Early Turn-taking Prediction with Spiking Neural Networks for Human Robot Collaboration

Tian Zhou and Juan P. Wachs

Abstract—Turn-taking is essential to the structure of human teamwork. Humans are typically aware of team members’ intention to keep or relinquish their turn before a turn switch, where the responsibility of working on a shared task is shifted. Future co-robots are also expected to provide such competence. To that end, this paper proposes the Cognitive Turn-taking Model (CTTM), which leverages cognitive models (i.e., Spiking Neural Network) to achieve early turn-taking prediction. The CTTM framework can process multimodal human communication cues (both implicit and explicit) and predict human turn-taking intentions in an early stage. The proposed framework is tested on a simulated surgical procedure, where a robotic scrub nurse predicts the surgeon’s turn-taking intention. It was found that the proposed CTTM framework outperforms the state-of-the-art turn-taking prediction algorithms by a large margin. It also outperforms humans when presented with partial observations of communication cues (i.e., less than 40% of full actions). This early prediction capability enables robots to initiate turn-taking actions at an early stage, which facilitates collaboration and increases overall efficiency.

I. INTRODUCTION

Turn-taking provides humans the fundamental structure to organize and coordinate their collaborative actions in conversations [1], collaborative task-solving [2], and shared haptic control [3]. During collaboration, each agent has the capability to comprehend the on-going task progress and their peer’s multi-modal communication cues, in order to predict whether, when, and how to size the next turn. Having a fluent and natural turn-taking process can enhance the collaboration efficiency and strengthen the communication grounding among team members [4], [5]. Such coordinated turn-taking behaviors stand out clearly in high-risk and high-paced tasks like surgery in the Operating Room (OR). Surgeons and nurses perform fast, fluent, and precise turn-taking actions when exchanging surgical instruments. Therefore, the OR scenario was selected as the test-bed for investigating human robot turn-takings.

The same turn-taking behaviors observed during human-human interactions are expected in human-robot interactions. When robots are designed to collaborate with humans, they are expected to understand the human’s turn-taking intentions, in order to determine when is a good time to engage in an interaction. When designing robotic nurses to collaborate with surgeons, they are expected to understand both the surgeon’s implicit communication cues (e.g., change of body posture like leaning forward) and explicit cues (e.g., uttering the word “scissors”) [6]. Figure 1 shows our human-robot system including the capability to understand both implicit and explicit communication cues for turn-taking.

A typical process where humans delegate the turn to a robot is illustrated in Figure 2. As the human is approaching the end of her turn, she starts exhibiting implicit cues (i.e., physiological and physical cues), followed by explicit cues (i.e., utterance). The collection of these multimodal signals indicates her intention to relinquish the turn. Simultaneously, the robot needs to capture these subtle communication cues and predict the end of her turn. Early and accurate predictions of such turn-giving intentions allow robots to begin preparatory movements to facilitate the turn transition process.

There has been research on developing algorithms to automatically recognize human’s turn-taking intentions, in both...
conversations [7] and physical collaborations [8]. However, current turn-taking recognition algorithms build on mathematically derived models such as Support Vector Machines [9], Decision Trees [10], and Conditional Random Field [7]. Although a certain level of recognition accuracy is achieved, it still has not reached the level of human competence [11]. Moreover, a model derived computationally and mathematically cannot be explained and interpreted well by humans due to different reasoning processes. Therefore, there is a need of a cognitive turn-taking reasoning model that reaches the level of human competence and can be easily interpreted.

To bridge this gap, the Cognitive Turn-Taking Model (CTTM) is proposed to simulate and predict human’s turn-taking intentions. The CTTM is built upon the brain-inspired Spiking Neural Network (SNN). The SNN framework has plasticity structure [12] to model turn-taking processes and predicts human’s turn-taking intentions based on multimodal observations. Compared to traditional neural networks, SNN has the advantage of modeling conduction delays of variable lengths. Therefore, it is suitable for time sensitive sequence modelling tasks such as gesture recognition [13], speech recognition [10], and seizure detection [14]. However, there are still challenges in using SNN for multimodal turn-taking prediction such as how to map the raw signals into neurons, how to deal with multimodal fusion problems, and how to create descriptive features for early prediction. Those challenges are tackled by the proposed CTTM framework.

To summarize, this paper makes the following three-fold contributions: 1) presents a formulation to rigorously define the human-robot collaborative task and the associated turn-taking events; 2) proposes the cognitive framework CTTM for early turn-taking prediction; 3) designs a multimodal interaction system between surgeons and robotic nurses.

The paper is organized as the following: Section II presents the related work, followed by definitions of the collaborative task and turn-events in Section III. The proposed CTTM framework is introduced in Section IV, and the experimental setup and results are presented in Section V. Finally, Section VI summarizes the paper with conclusions and future work.

II. RELATED WORK

This section presents the related work about turn-taking analysis, in the context of conversational turn-taking and physical turn-taking.

In conversational turn-taking analysis, linguistic cues such as pause duration [15] and pitch levels [16] have been found to play a key role in turn-taking regulation. The study of de Ruiter et al. [17] revealed that syntax and semantics cues are also important in projecting the end of a speaker’s turn. Non-verbal behaviors such as gaze and posture shifts have been investigated and found to be relevant to turn regulations [18].

Physical turn-taking refers to the process where a hybrid human-robot team take turns on a physical task. Turn-taking has been studied in robotics and HRI to regulate shared resources among team members. Those resources include time (i.e., only one person can work on the shared task at a time) and space (i.e., only one person can access the working space at a time). In the context of human robot interaction, the types and usage frequencies of various implicit communication cues have been studied in a robot-assisted assembly task [8]. The timing in multimodal turn-taking (i.e., speech, gaze, and gesture) was investigated through a collaborative Towers of Hanoi challenge accomplished by a hybrid human robot team [19]. Nonverbal cues for timing coordination in physical turn-taking were studied in manufacturing [20]. Such research focuses on turn-taking process modelling and the robot control, while neglecting predictions of the turn-taking timing and intention.

III. COLLABORATIVE TASK AND TURN-EVENTS DEFINITION

This section introduces the definitions of collaborative task and associated turn-events to formulate the early turn-taking prediction problem.

Consider a human agent $H$ working with a robot agent $R$ on a collaborative task $W$. $H$ and $R$ conduct and alternate through a sequence of subtasks $w^A$, where subscript $k$ indicates the subtask index and superscript $A$ indicates the agent responsible for this subtask (i.e., $A \in \{H, R\}$). The subtask is defined as $w^A_k \triangleq (g_k, z^H_k, z^R_k)$, where $g_k \in G$ is the action label, $z^H_k$ is the beginning time, and $z^R_k$ is the finishing time of the subtask. $G$ is the set containing all the action labels, such as delivering, retrieving, and exchanging tools. The subtask $w^A_k$ is treated as the “atomic” component in the definition since a turn only happens during the transition of two subtasks.

As the human $H$ conducts subtask $w^H_k$ and time goes on from $z^H_k$ to $z^H_{k+1}$, she gets closer in finishing the subtask and expresses an increased level of intention to give out the turn. Since this paper focuses on enabling robotic assistants to take turns from humans, only the turn transitions from $H$ to $R$ are considered. Each turn transition from $w^H_k$ to $w^R_{k+1}$ defines a turn-event $E_k$, in which the human is showing an intention to give out the turn (denoted as $E^{give}$). On the other hand, for the majority of subtask $w^R_k$, the human is focusing on the current action and shows no intention to relinquish the turn. This period implicitly defines a turn-event $E_k$, in which human intends to keep the turn (denoted as $E^{keep}$). Each turn-event $E_k \in \{E^{give}, E^{keep}\}$ spans a time window $[t^k_g, t^{k+1}_g]$, which are the starting and ending time for this turn-event. The collaborative task $W$, subtask $w^A_k$, and turn-events are illustrated in Figure 3.

![Illustration of the definition of collaborative task $W = \{w^A\}$ and turn-events $E_k \in \{E^{give}, E^{keep}\}$](image-url)
As the human is conducting subtask \( w_{H} \), she is monitored through \( M \) sensor channels which include physiological, physical, and neurological signals. The captured signal at time \( t \) is denoted as \( s(t) \in \mathbb{R}^{1 \times M} \). Given a turn-event \( E_{k} \) which spans time \([t_{k}, t_{k}']\), the captured signals within this time window are stacked together to form a matrix representation \( X_{k} \equiv [s(t_{k}': t_{k}')] \in \mathbb{R}^{L_{k} \times M} \), where \( L_{k} \) is the discrete event length \( (i.e., L_{k} = t_{k}' - t_{k}) \). For each stacked segment \( X_{k} \), a label \( y_{k} \in \{0, 1\} \) is assigned to indicate either turn-giving \((i.e., y_{k} = 1) \) when \( E_{k} \in \mathcal{E}_{\text{giving}} \) or turn-keeping \((i.e., y_{k} = 0) \) when \( E_{k} \in \mathcal{E}_{\text{keeping}} \). The turn-taking prediction algorithm \( \phi(\cdot) \) then returns an estimate \( \hat{y}_{k} = \phi(X_{k}) \in \{0, 1\} \). Moreover, the turn-taking intention should be predicted in an early stage given only partial observations \( (i.e., with only implicit cues before explicit cues even start) \). This way, the human’s intent to relinquish the turn can be recognized early and the robot can start moving early to facilitate the transition. Therefore, given partial observation \( X_{k}^{c} \in \mathbb{R}^{(\tau L_{k}) \times M} \) for only the beginning \( \tau \) fraction \((0 < \tau \leq 1) \), an early decision is made according to \( \hat{y}_{k}^{c} = \phi(X_{k}^{c}) \in \{0, 1\} \). The resultant dataset \( D^{*} = \{(X_{k}^{c}, y_{k}, \hat{y}_{k}) \mid X_{k}^{c} \in \mathbb{R}^{(\tau L_{k}) \times M}, y_{k} \in \{0, 1\}, \hat{y}_{k} = \phi(X_{k}^{c}) \in \{0, 1\}\} \) is then used for the following turn-taking analysis.

IV. COGNITIVE TURN-TAKING MODELS (CTTM)

This section presents the CTTM framework: Section IV-A discusses the network structure, Section IV-B presents the neuron mapping methods, and Section IV-C discusses the various aspects of SNN training.

A. SNN Structure

Conventional neural networks enforce the same conducting delay between consecutive layers, thus all the neurons of the same layer can only fire at the same time. This rigid structure is insufficient when modeling multimodal temporal sequences [20]. SNN, however, can model the variability of axonal conduction delays between neurons. Firings will take different amounts of time to traverse around the network based on the different conduction delays. This way, the asynchronous effect of multimodal signals can be modeled.

The adopted SNN structure and parameters are similar to the work of Izhikevich [21]. The network has 250 neurons, with 200 excitatory neurons \((i.e., can be stimulated)\) and 50 inhibitory neurons \((i.e., cannot be stimulated)\). Each excitatory and inhibitory neuron has 25 post synapses, connecting to 25 other neurons following a uniform distribution. Each synapse has a conduction delay in the range of \([1, 20]\ \text{ms}\) following a uniform distribution. The weights of the synaptic connections are initialized to \(+6\) for all post synapses after excitatory neurons, and \(-5\) for all post synapses after inhibitory neurons. Those weights represent how strong the synaptic connection is between two neurons, and are updated based on the Spike Timing-Dependent Plasticity (STDP) rule [22]. The maximum weight for each synaptic connection is set to be 10.

Each neuron model is depicted by a set of formulas, as given in (1). This neuron model, together with the constant parameters are the same with the simple spiking model proposed in [23]:

\[
\begin{aligned}
\frac{dv}{dt} &= 0.04v^2 + 5v + 140 - u + I \\
\frac{du}{dt} &= a(bv - u)
\end{aligned}
\]

with the auxiliary after-spike resetting:

\[
\text{if } v \geq +30 \text{mV, then } \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases}
\]

The variable \( v \) represents membrane potential of the neuron, and \( u \) represents a membrane recovery variable which provides negative feedback to \( v \). Variable \( I \) is the input DC current to this neuron, which was set to 20 mA when this neuron is stimulated by the input multimodal data. As illustrated in Figure 4, for recovery variable \( u \), \( a \) represents the time scale, \( b \) represents the sensitivity, and \( d \) represents after-spike reset increment. For membrane potential variable \( v \), \( c \) represents the after-spike reset value. The excitatory neurons are configured to exhibit regular spiking (RS) firing patterns with parameters \((a, b, c, d) = (0.02, 0.2, -65, 8)\), and the inhibitory neurons are configured to exhibit low-threshold spiking (LTS) firing patterns with parameters \((a, b, c, d) = (0.02, 0.25, -65, 2)\). Those parameters were chosen since they performed the best out of all excitatory-inhibitory neuron combinations proposed by [23] in our case.

Fig. 4. (left) dynamics of the simple spiking model. (middle) voltage response of regular spiking (RS) neuron to a step DC-current \( I=10 \text{mA} \). (right) voltage response of low-threshold spiking (LTS) neuron. Time resolution is 0.1ms. Electronic version of the figure and reproduction permissions are freely available at www.izhikevich.com.

B. Neuron Mapping

To model human turn-taking behaviors, it is necessary to map the input multimodal data to the SNN neurons. Previous research has attempted to map the discrete input data into the neurons on a one-to-one basis. For example, one image pixel corresponds to one neuron [24] or one level of orientation maps to 5 randomly chosen neurons [25]. However, mapping multimodal continuous-valued signals into SNN neurons requires a different approach since the network dimension grows exponentially to encode all possible combinations of continuous inputs. A small example of 5 discrete levels and 10 multimodal channels would lead to \(10^8\) neurons, which is intractable. This problem is solved by resorting to automatic channel quantization and decision-level fusion methods.

First, each of the \(M\) channels was quantized into \(V\) discrete levels. The 1-percentile threshold \((r_1)\) and the 99-percentile threshold \((r_{99})\) are used as fences to remove outliers. The \(V\) bins were divided within the \(1\% \sim 99\%\) range to
encode the continuous sensor signals. Given a sensor reading value \( s \), it was quantized to level \( q \) \((q = 0, 1, \ldots, V - 1)\), following \( q = \text{int}(\frac{s - c_{i-1}}{v_{i+1}} \times V) \). The quantization process is applied to each of the \( M \) channels of \( X_k \in \mathbb{R}^{L_k \times M} \). The result is \( \tilde{X}_k \in \mathbb{Q}_V^{L_k \times M} \), where \( \mathbb{Q}_V \) represents the quantized space with \( V \) levels. Now the partial observation is \( \tilde{X}_k \in \mathbb{Q}_V^{L_k \times M} \). For each quantized level \( q \), 5 excitatory neurons in the SNN were randomly allocated following a uniform distribution to represent it. When level \( q \) is active, all its 5 corresponding neurons are stimulated one by one at 1 ms intervals. The DC current of 20 mA provides the stimulation to variable \( I \) in (1). The value of \( V \) was set to be 40, to reach a total of \( 40 \times 5 = 200 \) excitatory neurons so that each excitatory neuron is mapped to a discrete level.

To encode the multimodal inputs, one SNN is constructed for each of the \( M \) channels. The outcomes from each SNN are combined in the end using a decision-level fusion. This approach follows the human brain mechanism (i.e., vision and hearing are independently processed and are fused in later cognitive decision-making stage).

C. SNN Training

Once the mappings between input data \( \tilde{X}_k \) and SNN neurons are established, the network needs to be trained.

The SNN training includes three stages, where the first stage trains the SNN network weights by feeding unsupervised training data repeatedly into the network. Each training observation activates corresponding neurons in sequence, and the network weights are updated following the STDP rule.

The second stage consists of soliciting network responses for different turn-taking classes, and the third stage focuses on constructing features from SNN responses.

1) SNN Synapse Weight Training: This training phase aims to generate the stereotypical network responses for different types of inputs \( \tilde{X}_k \). As mentioned earlier, one SNN network was constructed for each channel of information (i.e., one column of \( \tilde{X}_k \in \mathbb{Q}_V^{L_k \times M} \), denoted as \( \tilde{X}_{ki} \) for column \( i \)). Therefore, there are \( M \) SNN constructed in total, forming a SNN group and is denoted as \( S = \{S_i\}, i = 1, \ldots, M \). Given input \( \tilde{X}_{ki} \), its response to this SNN group is denoted as \( \tilde{G}_k = S(\tilde{X}_{ki}) \). \( \tilde{G}_k \) consists of \( M \) individual responses \( \tilde{G}_{ki} \) for each of the SNN networks, i.e., \( \tilde{G}_k \triangleq \{\tilde{G}_{ki}\}, i = 1, \ldots, M \) where responses \( \tilde{G}_{ki} \triangleq S_i(\tilde{X}_{ki}) \). \( \tilde{G}_{ki} \) denotes the firing maps (which neurons fired at what time) when input \( \tilde{X}_{ki} \) is presented to the model \( S_i \). As mentioned earlier, given an input \( \tilde{X}_{ki} \), the network was simulated for \( T \) milliseconds and each millisecond forms an atomic time unit. There are in total \( N \) neurons in the network which can be potentially fired. Therefore, \( \tilde{G}_{ki} \) is formed as a \( N \) by \( T \) Boolean matrix (i.e., \( \tilde{G}_{ki} \in \{0, 1\}^{N \times T} \)), where a value of 1 at cell \( (n, t) \) indicates that neuron \( n \) \((1 \leq n \leq N)\) fired at time \( t \) \((1 \leq t \leq T)\), i.e.:

\[
\tilde{G}_{ki}(n, t) = \begin{cases} 
1 & \text{neuron n fired at time t} \\
0 & \text{neuron n did not fire at time t} 
\end{cases} \tag{3}
\]

The firing map \( \tilde{G}_k = \{\tilde{G}_{ki}\} \) forms a compact and rich representation of the original signal \( \tilde{X}_k \), and is used to predict the turn-taking type \( (\hat{y}_k = \{0, 1\}) \). When given partial observation \( \tilde{X}_k^* \), its discretized version \( \tilde{X}_k^* \) is fed into the SNN group \( S \), generating a partial response \( \tilde{G}_k^* = S(\tilde{X}_k^*) \). This response is used to predict its turn-taking type \( \hat{y}_k \).

2) Signature Firing Patterns Training: This training phase focuses on creating descriptive features for the SNN network response \( \tilde{G}_{ki} \). Although \( \tilde{G}_{ki} \) can be directly used for classification purposes, it would cost unnecessary computational time and memory usage, as \( \tilde{G}_{ki} \) is a large sparse matrix with many zeros. To solve that problem, we proposed the Normalized Histogram of Neuron Firings (NHNF) descriptor to compactly represent \( \tilde{G}_{ki} \). Specifically, the total number of neurons (i.e. \( N \)) are evenly divided into \( B \) bins, where bin \( b \) \((b = 0, \ldots, B - 1)\) covers neurons indexes in the range \([bN/B, (b+1)N/B)\]. During a simulation of \( T \) ms, the total number of firings for neurons within bin \( b \) is counted and then divided by the simulation duration \( T \) to generate the descriptor \( h_{bi} \) for sample \( X_k \) and feature \( i \):

\[
(h_{bi}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n=bN/B}^{(b+1)N/B} \tilde{G}_{ki}(n, t) \tag{4}
\]

for \( i = 1, \ldots, M; k = 1, \ldots, K; b = 0, \ldots, B - 1 \). Dividing the histogram by the simulation duration \( T \) can make this descriptor duration-invariant and thus can work robustly for variable signal lengths. An illustration of the NHNF descriptor for a sample firing output is given in Figure 5. In the figure, the red dotted lines indicate the boundary of bins, and the green bars represent the histogram values.

Since there are \( M \) channels of information in total, \( M \) sets of histograms \( h_{bi} \) are generated for a given sample \( X_k \). The histograms \( h_{bi} \) for each bin \( (b) \) and each channel \((i)\)
are concatenated together to form the final feature descriptor for input $X_k$, denoted as $H_k$. Then $H_k$ is used to predict the turn-event type $\hat{y}_k$. When only partial responses $G_{\tau}$ are available, the NHNF descriptors were extracted from it (denoted as $H_{\tau}$) and used to predict the turn-event type $\hat{y}_{\tau}$.

V. EXPERIMENTS

The proposed CTTM framework was tested in a robotic scrub nurse scenario, where the surgeon’s turn-taking intentions must be predicted ahead of time. This section discusses the relevant aspects in the experiment setup, including surgical task setup (Section V-A), human sensing and signal processing (Section V-B), the performance (Section V-C), and finally the network visualization (Section V-D).

A. Surgical Task Setup

A simulation platform for surgical operations was used to capture turn-taking cues of surgeons, as shown in Figure 1. The platform consists of a patient simulator and a set of surgical instruments to conduct a mock abdominal incision and closure task [27]. In this collaborative task, the surgeon and nurse collaborate by exchanging surgical instruments. The surgeon performs operations while the nurse searches, prepares, and delivers the requested surgical instrument.

Participants were recruited to perform a mock surgical task. Twelve participants were recruited from a large academic institution, with the age range of 20 to 31 years ($M=25.7, SD=2.93$). After inform consent was given (IRB protocol 1305013664), participants completed a training session. Surgical instruments were introduced by repeated recitation of their names, and a training video of step-by-step instructions of the mock abdominal incision and closure task (10 mins) was given. After the video, the participant attempted a “warm-up” trial. Each participant repeated the surgical task 5 times to reach performance proficiency. Although a surgeon population was not used, the training sessions and repeated trials led to high face-fidelity data. This dataset is realistic enough to validate the early turn-taking prediction capability.

Each execution of the surgical task included in average 14 surgical instrument requests. The surgical request actions were annotated as turn-giving events ($E_{\text{give}}$), and the surgical operation actions were annotated as turn-keeping events ($E_{\text{keep}}$). Two annotators were presented with the recorded videos of the surgical tasks. The annotators independently determined the starting and ending time for each turn-event as well as its type. The main annotator labeled the entire dataset while the second annotator labeled a random 10% of selected segments. An inter-rater reliability of Cohen’s $\kappa = 0.95$ [28] was found, indicating high agreement between two sets of annotations. Overall, 846 turn-giving events ($y_k = 1$) and 1305 turn-keeping events ($y_k = 0$) are generated for the following turn-taking analysis.

B. Multimodal Human Sensing and Signal Processing

The participant in the surgeon role had her communication cues collected during the surgical operation for turn-taking analysis. Three sensors were used to capture the multimodal communication cues: the Myo armband, Epoc headset, and Kinect sensor. Each sensor captured multiple channels of information from the human, as illustrated in Figure 6. There were in total $M(M = 50)$ channels of raw signals, which were synchronized at a frequency of 20 Hz.

Preprocessing techniques were used to smooth and normalize the raw multimodal signals. Each of the $M$ channels of information was smoothed with Exponentially Weighted Moving Average approach with an empirical weight of 0.2 [29]. Then, each of the $M$ channels was normalized to zero mean and unit variance.

The feature construction and selection algorithm proposed by [30] was adopted. Each channel of the normalized signals was first convolved with a filter bank containing 6 filters, i.e., identity transformation, Sobel operator, Canny edge detector, Laplacian of Gaussian detector, and two Gabor filters. Then the correlation between each encoding with the turn-event labels was calculated through $\chi^2$ test of independence. The $m$ features of the largest test statistic values were retained as the final feature set to form the initial representation $X_k$.

In this experiment, the value of $m$ was set to 10 empirically. The selected top 10 features are shown in TABLE I.
C. CTTM Performance

To evaluate the performance of CTTM in predicting surgeons’ turn-taking intentions, computational experiments were conducted. The experiment setup followed the leave-one-subject-out (loso) cross validation. In each fold, the data from 11 subjects was used for training and data from the last subject was separated for testing. Such evaluation scheme can evaluate the algorithm’s generalization capability on unseen subjects. For accuracy measurement between prediction result \( y_k \in \{0, 1\} \) and ground truth \( y_k \in \{0, 1\} \), the \( F_1 \) score for turn-giving class was calculated (i.e., harmonic mean of precision and recall).

The CTTM can recognize the type of the turn-event given only partial observation \( X^t_k \in \mathbb{R}^{(\tau_u+k) \times M} \). An early decision was then made according to \( \hat{y}^t_k = \phi(X^t_k) \in \{0, 1\} \). To evaluate the algorithm’s early prediction performance, the \( F_1(\tau) \) value for \( \tau = 0.1, 0.2, \ldots, 1.0 \) was calculated and presented. The NHNF descriptor \( H^t_k \) was extracted from the beginning \( \tau \) fraction of input (i.e., \( X^t_k \)), with 50 bins for the histogram (i.e., \( B = 50 \)). Then the descriptor \( H^t_k \) was normalized so that each dimension had zero mean and unit variance. The normalized feature was then fed into a SVM classifier, which gave the prediction of turn-taking event type \( \hat{y}^t_k \). The hyper-parameters for the SVM was set based on a 5-fold within-group grid search process.

Four benchmark algorithms were implemented to compare and evaluate the proposed framework, as explained below.

The first benchmark, Dynamic Time Warping (DTW), is a traditional time-series modelling algorithm. It has been successfully applied to speech recognition [31] and early gesture recognition [32] but never to turn-taking tasks. The multidimensional DTW proposed in [33] was used with 1-norm distance measurement for two multi-dimensional signals. The k-nearest-neighbor classification scheme was applied (with 20 templates for each class). The DTW distance between the two feature vectors was calculated (i.e., \( X_u \) and \( X_k \)) and the label of the closer sample was used as the predicted label.

The second benchmark (Ishii) is a turn-taking prediction algorithm applied to human conservation [34]. Even though this framework targets a different turn-taking application, it can still be adopted into this case with minor modifications, as described below. In Ishii’s framework, each signal channel from \( X_k \) was normalized into the range of \([\mu - \sigma, \mu + \sigma]\) and then three descriptive features were extracted (i.e., average movement, average amplitude, and average frequency of movement) for each channel of signal. The SVM algorithm was then used for classification, with hyper-parameters selected based on 5-fold grid search.

The third benchmark (SNN-PNG) is a SNN-type algorithm [35] and is the closet algorithm to the proposed approach. The major difference is that the SNN-PNG framework is a template-based technique and uses PNG as features. In [35], PNG extracted from \( G_{ki} \) were used as features, and the Jaccard similarity and Longest Common Sequence (LCS) algorithm are used to measure the distance between the unknown pattern and the training templates. A nearest-neighbor approach was used for classification purposes, with 20 templates for each class.

The fourth benchmark (Human) reflects human performance. We used a “button-press” paradigm [36], where recorded videos of the surgical operation were played back to participants and paused at random times. At every pause, the participant was asked what he thinks that the surgeon wants to do (keep or relinquish the turn). The participants in this experiment used a cross-participant setting for data annotation (no self-annotation).

The performances of the proposed CTTM framework with the four benchmarks are shown in Figure 7. The CTTM framework greatly outperforms all the computational benchmark methods. Additionally, the CTTM performance surpasses human performance when little action is given (i.e., when \( \tau < 0.4 \)). After providing more observations, the human performance is better than CTTM, with an average \( F_1 \) score margin of 0.05. This prediction behavior can allow inference of human’s turn-taking intentions in an early stage, since only a few anchor neurons are required to fire to generate a sequence of signals that traverse through the network, forming a stereotypical response [24]. Similar observations have been reported in hand digit recognition tasks [24], [35] and gesture recognition [13] with SNN.

![Fig. 7. Performance of proposed CTTM framework compared to the benchmark algorithms.](Image)
D. Visualization of Fired Neurons

Visualizing the SNN responses allows better understanding of the patterns learned by the model. Figure 8 shows 6 neurons firing maps (i.e., $G_{ki}$) for each class of input. The SNN model for the first feature was selected here for visualization. The responses to turn-keeping inputs ($E_{keep}$) are on the top two rows, and the responses to the turn-giving inputs ($E_{give}$) are at the bottom two rows. Visual inspection reveals that the SNN responds differently to $E_{keep}$ and $E_{give}$ inputs. The $E_{give}$ inputs in general can fire more neurons in the trained SNN network compared to $E_{keep}$ inputs. This could mean that humans exhibit a coherent pattern when relinquishing their turn. The neurons in the CTTM framework fire in the presence of such pattern. On the other hand, the intention to keep the current turn (i.e., focusing on operation) can be diverse (since the operations can be very diverse) and cannot trigger enough firing. Additionally, responses in $E_{give}$ generally have a column-wise structure (either one column or two columns). This structure is generated when a group of neurons fire together in a time-locked pattern, forming a PNG as a signature of early turn-taking intent.

Fig. 8. Comparison of SNN responses to different classes of input. The SNN firing map is shown for the turn-keeping inputs (top two rows) and turn-giving inputs (bottom two rows). Each simulation lasts 250ms (x-axis) and there are 250 neurons (y-axis). Blue dot indicates neuron firing.

VI. CONCLUSIONS

In human robot interaction, turn-taking capability is critical to enable robots to interact with humans seamlessly, naturally, and efficiently. However, current turn-taking algorithms cannot help accomplish early prediction. To bridge this gap, this paper proposes the CTTM, which leverages cognitive models for early turn-taking prediction. More specifically, this model is capable of reasoning human turn-taking intentions, based on the neurons firing patterns in the SNN. The CTTM framework can reason about the multimodal human communication cues (both implicit and explicit) and predict a person’s intention of keeping or relinquishing the turn in an early stage. Such prediction can then be used to control robot actions.

The proposed CTTM framework was implemented in a surgical context, where a robotic scrub nurse predicted the surgeon’s turn-taking intentions to determine when to deliver surgical instruments. The algorithm’s turn-taking prediction performance is evaluated based on a dataset, acquired through a simulated surgical procedure. The proposed CTTM framework outperformed computational state-of-the-art algorithms and can surpass human performance when only partial observation is available (i.e., less than 40% of full action). Specifically, the proposed framework achieves a $F_1$ score of 0.68 when only 10% of full action is presented and a $F_1$ score of 0.87 at 50% presentation. Such early prediction capability is partially due to the implemented cognitive models (i.e., SNN) for early prediction. Such behavior would enable co-robots to work in a hybrid environment side by side with humans.

Future work includes 1) including more contextual information to improve early prediction capability (e.g., current task progress); 2) validating the framework in real surgeries; 3) transfer the CTTM framework to other scenarios, such as robot-assisted manufacturing.

References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, pp. 696–735, 1974.
[2] K. Inkpen, J. McGrenere, K. S. Booth, and M. Klawe, “Turn-taking protocols for mouse-driven collaborative environments,” in *Proceedings of Graphics Interface*, vol. 97, pp. 138–145, 1997.
[3] A. Chan, K. MacLean, and J. McGrenere, “Designing haptic icons to support collaborative turn-taking,” *International Journal of Human–Computer Studies*, vol. 66, pp. 333–355, May 2008.
[4] N. Sebanz, H. Bekkering, and G. Knoblich, “Joint action: bodies and minds moving together,” *Trends in cognitive sciences*, vol. 10, no. 2, pp. 70–76, 2006.
[5] K. L. Marsh, M. J. Richardson, and R. C. Schmidt, “Social connection through joint action and interpersonal coordination,” *Topics in Cognitive Science*, vol. 1, no. 2, pp. 320–339, 2009.
[6] C. Chao and A. Thomaz, “Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration,” *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1330–1353, 2016.
[7] I. De Kok and D. Heylen, “Multimodal end-of-turn prediction in multiparty meetings,” in *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 91–98, ACM, 2009.
[8] E. Calisgan, A. Haddadi, H. Van der Loos, J. Alcazar, and E. Croft, “Identifying nonverbal cues for automated human-robot turn-taking,” in *2012 IEEE RO-MAN*, pp. 418–423, Sept. 2012.
[9] H. Arisikere, E. Shriberg, and U. Ozertem, “Enhanced End-of-Turn Detection for Speech to a Personal Assistant,” in *2015 AAAI Spring Symposium Series*, Mar. 2015.
[10] S. Loiselle, J. Rouat, D. Pressnitzer, and S. Thorpe, “Exploration of rank order coding with spiking neural networks for speech recognition,” in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, 2005., vol. 4, pp. 2076–2080, IEEE, 2005.
[11] P. Heeman and R. Lünsford, “Can Overhearers Predict Who Will Speak Next?,” in *2015 AAAI Spring Symposium Series*, Mar. 2015.
[12] W. Maass, “Networks of spiking neurons: the third generation of neural network models,” *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[13] J. Botzheim, T. Obo, and N. Kubota, “Human gesture recognition for robot partners by spiking neural network and classification learning,” in *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS)*, 2012 Joint 6th International Conference on, pp. 1954–1958, IEEE, 2012.

[14] S. Ghosh-Dastidar and H. Adeli, “Improved spiking neural networks for EEG classification and epilepsy and seizure detection,” *Integrated Computer-Aided Engineering*, vol. 14, no. 3, pp. 187–212, 2007.

[15] D. Schlangen, “From reaction to prediction: Experiments with computational models of turn-taking,” in *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.

[16] N. G. Ward, O. Fuentes, and A. Vega, “Dialog prediction for a general model of turn-taking,” in *INTERSPEECH*, pp. 2662–2665, Citeseer, 2010.

[17] J.-P. d. Ruiter, H. Mitterer, and N. J. Enfield, “Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation,” *Language*, vol. 82, pp. 515–535, Sept. 2006.

[18] E. Padilha and J. Carletta, “Nonverbal behaviours improving a simulation of small group discussion,” in *Proceedings of the First International Nordic Symposium of Multi-modal Communication*, pp. 93–105, Citeseer, 2003.

[19] C. Chao and A. Thomaz, “Timed petri nets for multimodal interaction modeling,” in *ICMI 2012 Workshop on Speech and Gesture Production in Virtually and Physically Embodied Conversational Agents*, 2012.

[20] J. W. Hart, B. Gleeson, M. Pan, A. Moon, K. MacLean, and E. Croft, “Gesture, Gaze, Touch, and Hesitation: Timing Cues for Collaborative Work,” Feb. 2015.

[21] E. M. Izhikevich, “Polychronization: computation with spikes,” *Neural computation*, vol. 18, no. 2, pp. 245–282, 2006.

[22] J. Sjstrm and W. Gerstner, “Spike-timing dependent plasticity,” *Spike-timing dependent plasticity*, vol. 35, 2010.

[23] E. M. Izhikevich, “Which model to use for cortical spiking neurons?,” *IEEE transactions on neural networks*, vol. 15, no. 5, pp. 1063–1070, 2004.

[24] B. Rekabdar, M. Nicolescu, M. Nicolescu, and R. Kelley, “Scale and translation invariant learning of spatio-temporal patterns using longest common subsequences and spiking neural networks,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2015.

[25] M. Beyeler, N. D. Dutt, and J. L. Krichmar, “Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule,” *Neural Networks*, vol. 48, pp. 109–124, 2013.

[26] E. M. Izhikevich, “Simple model of spiking neurons.” *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[27] S. N. Martyak and L. E. Curtis, “Abdominal incision and closure,” *The American Journal of Surgery*, vol. 131, pp. 476–480, Apr. 1976.

[28] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 01, 1960.

[29] J. M. Lucas and M. S. Saccucci, “Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements,” vol. 32, pp. 1–12, Feb. 1990.

[30] L.-P. Morency, I. de Kok, and J. Gratch, “Context-based recognition during human interactions: automatic feature selection and encoding dictionary,” in *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 181–188, ACM, 2008.

[31] W. H. Abdulla, D. Chow, and G. Sin, “Cross-words reference template for DTW-based speech recognition systems,” in *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, vol. 4, pp. 1576–1579, IEEE, 2003.

[32] A. Mori, S. Uchida, R. Kurazume, R.-i. Taniguchi, T. Hasegawa, and H. Sakoe, “Early recognition and prediction of gestures,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3, pp. 560–563, IEEE, 2006.

[33] G. A. ten Holt, M. J. Reinders, and E. A. Hendriks, “Multi-dimensional dynamic time warping for gesture recognition,” in *Thirteenth annual conference of the Advanced School for Computing and Imaging*, vol. 300, 2007.

[34] R. Ishii, S. Kuman, and K. Otsuka, “Predicting next speaker based on head movement in multi-party meetings,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 2319–2323, IEEE, 2015.

[35] B. Rekabdar, M. Nicolescu, M. Nicolescu, and S. Louis, “Using patterns of firing neurons in spiking neural networks for learning and early recognition of spatio-temporal patterns,” *Neural Computing and Applications*, vol. 28, no. 5, pp. 881–897, 2017.

[36] L. Magyari, M. C. Bastiaansen, J. P. de Ruiter, and S. C. Levinson, “Early anticipation lies behind the speed of response in conversation,” *Journal of Cognitive Neuroscience*, 2014.