.
Table 3  The performance of the four models.

| Model      | Precision | Recall   | F1     |
|------------|-----------|----------|--------|
| Contours   | 0.6022    | 0.6852   | 0.6146 |
| Low-pitch  | 0.6417    | 0.7000   | 0.6629 |
| Prosodic   | 0.6016    | 0.7189   | 0.6537 |
| Multimodal | 0.6947    | 0.6220   | 0.6566 |

As the number of the speaker’s smile and head movement is limited compared to prosodic features, there is not a separate model trained by visual features. Visual features are combined with prosodic features in the Multimodal model.

The SVM classifier is applied to build the prediction models because of its excellent performance on generalization and classification. The models can classify the segments in which the virtual agent listener should give a backchannel. All the speakers’ video segments are randomly divided into 5 folds, 4 of them are used as the training set and the remaining 1 is used as the testing set. This process repeats 5 times so that each segment is used as the testing data once. In SVM classification, the feature values are normalized into [-1, 1] and the radial basis function (RBF) kernel is used. Experimental results are obtained by 5-fold cross-validation and shown in Table 3.

6.4 Results

During the experiment, four prediction models are evaluated. Among them, the Low-pitch model gets the highest F1 (F1 = 0.6629), which indicates that low-pitch regions are the most important features for backchannel prediction in Mandarin. The Prosodic model achieves the best recall (recall = 0.7189) which indicates that pause and contour features help a lot in finding more correct results, as recall is a measure of completeness. Finally, the Multimodal model achieves comparable performance (F1 = 0.6566) and the highest precision (precision = 0.6947) which demonstrates that visual features indeed contribute to improve precision in multimodal prediction.

7. Subjective Evaluation

In order to further evaluate the performance of the proposed prediction models, a subjective evaluation experiment is conducted to assess whether the virtual agent driven by the models can be accepted by the users. The prediction models are compared with an original listener model and a random model. We expect the Multimodal version to achieve best performance among other prediction models since it takes more information from the speaker into account.

7.1 User Study

Videos which illustrate interactions between a human speaker and a virtual agent listener [32] are composed for the subjective evaluation experiment (shown in Fig. 5). Ten speaker video fragments are randomly selected from the pre-recorded conversation videos and used to compose videos. For each speaker video fragment, five versions of back-channel feedback are animated. All five versions use the agent listener’s head nods as backchannel feedback to the human speaker. The only difference between them is the timing. The five versions are introduced as follows:

- **Original**: the feedback timing of the virtual listener is the same as that of the corresponding human listener.
- **Multimodal**: the feedback timing of the virtual listener is predicted by the Multimodal model.
- **Low-pitch**: the feedback timing of the virtual listener is predicted by the Low-pitch model.
- **Prosodic**: the feedback timing of the virtual listener is predicted by the Prosodic model.
- **Random**: backchannel feedback of the virtual listener is randomly generated in the speaker’s pauses. The total number of the generated backchannels is the same as that of the original human listener’s backchannels.

The Contours model is not included in the subjective evaluation experiment because it is not outstanding in precision, recall or F1 score.

Twenty untrained participants are recruited to evaluate the performance of the virtual agent listener. There are 7 females and 13 males in the selected participants, all of whom are native Chinese speakers with ages between 22 and 34. They are all randomly chosen among the post-graduates in Beihang University, all unfamiliar with the human speakers in the video fragments.

They need to watch fifty composed video fragments (ten speaker video fragments, each speaker video is corresponding with five versions of virtual listener videos). Before the experiment, the participants were told that they were going to watch some videos showing conversations between a human speaker and a virtual agent listener; the virtual listener attempted to interact with the human speaker by occasional head nod. The participants had no idea about the details of the different versions. After watching each video, the participants evaluated the timing of the agent’s head nods by answering three questions:

- **Precision**: How often do you think the virtual listener nodded at an inappropriate time? (1 (always inappropriate) – 7 (never inappropriate))
Table 4 The subjective evaluation results.

|                | Precision       | Recall          | Overall Unhindered |
|----------------|-----------------|-----------------|-------------------|
| Original       | 4.848 ± 0.999   | 4.761 ± 1.160   | 5.677 ± 1.129     |
| Multimodal     | **5.003 ± 1.058** | **4.970 ± 1.127** | **5.736 ± 1.064** |
| Low-pitch      | 3.697 ± 0.937   | 5.075 ± 0.916   | 5.398 ± 0.937     |
| Prosodic       | 3.505 ± 1.014   | 4.954 ± 1.091   | 4.616 ± 0.924     |
| Random         | 3.547 ± 1.135   | 2.971 ± 1.027   | 3.227 ± 0.948     |

- **Recall**: How often do you think the virtual listener missed nodding opportunities? (1 (always missed) – 7 (never missed))
- **Overall Unhindered**: Do you think the virtual listener makes the dialogue more unhindered by using head nods in the conversation? (1 (no, not at all) – 7 (yes, absolutely))

7.2 Results

The mean and standard deviation of the evaluation scores are calculated to verify the validity of the model. The result is shown in Table 4.

7.2.1 Precision

As shown in Table 4, the Multimodal version outperforms all other versions in precision evaluation, with a score of 5.003. The Original receives 4.848 while being slightly lower. Meanwhile, the other three versions’ scores are all below 3.7, which is significantly lower compared to the Multimodal one.

For the Original version, the feedback timing of the virtual listener is the same as that of the corresponding human listener. Since the original human listener produces less backchannel feedback at inappropriate time, the precision score of the Original version is relatively high. The Multimodal model gets the highest precision in Sect. 6.4, which explains the reason why the Multimodal version outperforms the other three versions and achieves comparable performance with the Original version.

7.2.2 Recall

For the recall question, the Random version gets the lowest score of 2.971. The other four versions’ performances are close to each other and are all above 4.7, much higher than the Random one, according to Table 4.

The result is well explained by the fact that the Multimodal model, the Low-pitch model and the Prosodic model are trained from the PCS data which consist of appropriate backchannels from multiple PCS listeners. The Original version gets the lowest recall score in the four better versions, which indicates that the original listeners have missed some potential backchannels. This shortcoming has been overcome by using the PCS method, so the proposed prediction models miss less feedback.

7.2.3 Overall Unhindered

While comparing the overall unhindered score in Table 4, the Original version, the Multimodal version and the Low-pitch version received 5.677, 5.736 and 5.398, much higher than the Prosodic version and the Random version (4.616 and 3.227 respectively).

This result demonstrates that the virtual listener driven by the proposed prediction models can indeed make the conversation more unhindered, especially the Multimodal version, which has outstanding performance in all the above three questions.

8. Conclusion and Future Work

Recent advancements in human-computer interaction has brought about the need to endow virtual listeners with the ability to actively respond to human speakers to express attentiveness, rather than passively receiving orders. This paper discusses the modelling of the backchannel behavior of human listeners in Mandarin conversations as well as its application and evaluation on virtual agent listeners in a complete and thorough way.

Firstly, a corpus of multimodal dyadic Mandarin daily conversations is recorded as the base of the research. The corpus focuses on reflecting how human listeners would naturally act during a face-to-face interaction by applying as few restrictions to both sides as possible. The speakers’ smile and head movements as well as the listeners’ backchannels are located and annotated. Secondly, since using the corpus alone would bring individual differences into the backchannel model, responses from multiple human listeners are collected to eliminate such inconsistency via PCS algorithm. The elicited backchannel timings are then used together with several feature sets to train a series of classifier models which are capable of predicting backchannel feedback timings. Experiments on the testing data set show that the Multimodal version, which is based on both audio and visual features, performs the best. A subject evaluation with human participants is conducted to test the different prediction methods in action. During the test, the predicted timings are used to drive a virtual listener. The evaluation result agrees with previous experiments that the Multimodal version again brings the best overall performance amongst others.

This coincides with our assumption that more input cues give better backchannel timings, which is intuitively correct and proves that Mandarin backchannel feedback also relies on both audio and visual cues. On the other hand, albeit the prediction models that take only audio features into account act not as well as the Multimodal one, they still perform decently. Actually, their lack of dependency on visual input might come handy in applications such as telephone or voice assistants. The Multimodal version, however, can benefit from human speaker’s change in head pose or facial expression and yields a more accurate prediction in cases...
like training, entertainment or home care.

Future work includes natural language processing which involves key word detection and semantic understanding to predict backchannels occurring not only in the speaker’s pauses but also during continuous speech. Also, some limitations of the study involve the perception of psychophysiological assessment related to the feeling of being “in relationship” with the other party. Future studies will detect the psychophysiological arousal during the interaction. Furthermore, since different backchannel actions often carry different meanings, the diversity of backchannel behaviors and its impact on human speakers will be specifically modelled in further study.

Acknowledgements

This work is supported by the National Nature Science Foundation of China (No.610103097), National Research Foundation for the Doctoral Program of Higher Education of China (No.20121102130001) and the Fundamental Research Funds for the Central Universities (No.YWF-14-DZXY-015).

References

[1] A. Gulz and M. Haake, “Design of animated pedagogical agents—A look at their look,” Int. J. Human-Computer Studies, vol.64, no.4, pp.322–339, 2006.
[2] X. Mao and Z. Li, “Agent based affective tutoring systems: A pilot study,” Computers & Education, vol.55, no.1, pp.202–208, 2010.
[3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Embodied Conversational Agents, The MIT Press, 2000.
[4] Z. Li and X. Mao, “EEMML: The emotional eye movement animation toolkit,” Multimedia Tools and Applications, pp.1–21, 2011.
[5] R. Beale and C. Creed, “Affective interaction: How emotional agents affect users;” Int. J. Human-Computer Studies, vol.67, no.9, pp.755–776, 2009.
[6] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B.D. Carolis, “From greta’s mind to her face: modelling the dynamics of affective states in a conversational embodied agent;” Int. J. Human-Computer Studies, vol.59, no.1–2, pp.81–118, 2003.
[7] M. Ullman, “Contributions of memory circuits to language: The declarative/procedural model,” Cognition, vol.92, no.1, pp.231–270, 2004.
[8] A.T. Dittmann and L.G. Llewellyn, “Relationship between vocalization and head nods as listener responses;” J. Personality and Social Psychology, vol.9, no.1, p.79, 1968.
[9] V. Yngve, On getting a word in edgewise, pp.567–577, Chicago Linguistic Society, 1970.
[10] S. Duncan, “On the structure of speaker–auditor interaction during speaking turns,” Language in Society, vol.3, no.02, pp.161–180, 1974.
[11] J. Bavelas, L. Coates, and T. Johnson, “Listeners as co-narrators;” J. Personality and Social Psychology, vol.79, no.6, p.941, 2000.
[12] R. Nishimura and S. Nakagawa, “Response timing generation and response type selection for a spontaneous spoken dialog system;” Eleventh Biannual IEEE Workshop on Automatic Speech Recognition and Understanding, pp.462–467, 2009.
[13] L. Huang, L. Morency, and J. Gratch, “Virtual Rapport 2.0;” Intelligent Virtual Agents, pp.68–79, Springer, 2011.
[14] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pamm, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wöllmer, “Building autonomous sensitive artificial listeners,” IEEE Trans. Affective Computing, vol.3, no.2, pp.165–183, 2012.
[15] Y. Sakai, Y. Nonaka, K. Yasuda, and Y. Nakano, “Listener agent for elderly people with dementia;” Proc. Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp.199–200, 2012.
[16] I. de Kok and D. Heylen, “The MultiLis corpus—dealing with individual differences in nonverbal listening behavior,” Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues, pp.362–375, 2011.
[17] N.G. Ward and J.L. McCartney, “Visualizations supporting the discovery of prosodic contours related to turn-taking,” Feedback Behaviors in Dialogue, pp.85–89, 2012.
[18] X. Mao, N. Luo, and Y. Xue, “Emphasizing on the timing and type: enhancing the backchannel performance of virtual agent;” Proc. 4th International Conference on Agents and Artificial Intelligence, pp.259–263, 2012.
[19] R. Poppe, K. Truong, D. Reidsma, and D. Heylen, “Backchannel strategies for artificial listeners,” Intelligent Virtual Agents, pp.146–158, 2010.
[20] N. Ward and W. Tsukahara, “Prosodic features which cue backchannel responses in English and Japanese,” J. Pragmatics, vol.32, no.8, pp.1177–1207, 2000.
[21] L. 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, 2010.
[22] L. Huang, L. Morency, and J. Gratch, “Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior,” Proc. 9th International Conference on Autonomous Agents and Multiagent Systems, pp.1265–1272, 2010.
[23] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “ELAN: A professional framework for multimodality research,” Proc. LREC, pp.1556–1559, 2006.
[24] R. Poppe, K. Truong, and D. Heylen, “Backchannels: Quantity, type and timing matters,” Intelligent Virtual Agents, pp.228–239, 2011.
[25] P. Clancy, S. Thompson, R. Suzuki, and H. Tao, “The conversational use of reactive tokens in English, Japanese, and Mandarin,” J. Pragmatics, vol.26, no.3, pp.355–387, 1996.
[26] N. Cathcart, J. Carletta, and E. Klein, “A shallow model of backchannel continuers in spoken dialogue,” Proc. Tenth Conference on European Chapter of the Association for Computational Linguistics, pp.51–58, 2003.
[27] H. Koiso, Y. Horiiuchi, S. Tutiya, A. Ishikawa, and Y. Den, “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs,” Language and Speech, vol.41, no.3–4, pp.295–321, 1998.
[28] T. Ohgusa, M. Nishida, Y. Horiiuchi, and A. Ishikawa, “Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue,” Ninth European Conference on Speech Communication and Technology, pp.33–36, 2005.
[29] N. Kitaoaka, M. Takeuchi, R. Nishimura, and S. Nakagawa, “Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems,” J. Japanese Society for Artificial Intelligence, vol.20, no.3, pp.220–228, 2005.
[30] R. Nishimura, N. Kitaoaka, and S. Nakagawa, “A spoken dialog system for chat-like conversations considering response timing,” Text, Speech and Dialogue, pp.599–606, 2007.
[31] R. Maatman, J. Gratch, and S. Marsella, “Natural behavior of a listening agent,” Intelligent Virtual Agents, pp.25–36, 2005.
[32] M. Mancini and C. Pelachaud, “Implementing distinctive behavior for conversational agents,” Gesture-Based Human-Computer Interaction and Simulation, pp.163–174, 2009.
Xia Mao was born in Yiwu, Zhejiang province, China in 1952. She received her M.S. degree and Ph.D. degree from Saga University, Japan in 1993 and 1996 respectively. She is currently a professor at School of Electronic and Information Engineering, Beihang University, Beijing, China. Her current research interests include affective computing, artificial intelligence, pattern recognition and Human-Computer Interaction. So far, she has published over 140 pieces of papers both domestically and overseas, many of them have been cited by the SCI, EI, ISTP etc. Dr. Mao is leading several projects supported by the National High-tech Research and Development Program (863 Program), National Natural Science Foundation and Beijing Natural Science Foundation.

Yiping Peng received B.Sc. from and is currently a post-graduate at School of Electronic and Information Engineering, Beihang University, Beijing, China. His research interests are virtual agents in human-computer interaction and speech emotion recognition.

Yuli Xue received her B.S. degree and Ph.D. degree in School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2003 and 2009 respectively. Her current research interests include human computer interaction, pattern recognition, and image processing.

Na Luo received B.Sc. from and is currently a post-graduate at School of Electronic and Information Engineering, Beihang University, Beijing, China. Her research interests are virtual agents in human-computer interaction and speech signal processing.

Alberto Rovetta was appointed Full Professor in the Group of Machine applied Mechanics with a chair in Robot Mechanics, Politecnico di Milano. From 1991 he worked as a Professor in Robot Mechanics. He is the coordinator of the Robotics Laboratory of Politecnico di Milano, one of whose invention was the first voice-controlled robot, Gilberto. His current research fields are emotional interaction in telecommunications, nanotechnologies as well as environmental robotics.