ABSTRACT
Virtual humans, with realistic behaviors and increasingly human-like social skills, evoke in users a range of social behaviors normally only seen in human face-to-face interactions. One of the key challenges in creating such virtual humans is giving them human-like conversational skills. Traditional conversational virtual humans usually make turn-taking decisions depending on explicit cues, such as "press-to-talk buttons", from the human users. In contrast, people decide when to take turns by observing their conversational partner's behavior. In this paper, we present a multimodal end-of-turn prediction model. Instead of recording face-to-face conversations, we collect the turn-taking data using Parasocial Consensus Sampling (PCS) framework, where participants are guided to interact with media representation of people parasocially. Then, we analyze the relationship between verbal and nonverbal features and turn-taking behavior using the consensus data and show how these features influence the time people use to take turns. Finally, we present a probabilistic multimodal end-of-turn prediction model learned from the consensus data, which enables virtual humans to make real-time turn-taking predictions. The evaluation results show that our model achieves a high accuracy and takes human-like pauses, in terms of length, before taking its turns. Our work demonstrates the validity of Parasocial Consensus Sampling and generalizes this framework to model turn-taking behavior.

1. INTRODUCTION
Virtual humans play important roles in today's immersive virtual worlds. With realistic behaviors and increasingly human-like social skills, virtual humans evoke in users a range of social behaviors normally only seen in human face-to-face interactions. Previous studies have shown the effectiveness of virtual humans is enhanced by their realism [32] and has practical benefit for a host of applications, such as training [26], education [25] and health care [28].

One of the key challenges in creating virtual humans is giving them human-like conversational skills. Human conversation is a cooperative and fluent activity. The speaker monitors feedback from the listener and adjusts his behaviors accordingly while producing utterances; the listener provides moment-to-moment feedback to alter and serve to co-construct the subsequent speech while processing the utterances from the speaker. People rarely speak simultaneously. Rather, the roles of speaker and listener are regulated seamlessly by a negotiation process of turn-taking. And considerable research is directed at understanding this mechanism [1-12] and integrating it into virtual humans [15-18].

The fluidity of natural conversation presents a considerable challenge for virtual humans. On one hand, communication is multimodal: information is manifest in different channels, such as speech and gaze, and these channels may unfold under different time scales. Many existing conversational agents focus on a single channel (e.g. speech) [13,14,20] or depend on hand-crafted rules based on literature to integrate information from different channels [15-18]. On the other hand, effective communication involves forecasting what one's conversational partner will do in the future. For example, Sacks et al. [1] argued that the smooth exchange of turns in conversations is due to the conversational partner's ability to anticipate when the transition of speaker and listener roles may occur so that they are prepared in advance to talk at the right moment. However, many conversational virtual humans depend on explicit cues (e.g. press-to-talk buttons) from the human users to take turns without making predictions based on the conversational partner's behavior. This rigid turn-taking mechanism may disrupt the smooth exchange of turns in conversation and make the interaction unnatural.

This paper makes two primary contributions. First, we present a multimodal end-of-turn prediction model, drawing on prior findings from social psychology and linguistic literature on nonverbal signals and turn-taking behavior [1-12]. The model significantly outperforms other methods that have been proposed in the literature. Second, we demonstrate the effectiveness of a novel methodology for learning such models. We learn this model automatically from data collected using the technique of Parasocial Consensus Sampling (PCS) recently proposed by Huang and colleagues [21]. Previously, PCS has been successfully applied to the problem of predicting listener backchannel feedback [21,22]. Here we reinforce the viability of PCS by demonstrating its effectiveness on the novel domain of end-of-turn prediction.

In the remainder of the article we review previous literature on turn-taking and the PCS framework. We then describe how we collected a corpus of turn-taking data using PCS. Next, we analyze the relationship between verbal and nonverbal features and turn-taking behavior using the consensus view and show how these features influence the turn-taking time. Finally, we present a probabilistic multimodal end-of-turn prediction model learned from the parasocial consensus. This model is designed to enable virtual humans to make real-time turn-taking predictions. It incorporates features from different channels and models their influence on turn-taking behavior. Evaluation results show that our model achieves a high accuracy and has the human-like pauses, in terms of length, before taking its turns.
1.1 Related Work

Turn-taking has its root in conversation analysis. Much work [1,5,6] in this area focused on analyzing the syntactic, pragmatic and semantic structure of turns. Sacks et al. [1] proposed the idea of turn-constructional unit and considered the completion of a turn-constructional unit as a possible place to take the turn. They argued that it is the projectable completion of turn-constructional units that enables human beings to anticipate the end of turns. Duncan et al. [2] suggested the completion of grammatical clauses and special phrase, like “you know”, as useful turn-taking cues. Ford and Thompson [6] further stated that syntactic completion points, when co-occurring with intonational and pragmatic completion points, are always places of turn transition. Therefore, completed syntactic structures like “sentences, clauses, phrases, and even one-word constructions” [1] are considered as an informative cue for predicting turn-taking.

Besides verbal information, prior research has also conducted a variety of experiments exploring the relationship between nonverbal behaviors and turn-taking. Most behaviors examined belong to one of the following categories: prosody, eye gaze or head gesture.

Duncan et al. [2] found that a rising or falling pitch at the end of a sentence serves as a turn-taking signal; similarly, Beattie [7] claimed that a falling intonation pattern at the end of a clause always indicates the end of a turn. However, Caspers [8] argued that intonation cues are not generally used as turn yield signals; rather, the combination of a rising pitch followed by a high level boundary tone is used as a turn-holding signal. Models based on prosody information are used in several conversational systems to address the problem of end-of-turn detection. Jonsdottir et al. [13] presented a reinforcement learning model that learned the optimal pause duration by relying on pitch slope and pitch value, and that could take turns with human-like speed and reliability. However, their model was not evaluated on real conversation data and whether the same performance can be achieved in practice remains unclear. Raux et al. [20] proposed an algorithm to optimize the pause thresholds using prosodic and dialogue information and applied it to a real world dialogue system, reducing its latency by up to 24%. Schlangen [14] used acoustic features together with syntactic features to predict end-of-turn. He gradually reduced the length of the pause threshold, turning the problem from detection to prediction. When the pause threshold was reduced to 0, the F1 score reduced to 0.355. These systems demonstrated that prosody features serve as informative cues for turn-taking prediction.

In the visual channel, gaze has been observed to be a useful way to coordinate turns in conversation. When one conversant finishes his current turn, he will often look towards the other; later, the established mutual gaze between them will be broken by the other conversant when she begins to talk. This is called “mutual gaze break” [9]. It is found that looking-towards (“gaze back”) and looking-away (“gaze aversion”) are correlated to end of turn and beginning of turn, respectively. In [9], this pattern occurs at approximately 42% of the turn exchanges. Several virtual agent systems have incorporated such gaze patterns to coordinate the conversation using rules like “speaker looks away from the hearer at the beginning of a long turn and looks towards the hearer at the end of the turn” [15,16]. Head movements sometimes accompany the change of eye gaze and they also act as turn-yielding signals [3]. Speakers often make tiny head nods or shakes at the end of turn to elicit confirmation from the listener [10], or to signal to the listener the turn-taking channel is open again.

Given the wide variety of potential turn-taking cues, one might expect that combining information from both auditory and visual channels should lead to the best results for end-of-turn prediction. Thorisson's Gandalf agent [17] was a pioneer among multimodal turn-taking models. It is a layered architecture with several update loops, each designed for a different sensor and operating at a different speed. The model combined unimodal features according to some heuristic rules and the turn-taking decisions were made by a rule-based model. Subsequently, Bohus and Horvitz [18] proposed a computational framework for modeling and managing turn-taking in multi-party interaction, leveraging audio-visual and contextual information to make real-time decisions. Currently, their system relies on a set of rules that take into account the basic turn-taking norms and turn-taking context, such as the current floor state and the intention of each participant. Recently, de Kok et al. [19] demonstrated the possibility to apply machine learning techniques to fuse features from multiple channels to predict end-of-turn in multi-party meetings, though with mixed results. In this paper, we take advantage of both the multi-modalities approach and the data-driven approach to learn a probabilistic multimodal end-of-turn prediction model for conversational virtual humans.

1.2 Background: Parasocial Consensus Sampling

Traditionally, virtual humans learn from annotated recordings of face-to-face interactions. However, as suggested in [21], there are some drawbacks with such data. For example, human behavior contains variability and not all human data should be considered as positive examples of the behavior that the virtual human is attempting to learn. If the goal is to make the virtual human learn to take turns properly, it is necessary to realize that many face-to-face interactions fail in this regard, resulting in interruptions or long mutual silence. To address this and other issues, Huang et al. [21] proposed a new data collection framework called Parasocial Consensus Sampling (PCS) that results not only in examples of a target behavior (such as turn-taking), but also probability ratings that characterize the quality of specific examples.

The theory of Parasocial Consensus Sampling is based on the concept of parasocial interaction which was first proposed by Horton and Wohl [27], where they argued humans tend to interact with media representation of people as if they were interacting face-to-face with another person. PCS was proposed by exploiting this phenomenon. In this framework, instead of recording face-to-face interactions, participants are guided to interact with media representation of people, such as pre-recorded speaker videos, parasocially. In such way, multiple independent participants are able to experience the same social situation and provide parasocial responses to the same event. Later, these individual responses are aggregated into a consensus view of how a typical individual would respond in that given
situation. By eliciting multiple perspectives, this approach can help tease apart what is idiosyncratic from what is essential and help reveal the strength of cues that elicit social responses. PCS has been successfully applied to collect and model listener backchannel feedback, for example nods or paraverbs like "uh-huh", in face-to-face interactions [21,22]. In their experiments, participants were guided to interact with the pre-recorded speaker videos and asked to try to show interests to what has been told by providing backchannel feedback to the speaker. Whenever they felt like to give backchannel feedback, they pressed a button. Evaluations showed that the data collected in this approach can generate better backchannel feedback and can be used to learn a better backchannel prediction model.

Similar approaches have already been applied in previous work [11,12] to investigate turn-taking behaviors. Barkhuysen et al. [12] showed participants fragments of video clips and asked them to press a button as soon as possible to indicate that the speaker had finished his or her utterances. Ruiter et al. [11] performed a subjective evaluation to test the importance of prosody and syntax information in anticipating the end of turn. They presented fragments of conversations to participants and asked them to press a button at the moment they thought that the speaker had finished his or her turn. Their experiment showed that the performance of the participants is very similar to that of the real listeners in the original conversations regarding turn-taking speed, which proved the validity of their approach. In our work, we apply the PCS framework to model turn-taking behaviors.

2. Parasocial Consensus Sampling for Turn-taking Behavior
Parasocial Consensus Sampling consists of five key elements: interactional goal, target behavioral response, media, target population and measurement channel. We define them as follows for collecting turn-taking behavior.

Interactional Goal: This is the intended goal of the virtual human's interactional behaviors. In our work, we want the virtual human to learn a turn-taking strategy similar to that of a human and avoid interruptions and long mutual silences during the conversation.

Target Behavioral Response: This is the response that we want the virtual human to learn to generate. Here it is a prediction on when to take the conversational turn.

Media: This is the stimuli that will be presented to participants in order to stimulate their parasocial responses. We use videos taken from a series of interviews (described in Section 2.1).

Target Population: This is the population of individuals we wish the virtual human to approximate. In this experiment, participants are drawn from general public.

Measurement Channel: This is the mechanism by which we measure the parasocial responses, such as visual channel (e.g. videotaping), auditory channel (e.g. voice recording) or mechanical channel (e.g. press buttons). We ask the participants to press a button when they feel it is appropriate to take a turn.

2.1 PCS Data Collection
The videos used in our work are from an interview data set collected from a previous unrelated study on intimate self-disclosure [31]. In the dataset, participants recruited from the general population were asked a series of 10 increasingly intimate questions, either by a human confederate via a two-way video conference or by a virtual human interviewer controlled by a human confederate through a wizard-of-oz interface. In the virtual human condition, the interviewer pressed a button to retrieve pre-recorded voice messages when he wanted to take turns and the virtual human displayed corresponding lip-sync movements at the same time. In both conditions, the interviewer listened silently until they felt it was appropriate to proceed to the next question. All interviewee's behaviors were videotaped.

To strengthen the parasocial effect, we created an interactive interface that heightened the feeling that coders were engaged in an actual interview, although all coders knew they were interacting with pre-recorded media. This coding procedure is illustrated in Figure 2. Interviewee videos are displayed in a web browser. Coders press the “S” key to “ask” a question (upon pressing the key they hear an audio recording of a question asked in the original interview) and see the interviewee’s reactions as they respond to the question. At any point during the answer, the coder may press “Space” key to indicate when it is a proper point to interrupt. At this point, the interface automatically loads a new randomly selected question and waits till the coder presses “S”
again to ask this next question. If the coder does not press “Space” key, the video will keep playing until reaching the end. At this point the video ends and the coder is prompted to push “S” to ask the next question.

We recruited 9 participants from a local temporary agency. First, they signed confidentiality forms protecting the video materials they were about to watch. Then, they were taught how to use the interactive tool and provided instruction regarding the task: "You are going to watch some videos clips where an interviewee is answering questions. While watching, please press space bar as soon as you think the interviewee has finished his answer to the current question". Every participant watched all videos in one day. None of them reported difficulty in finishing the task or following the routines of our interactive tool.

2.2 Building Consensus

There are 46 videos in total in our data set, each contains 10 questions and was watched by all 9 participants. There are 4140 (46×10×9) turn-taking opportunities (i.e. if each participant took a turn at the end of the answer to each question). The turn-taking data we collected contains approximately 3300 records, which suggests that individual data contains variability. Not everyone agreed to take a turn at the end of each question. By combining all individual data together, we can develop a more reliable version of how a typical individual would take turns, which teases apart what is idiosyncratic from what is essential.

We compute a histogram of multiple participants’ responses to build the consensus. We did this by converting the video time line into samples at a sample rate of 30Hz. Each turn-taking point (the time when a participant pressed the button) is centered in a 1 second window, that is, a window of 30 samples. Whenever a turn took place on a sample, the histogram of that sample increases by 1. Therefore, each sample is associated with a number indicating the number of participants who agreed to take the turn here. Figure 1 shows an example of a parasocial consensus. The parasocial coders identified two possible turn-taking points (the one within the pause following “yoga:” and the other one within the pause following “music”), but they are associated with different probabilities. Clearly, the consensus view favors the second one probably due to the fact that the speaker is averting gaze during the first place. Overall, the histogram can be considered as a measurement of probability to take the turn. By setting a proper consensus level, we can separate the turn-taking places with low probabilities from the ones with high probabilities, which is better for further analysis and learning the prediction model.

The consensus level can be seen as a tool to remove outliers (i.e. the ones with low coder agreement) from the data. The higher the consensus level is, the fewer and more reliable turn-taking places will be selected. There are 46 videos in the data set, each containing 10 questions; therefore we have 460 turn-taking opportunities. Following the recommendation from [21], we choose the consensus level so that the number of the selected turn-taking from the consensus data is closest to 460. By testing on different values, the optimal consensus level is 3 with 479 turn-taking places.

3. Analysis of Multimodal Patterns

One potential advantage of PCS is that by automatically removing atypical responses, it may be easier to identify speaker cues that trigger or inhibit turn-taking. Before attempting to learn a model, we first explore the relative impact of the different turn-taking cues that have been proposed in the literature.

3.1 Feature Extraction

As described in Section 1.1, gaze, nods, prosody and syntactic features are all argued to impact turn-taking. In preparation for analysis, interviewee's answers were transcribed and gaze and head nods were manually annotated. Prosodic features (pause and pitch contour) are labeled automatically by Aizula [29]. We use similar intonation patterns as mentioned in [13], where the pitch contour in the most recent tail (last 300 msec) of speech right before a pause is used to find the slope and the average pitch value. The slope is categorized into 3 classes: Up, Straight, and Down; the average pitch value in the tail is categorized into Above, At, and Below by comparing with the average pitch value of the whole speech segment before that pause. The completion of syntactic structures is extracted from the transcripts. Since the basic units in the transcripts are always complete clauses, their boundaries can be considered as syntax completion points. We end up with 10 features in the analysis (See Table 1.)

3.2 Analysis

We classify pauses as a turn-taking pause if and only if a peak from the parasocial consensus occurs within the duration of the pause; otherwise we classify it as a non-turn-taking pause. There are 1434 pauses in our data, of which 422 are the turn-taking pauses. Of all turn-taking pauses, 95% are longer than 1.5s; while only 30% of non-turn-taking pauses are longer than that.

Our analysis is based on two measurements: (1) the frequency of co-occurrence between each of the 10 features and turn-taking and non-turn-taking pauses (as shown in Table 1) and (2) the average pause duration before taking a turn when one of the 10 features is present.
**Gaze:** As shown in Table 1, 59% of non-turn-taking pauses co-occur with looking-away, whereas the percentage is only 3% for turn-taking pauses. This suggests that looking-away is a very effective turn-holding signal [4]. By looking away, the speaker closes the turn-taking channel temporarily and thus can focus on recalling information while holding the floor. For turn-taking pauses, 27% of them co-occur with looking-towards; while the percentage is only 12% for non-turn-taking pauses. Therefore, looking-towards is not randomly distributed among pauses, it provides discriminative information. Interestingly, at turn-taking pauses, we find different gaze patterns influence the pause duration people wait to take turns. When the interviewee is staring at the interviewer, the average pause duration is about 1.3s; when looking-towards happens, the average pause duration is about 1.7s; in the few cases where looking-away happens, the average pause duration is about 2.0s. This suggests people employ different turn-taking strategies by observing different gaze patterns. Looking-away makes people hesitate to take turns; people make turn-taking decision more quickly when the interviewer is staring at him or her. Although looking-towards is a predictive cue for turn-taking, the average pause duration in this condition is not the shortest. We find people tend to wait for a period of time after the interviewee looks back. This may be because the participant needs to confirm there is no more information forthcoming after the interviewee's gaze aversion. The average waiting time is about 0.8s.

**Prosody:** As Table 1 shows, 73% of turn-taking pauses associated with rising or falling pitch contours; the percentage is only 46% for non-turn-taking pauses. This suggests that the slope of pitch contour is a predictive cue for turn-taking. This is in line with Duncan et al. [2] where they considered rising or falling intonation as one of the turn-yielding signals. For the average pitch value, we do not observe significantly discriminative information. At turn-taking pauses, 38% co-occur with low pitch, 14% co-occur with high pitch, and 48% co-occur with average pitch; a similar pattern occurs with non-turn-taking pauses, 32% co-occur with low pitch, 11% co-occur with high pitch, and 57% co-occur with average pitch. Similarly to the impact of gaze patterns on pause duration, different slope patterns influence the time people use to take turns as well. We find the average pause duration is 1.37s when the slope is down, 1.45s when the slope is up, and 1.56s when it is flat. People tend to be more hesitated to take turns when the slope is flat.

**Nod:** We find that head nods happen at 17% of turn-taking pauses but only at 8% of non-turn-taking places, as shown in Table 1. This suggests that head nod serves as a predictive cue for turn-taking. The average pause duration people wait to take turns is about 1.7s when head nod occurs, while it is shorter (about 1.4s) when there is no head nod. In our data, we observe that people usually wait until the head gesture is finished before taking a turn, which explains why the duration is longer when head nods appear. However, they immediately take the turn after the gesture finishes. The average gap between turn-taking and the end of a head nod is about 0.2s.

**Syntax:** Completed syntactic structures like "sentences, clauses, phrases, and even one-word constructions" [1] are considered to be possible turn-taking opportunities. In our data, the percentage of turn-taking pauses that co-occur with syntax completion is high, but there are also 64% non-turn-taking pauses following completed syntactic structures. In interviews, a turn is usually consist of several complete clauses; thus, many syntax completion points are inside a turn. Although it is a necessary condition for turn-taking, it is not sufficient enough.

**3.3 Summary**

By analyzing different turn-taking cues, we find the occurrences of looking-away, looking-towards and head nods are very informative cues for turn-taking prediction and they also impact the length of pause people wait to take turns. This is in line with previous studies [4,9-11]. Prosodic features provide useful information for turn-taking prediction as well, but we find slope is the only informative cue. This indicates prosodic features only may not be enough to predict end-of-turn. Syntax completion points co-occur with turn-taking pauses; however, they are not sufficient cues because a turn is usually consist of several complete clauses in our data. The analysis suggests that combining features from different channels should lead to the best results for turn-taking prediction. Therefore, all nonverbal features are included in the initial feature set for learning the prediction model. Syntax is not included because it is hard to get such information in real time.

### Table 1. The percentage of turn-taking pauses and non-turn-taking pauses that co-occur with different features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Turn-taking Pauses (422)</th>
<th>Non-turn-taking Pauses (1012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Looking-away</td>
<td>3% (11)</td>
<td>59% (598)</td>
</tr>
<tr>
<td>Looking-towards</td>
<td>27% (114)</td>
<td>12% (123)</td>
</tr>
<tr>
<td>Nods</td>
<td>17% (71)</td>
<td>8% (77)</td>
</tr>
<tr>
<td>Pitch Slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up</td>
<td>38% (160)</td>
<td>22% (227)</td>
</tr>
<tr>
<td>Down</td>
<td>35% (149)</td>
<td>24% (238)</td>
</tr>
<tr>
<td>Straight</td>
<td>27% (113)</td>
<td>54% (547)</td>
</tr>
<tr>
<td>Average Pitch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Above</td>
<td>14% (60)</td>
<td>11% (108)</td>
</tr>
<tr>
<td>Below</td>
<td>38% (162)</td>
<td>32% (321)</td>
</tr>
<tr>
<td>Value</td>
<td>At</td>
<td>48% (200)</td>
</tr>
<tr>
<td>Symtax Completion</td>
<td>98% (416)</td>
<td>64% (648)</td>
</tr>
</tbody>
</table>

The absolute number is shown in parentheses.

### 4. Multimodal End-of-Turn Prediction Model

The goal of the predictive model is to predict when virtual humans should take turns in real time. Conditional Random Field (CRF) [30] is used because of its advantages in modeling the sequential aspects of human behavior. It takes as input a sequence of observations, such as gaze, nods and prosody, and returns a sequence of probabilities of states (e.g. taking turn or not). Human behavior is not always synchronized. For example, the listener may decide to take a turn because the speaker looks-towards him 0.8s ago. To capture this asynchrony, a feature encoding technique [23] is used to encode nonverbal features to model the long-range dependency (Section 4.1). Further, an automatic feature selection algorithm [24] is utilized to find the most discriminative features to reduce the overfitting problem that is usually due to the high complexity of model and limited training data (Section 4.2). By incorporating the two techniques in the learning process, we build a probabilistic multimodal end-of-turn prediction model (Section 4.3). The evaluation results
show our model achieves a high accuracy and takes human-like pauses, in terms of length, before taking its turns (Section 4.4 and 4.5).

4.1 Feature Encoding

Morency et al. [23] first proposed the concept of encoding dictionary in listener backchannel prediction. By using different encoding dictionaries, like binary, step and ramp, their model successfully captured the different relationship between speaker features and listener backchannel feedback. The three encoding dictionaries that are used in our model are discussed briefly below. Please refer to [23] for details.

**Binary Encoding:** The value of the encoded feature is 1 between the start time and the end time and 0 elsewhere.

**Step Encoding:** This encoding generalizes binary encoding by adding two parameters: width and latency. Width represents the length of the encoded feature and the latency represents the delay between the start of the feature and its encoded version.

**Ramp Encoding:** This encoding is similar to step encoding except that the encoded value increases within a period of time (width) linearly.

To apply encoding dictionaries, the only needed information is the starting time except the binary encoding. Binary encoding simply detects when a feature is on or off. In all cases, no future information is needed. The width and latency parameters for step and ramp encoding are pre-defined. Therefore, all feature encoding can be done in real time.

4.2 Feature Selection

The feature selection algorithm is based on regularization which is often used in machine learning problems to prevent overfitting. Regularization smooths the model by introducing a penalty term into the optimization function. The penalty will increase if the complexity of the model increases. This is useful when the model has a high complexity and the size of the training data is limited. One of the most used regularization forms is L1 regularization which is defined as:

\[ R(\theta) = \lambda \| \theta \|_1 = \lambda \sum \theta_i \]

where \( \theta \) is the model parameters (here they are the weights of features) and \( \lambda > 0 \). L1 regularization has been used for feature selection in nonverbal behavior analysis before [24]. The feature selection process starts with a high regularization parameter \( \lambda \) where all feature weights are zeros and then gradually reduces the value of \( \lambda \) until all weights become non-zeros. During the process, the features whose weights turn to non-zeros earlier are considered more important. The number of features is selected automatically by comparing the performance of models on the validation set.

4.3 Learning

The initial feature set contains looking-away, looking-towards, head nod, pause and pitch patterns. Each feature is sampled at a rate of 30Hz and encoded by 13 encoding. Using the feature selection algorithm described in section 4.2, 10 features are automatically selected as our final feature set. We randomly split all videos into 4 folds, 3 of them are used as developing set and the remaining one is used as testing set. The process repeats 4 times so that every video appears in the testing set once. Following [30], we train the CRF model and optimize L2 regularization term (\(10^3\), \( k=1..3\)) by applying 10-fold cross validation on the developing set during the training process.

The performance of the model is measured by F1 score, which is the weighted harmonic mean of precision and recall. Precision measures the percentage of correct turn-taking predictions among all predictions; recall measures the percentage of turn-taking places in the testing set that are correctly predicted. A prediction is considered correct if the predicted turn-taking occurs within the 1 second window around the ground truth.

4.4 End-of-Turn Prediction Models

We compare the performance of two models learned from PCS (PCS-Multimodal model and PCS-Pause model) with two baseline models (prosody model and syntax model) from previous literature.

**PCS-Multimodal:** This is the model we learned in Section 4.3.

**PCS-Pause:** The pause model is created by choosing an optimal length of pause duration. It classifies a pause to be a turn-taking pause if its duration is longer than the threshold. As analyzed in Section 3, the optimal pause threshold is selected to be 1.5s, which is the average time that the participants took to take turns during parasocial interactions.

**Prosody model:** Prosody model is trained the same way as PCS-Multimodal model but with only prosodic features. These prosodic features were based on [13].

**Syntax model:** Syntax model is based on the previous work of Sacks et al. [1], where syntax completion points, such as the end of "sentences, clauses, phrases, and one-word constructions", are suggested as possible turn-taking places. The syntax completion points are determined by the transcripts of the interviewee's speech. Since the basic units in the transcripts are always complete clauses, their boundaries can be considered as syntax completion points.

4.5 Results and Discussion

Predictions from each model are evaluated by two criteria. The first one is turn-taking pause prediction, where the results are compared with the confederate interviewer's turn-taking behavior. The predicted time is considered correct if happening during the same pause as when the confederate interviewer took a turn (see Section 2.1). One should note the confederate interviewer's turn-taking behavior is only one of the possible

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCS-Multimodal</td>
<td>0.78</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>PCS-Pause Model</td>
<td>0.59</td>
<td>0.90</td>
<td>0.71</td>
</tr>
<tr>
<td>Prosody Model [13]</td>
<td>0.58</td>
<td>0.77</td>
<td>0.67</td>
</tr>
<tr>
<td>Syntax Model [1]</td>
<td>0.29</td>
<td>0.97</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 2. Evaluations for Turn-taking pause prediction: F1 score of PCS-Multimodal is significantly better than that of the other three models.
ways to behave. That is why we also compare with the parasocial consensus which aggregates multiple coders.

The second criterion is turn-taking time prediction, where the results are compared with the turn-taking opportunities from the parasocial consensus (see Section 2.2). This is a more challenging criterion, since the predicted time is considered correct only if it is within the 1 second window around the turn-taking opportunities from parasocial consensus. Note that the syntax model cannot predict exact turn-taking time, since it is evaluated at the clause level. For this reason, we did not include it in this condition. We also briefly discuss the feature selection results from Section 4.2.

**Turn-taking pause prediction:** As Table 2 shows, F1 score of the PCS-Multimodal model is better than that of other three models. Paired T-Test comparisons between PCS-Multimodal model and the other three models (p = 0.05 for PCS-Pause, p < 0.01 for the other two) suggest the difference is statistically significant. This indicates syntax or prosody only cannot provide enough information to predict the turn-taking pauses. In our data, a turn is usually consist of several clauses, where the syntax completion points are not always followed by the transition of turns. As suggested in [5], turn-constructional units in some activities, like story-telling, where a turn is always consist of several of them, should be categorized into non-final and final ones and the end of non-final turn-constructional units are not followed by transition of turns. Although [13] showed prosodic features only are good at end-of-turn prediction, it is not the case in our data where the change of prosody may sometimes be caused by the completion of clauses inside a turn but not the end of turn. By leveraging the multimodal features, our PCS-Multimodal model performs the best.

**Turn-taking time prediction:** PCS-Multimodal model is significantly better than prosody model (p=0.02) but not PCS-Pause model in predicting turn-taking time. However, the PCS-Pause model cannot represent the flexibility of human behavior; that is, in real conversations, the turn-taking time varies based on the other conversational partner's behavior (see Section 3). PCS-Multimodal model is able to generate the human-like turn-taking strategy by taking into account the interviewee's nonverbal features. For example, in Figure 3, the interviewee is staring at the interviewer when she answers the question, the model waits for only 0.9 second to take the turn; in Figure 4, the interviewee looks towards the interviewer about 1 second after he finishes his answer, the model waits for about 2 seconds to take the turn. The variation of the predictions from PCS-Multimodal model matches the trend we find in Section 3.

**Feature Selection:** Interestingly, by looking back at the PCS-Multimodal model, we find the automatically selected features match our analysis in Section 3. For example, looking-away with binary encoding is selected, which matches the close correlation between looking-away and non-turn-taking pauses; the rising (Up) and falling (Down) pitch slopes are both selected, which matches the observation that these two patterns are correlated with turn-taking pauses. By applying the feature selection algorithm on the parasocial consensus, we are allowed to reveal the multimodal cues for turn-taking prediction in an automatic way.

These results show that our PCS-Multimodal model performs better than other models and takes human-like pauses, in terms of length, before taking its turns. It also demonstrates the validity of PCS framework in modeling human turn-taking behavior.

### Table 3. Evaluation for Turn-taking time prediction: F1 of PCS-Multimodal is better than that of the other two models.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCS-Multimodal Model</td>
<td>0.43</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>PCS-Pause Model</td>
<td>0.32</td>
<td>0.46</td>
<td>0.38</td>
</tr>
<tr>
<td>Prosody Model [13]</td>
<td>0.33</td>
<td>0.42</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Figure 3. Turn-taking time with staring gaze:** The interviewee is staring at the interviewer at the end of this answer, our model predicts the turn-taking time about 0.9s after the end of the answer and the participants in parasocial interaction took about 1.0s.

**Figure 4. Turn-taking time with looking-towards:** The interviewee looks towards the interviewer at the end of his answer, our model predicts the turn-taking time about 2.0s after the end of answer and the participants in parasocial interaction took about 1.9s.

5. **Conclusion**

In this paper, we presented a multimodal end-of-turn prediction model learned from parasocial consensus. By analyzing the consensus view of how a typical individual would take turns, we found multimodal features provide predictive information for learning the end-of-turn prediction model. The learned model can predict when conversational virtual humans should take turns in real time. It achieved a high accuracy in predicting the turn-taking pauses and had a human-like turn-taking strategy. By applying the Parasocial Consensus Sampling (PCS) framework in collecting and modeling turn-taking behavior, we validate this new methodology further and generalize it to turn-taking behavior modeling. In future, we are planning to explore other nonverbal behaviors, such as facial expressions and hand gestures, for turn-taking prediction.
6. REFERENCES


