A Multimodal End-of-Turn Prediction Model: Learning from Parasocial Consensus Sampling (Extended Abstract)

Lixing Huang
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA, 90094
lhuang@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA, 90094
morency@ict.usc.edu

Jonathan Gratch
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA, 90094
gratch@ict.usc.edu

ABSTRACT

Virtual human, with realistic behaviors and 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 to give them human-like conversational skills, such as turn-taking skill. In this paper, we propose a multimodal end-of-turn prediction model. Instead of recording face-to-face conversation data, we collect the turn-taking data using Parasocial Consensus Sampling (PCS) framework. Then we analyze the relationship between verbal and nonverbal features and turn-taking behaviors based on 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, which enables virtual humans to make real-time turn-taking predictions. The result shows that our model achieves a higher accuracy than previous methods did.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Measurement, Design, Experimentation

Keywords

Virtual Human, Multimodal, Turn-taking, Parasocial Consensus

1. INTRODUCTION

Human conversation is a cooperative and fluent activity. People rarely speak simultaneously. Rather, the roles of speaker and listener are regulated seamlessly by a negotiation process of turn-taking. Considerable research is directed at understanding this mechanism and integrating it into virtual humans. The fluidity of natural conversation presents a considerable challenge for building virtual humans. On one hand, communication is multimodal: information is manifest in different channels and these channels may unfold under different time scales; on the other hand, effective communication involves forecasting what one's conversational partner will do in the future. Sacks et al. [1] argued that the smooth exchange of turns in conversation 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.

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. Second, we demonstrate the effectiveness of a novel methodology, Parasocial Consensus Sampling (PCS), for learning such models. PCS [2] was recently proposed and applied to the problem of predicting listener backchannel feedback successfully. Here we reinforce the viability of this general framework by demonstrating its effectiveness on the novel domain of end-of-turn prediction. The experiment shows that our model trained on the data from Parasocial Consensus Sampling achieves a higher accuracy than previous methods.

2. Parasocial Consensus Sampling

Traditionally, virtual humans learn from annotated recordings of face-to-face interactions. However, as suggested in [2], 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. [2] proposed the Parasocial Consensus Sampling 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 this way, multiple independent participants are able to experience the same social situation and provide parasocial responses to the same event.

Cite as: A Multimodal End-of-Turn Prediction Model: Learning from Parasocial Consensus Sampling, Lixing Huang, Louis-Philippe Morency and Jonathan Gratch (Extended Abstract), *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1289 - 1290.

		Turn-taking Pauses (422)	Non-turn-taking Pauses (1012)
Looking-away		3% (11)	59%(598)
Looking-towards		27%(114)	12%(123)
Nods		17%(71)	8%(77)
Pitch Slope	**Up**	38%(160)	22%(227)
	Down	35%(149)	24%(238)
	Straight	27%(113)	54%(547)
Average Pitch Value	**Above**	14%(60)	11%(108)
	Below	38%(162)	32%(321)
	At	48%(200)	57%(583)
Syntax Completion		98%(416)	64%(648)

Table 1. The percentage of turn-taking pauses and non-turn-taking pauses that co-occur with different features. The absolute number is shown in parentheses.

Later, these responses can be aggregated to form the 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. For details of PCS, please refer to [2].

3. Analysis of Multimodal Patterns

As described in the literature, gaze, nods, prosody and syntactic features are all argued to impact turn-taking behavior. Before attempting to learn the prediction model, we first explore the relative impact of these different turn-taking cues, which are shown in Table 1. We find the occurrences of looking-away, looking-towards and head nods are very informative cues; prosodic features (pitch slope) provide useful information as well. 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

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) [3] is used because of its advantages in modeling the sequential aspects of human behavior. In the training process, features are first encoded using encoding dictionaries [4] to capture the asynchrony. While testing, the model takes as input a sequence of encoded features and output a sequence of probabilities of states (taking turn or not), from which we can induce the predicted turn-taking time.

4.1 Results and Discussion

We compare the performance of two models learned from PCS with two baseline models. **PCS-Multimodal**: This is the model we learned previously; **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; **Prosody Model**: Prosody model is trained the same way as PCS-Multimodal model but with only prosodic features [5]; **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 predicted time is considered correct if happening during the turn-taking pause.

	Precision	Recall	F1
PCS-Multimodal Model	0.78	0.81	0.80
PCS-Pause Model	0.59	0.90	0.71
Prosody Model [5]	0.58	0.77	0.67
Syntax Model [1]	0.29	0.97	0.45

Table 2. Evaluations for Turn-taking pause prediction: F1 score of PCS-Multimodal is significantly better than that of the other three models.

As Table 2 shows, F_1 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. By leveraging the multimodal features, our PCS-Multimodal model performs the best. In this paper, Parasocial Consensus Sampling (PCS) framework is applied in collecting and modeling turn-taking behavior, we validate this new methodology further and generalize it to turn-taking behavior modeling.

5. ACKNOWLEDGEMENTS

This study was partially funded by the German Academic Exchange Service and by the U.S. Army Research, Development, and Engineering Command and the National Science Foundation under grant # IIS-0916858. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

6. REFERENCES

[1] Sacks, H., Schegloff, E., Jefferson, G. 1974. A Simplest Systematics for the Organization of Turn-taking for Conversation. Language, vol. 50, pp. 735-996.

[2] Huang, L., Morency, L.-P., Gratch, J. 2010. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In Proceedings of 9th International Conference on Autonomous Agent and Multiagent Systems (Toronto, 2010)

[3] Lafferty, J., McCallum, A., Pereira, F. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of 18th International Conference on Machine Learning, 2001.

[4] Morency, L.-P., de Kok, I., Gratch, J. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In Proceedings of the 8th International Conference on Intelligent Virtual Agents (Tokyo, 2008)

[5] Jonsdottir, G.R., Thorisson, K.R., Nivel, E. 2008. Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In Proceedings of International Conference on Intelligent Virtual Agent (Tokyo, 2008)