Estimating Rapport in Conversations: An Interpretable and Dyadic Multi-Modal Approach

Gustav GRIMBERG [a,1], Thomas JANSSOONE [b], Chloé CLAVEL [c] and Justine CASSELL [b]

[a] *École Normale Supérieure de Paris, PSL*
[b] *Carnegie Mellon University & Inria, Paris*
[c] *Institut Polytechnique de Paris, Telecom Paris, LTCI*

Abstract. The concept of rapport has attracted increasing attention in the human-agent interaction community. In this paper, we propose an interpretable rapport estimator based on features from the social sciences. We discuss how this can provide insight into how different conversational features impact rapport and ultimately inform behavior generation in virtual agents.

Keywords. Social Signal Processing, Interpretable Machine Learning

1. Introduction

The field of human-computer interaction (HCI) is increasingly leaving behind the metaphor of computer as tool, and adopting the metaphor of computer as conversational partner. In this context, there is increased focus on incorporating into virtual agents human-human conversational phenomena well-studied in the social sciences, such as grounding, turn-taking, conversational strategies, emotion, and even rapport, the feeling of connection or harmony between participants in a conversation [1]. Here we train a rapport estimator relying on insights from the social sciences and demonstrate how this, in conjunction with state-of-art methods for interpreting machine learning models, can provide insights into how different conversational features impact rapport and, ultimately, how to build these features into human-agent conversation.

[1] Corresponding Author. E-mail: gustav.grimberg@ens.psl.eu.

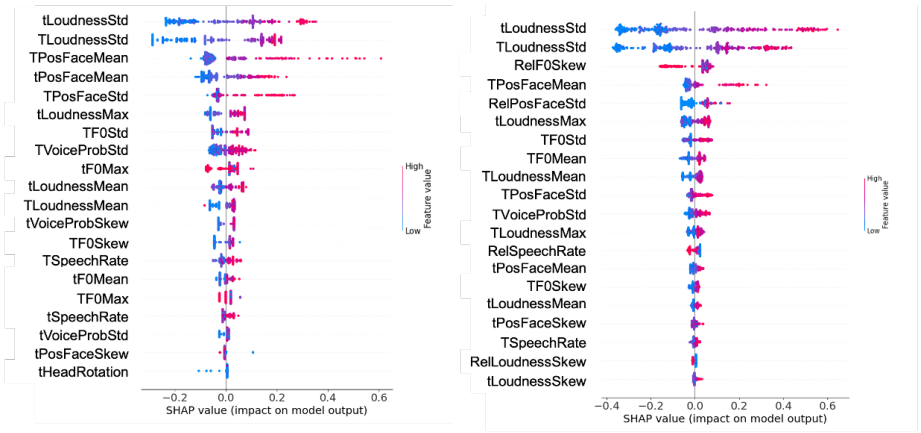
Table 1. Overview of features used for rapport estimations

Modality	Feature type	Feature name	Description	Key
Speech activity	Individual	SpeechRate Tutee/Tutor	The speech rate for the tutee or tutor	[t/T]SpeechRate
	Relative	RelSpeechRate	The relative speech rate computing for the tutee and tutor	RelSpeechRate
Prosody	Individual	$F_0[\mu\ \sigma$ skew] Tutee/Tutor	The mean, standard deviation and the skewness of the fundamental frequency over the slice for the tutee or tutor	[t/T]F0[Mean/Std/Skew]
		Loudness [$\mu\ \sigma$ skew] Tutee/Tutor	The mean, standard deviation and the skewness of the loudness over the slice for the tutee or tutor	[t/T]Loudness[Mean/Std/Skew]
		VoiceProb [$\mu\ \sigma$ skew] Tutee/Tutor	The mean, standard deviation and the skewness of the voicing probability over the slice for the tutee or tutor	[t/T]VoiceProb[Mean/Std/Skew]
	Relative	Rel$F_0[\mu\ \sigma$ skew]	The relative mean, standard deviation and the skewness of the fundamental frequency over the slice	RelF0[Mean/Std/Skew]
		RelLoudness [$\mu\ \sigma$ skew]	The relative mean, standard deviation and the skewness of the loudness over the slice	RelLoudness[Mean/Std/Skew]
		RelVoiceProb [$\mu\ \sigma$ skew]	The relative mean, standard deviation and the skewness of the fundamental voicing probability over the slice	RelVoiceProb[Mean/Std/Skew]
Visual	Individual	EyeGazeChange Tutee/Tutor	Eye gaze changes for the tutee or tutor	[t/T]GazeChange
		HeadMovement Tutee/Tutor	Change in head movement for the tutee or tutor	[t/T]HeadRotation
		PosiFace[$\mu\ \sigma$ skew] Tutee/Tutor	The mean, standard deviation and skewness of the facial positivity indicator over the slice for the tutee or tutor	[t/T]PosFace[Mean/Std/Skew]
	Relative	RelPosiFace[$\mu\ \sigma$ skew]	The relative mean, standard deviation and the skewness of the positive face indicator over the slice	RelPosFace[Mean/Std/Skew]

2. Machine Learning Models and Results

The machine learning models presented here were trained on data originally collected in order to investigate how rapport emerges in teenagers tutoring one another in algebra [2]. The current analysis relies on data from 14 dyads who met twice over videoconference, and the data was annotated for rapport in 30-seconds slices [2].

From these data, we extracted nonverbal and acoustic features from each individual student (such as a speech rate) as well as features from the dyad (such as relative speech rate), inspired by the social sciences (e.g. [1,3,4]), using OpenFace [5] and openSMILE [6] (see Table 1). We trained a number of machine learning models on these features to estimate rapport for each slice, and selected the best performing models. Then, we deployed SHAP [7], a game-theory-based model-agnostic algorithm that uncovers the role of the features in the models' decisions. Figure 1 shows the performance of two representative models.



(a) Model 1 (Gradient Boosting Regressor), trained on only the individual features. MAE = 1.072.

(b) Model 2 (Random Forrest Regressor), trained on both individual and relative features. MAE = 1.074.

Figure 1. SHAP summary plot of models. The vertical axis is the mean contribution of the feature over the model decision, the horizontal axis is how the distribution of features influences the model decision.

3. Discussion

From Fig 1a, we observe how individual features affect rapport estimations. For example, we see that high values of the positive facial expressions feature, especially for the tutor, affect rapport positively, which could be indicative of the salutary effect of positive feedback on the part of the tutor. In Fig 1b, we see the SHAP diagram for the model trained on both individual and relative features. While we see that the most important features are individual, some relative features rank higher than their corresponding individual features. This is the case for relative changes in positive facial expressions, for instance, where we observe that higher values of relative changes in positive facial expressions - suggesting synchrony between the participants with respect to this feature - impact rapport estimations positively. Being able to interpret the rapport estimator's decisions in this way may allow us to now focus on generating these behaviors in a tutoring embodied conversational agent (ECA). Ultimately, this may improve the ability of virtual agents to build rapport with their human users, and thereby improve both the nature of the conversation and performance on a shared goal [8].

References

[1] Spencer-Oatey H. (Im) Politeness, face and perceptions of rapport: unpackaging their bases and interrelationships. 2005.

[2] Madaio M, Lasko R, Ogan A, Cassell J. Using Temporal Association Rule Mining to Predict Dyadic Rapport in Peer Tutoring. International Educational Data Mining Society. 2017.

[3] Harrigan JA, Oxman TE, Rosenthal R. Rapport expressed through nonverbal behavior. Journal of nonverbal behavior. 1985;9(2):95-110.

[4] Tickle-Degnen L, Rosenthal R. The nature of rapport and its nonverbal correlates. Psychological inquiry. 1990;1(4):285-93.

[5] Baltrusaitis T, Zadeh A, Lim YC, Morency LP. Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE; 2018. p. 59-66.

[6] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia; 2010. p. 1459-62.

[7] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems; 2017. p. 4768-77.

[8] Raphalen Y, Clavel C, Cassell J. "You might think about slightly revising the title": Identifying Hedges in Peer-tutoring Interactions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL); 2022. .