NAVIGATING AND REACHING THERAPEUTIC GOALS WITH DYNAMICAL SYSTEMS IN CONVERSATION-BASED INTERVENTIONS

Victor Ardulov†, Shrikanth Narayanan

Signal Analysis and Interpretation Lab
University of Southern California
ardulov@usc.edu, shri@ee.usc.edu

ABSTRACT
Modern human behavioral signal processing and machine-learning methods have introduced novel ways for representing and estimating internal states of people in goal-based conversational interactions, such as psychotherapy. By combining these methods with systems theoretic approaches, we demonstrate how canonical approaches to control policy design can be utilized for improving the quality of goal-oriented talk-based interactions.

1. INTRODUCTION
Efficiently regulating behavior through conversation is a critical skill in a number of fields, such as law and psychotherapy. In these interactions, often dyadic, each interlocutor assesses the others’ responses and reactions to gauge where the interaction is currently, and attempts to close the gap on where they would like to be. In many regards, similarities can be drawn to negotiations, where the interlocutors goals are in opposition to one another and consequently their conversation and actions are driven towards convincing the other to best their partner [1, 2]. In contrast, the presented work explores methods for a holistic approach to model conversational partners that are trying to achieve an aligned goal. Our examples present use-cases in forensic interviewing of victims and psycho-therapy where successful practitioners are able to help their conversational partners reach desired behavioral states by observing multi-modal signals representing underlying states, and appropriately adjusting their own behavior through language, posture, and speaking patterns. This simultaneous estimation and regulation demonstrates how stripped down to its fundamental components, a conversation might be conceptualized as a state-estimation and navigation problem.

Modern machine learning (ML) methods, particularly deep neural networks (DNNs), have provided a paradigm through which behavioral signals such as language, and speech can be represented in a semantically contextualized embedding. These embeddings, typically a high-dimensional vector space, go hand-in-hand with the state-estimation and navigation perspective and invite a dynamical-system inspired analysis of conversations. As an example, language used by an individual might be considered analogous to an observed signal from a sensor, and the corresponding vector representation from that language can be utilized to represent a component of the speaker’s internal state. Extending this perspective further, the speech signal of interaction participant might be instead considered as an input into the system of their conversational partner, used to navigate the interaction to a desired behavioral condition.

This paper outlines two case-studies: first in child forensic interviewing [3] where conversation-level dynamics are utilized to identify patterns in the way children respond to input from an interviewer. Secondly, in cognitive behavioral therapy [4] where local interactions could be used to identify moments of appropriate application of clinical skills and used to adequately predict conversational outcomes. While there are differences in the extracted signals, both of these works demonstrate how re-imagining of conversational interactions as control-affine dynamical systems provided a holistic framework for the analysis and optimization of social interaction strategies with respect to conversational outcomes and constraints.

2. PRIOR WORK
Behavioral Signal Processing (BSP) is a framework under which signal processing techniques developed for problem spaces such as localization, navigation, and communication, are used to describe observed behaviors and underlying models of behavior expression [5, 6]. This perspective is critical to adequately modelling the multi-modal, linguistic and paralinguistic, elements of conversations. Relevant to the presented work, BSP provides a foundation upon which our methods build control-affine dynamical systems to describe the evolution of an interlocutors’ behavior throughout a conversation with input from a guiding party.

The inherently intertwined nature of affect, emotion, and behavior, enables the BSP framework to be naturally extended
for analysis within the domain of affective computing. Behavioral expression can be motivated by affect, and affect in turn is critical in the internalization and regulation of behavior. These inter-dependencies were critical to the realization of psycho-linguistic norms, which are representations which capture the associated affective content of language. In the work presented EmotiWord dictionary will be references for measuring emotional content in a child and interviewers language use for the analysis of child forensic interviews [7].

Work exploring the utilization of DNNs to predict empathy scores for therapy sessions, was later built atop to describe utterance level therapeutic skills using attention mechanisms [8, 9]. These works support the notion that multi-modal signals can be leveraged to capture semantically meaningful latent representations of conversational states. This work in turn lead to better automated feedback during therapist training allowing for end-to-end pipelines which takes audio information and produces skill codes and assessment ratings in order to improve a clinician therapeutic ability [10]. Thsi work further builds upon these methods by explicitly considering the role of one interlocutor as providing a guiding signal.

3. CASE STUDY 1: CHILD FORENSIC INTERVIEWS

Child Forensic Interviews (CFI) are a two phase approach to elicit reliable and inadmissible testimony from a child during a court proceeding. The protocol is designed to induce narrative responses to open-ended questions, by first presenting innocuous questions and prompts (e.g. “Tell me about your last birthday”), then focusing on concrete details and events pertaining to the crime in question. Interviewers are tasked to gauge the child’s readiness to disclose and must navigate the interaction towards detail-rich, verbally productive, recollections of events when prompted, without leading questions or having the child disengage or become retraumatized. The task is further complicated, as the child’s developmental state and language ability is often atypical, a factor that makes these children further vulnerable.

3.1. Data

A total of 200 forensic interviews are analyzed. The children are separated into 2 age groups: Early Childhood (EC) (4-6 years) and Late Childhood (LC) (10-12 years). Expertly transcribed interviews are used for per-utterance lexical feature extraction, including “agenda” terms, normalized counts of psycho-linguistic norms for valence, arousal, and pleasantness, and verbal productivity. The agenda terms and verbal productivity are computed metrics for measuring information disclosed by the child in accordance to prior work [11].

Extending work from Gentle Align, audio from these interviews are force-aligned over multiple iterations to extract turn level prosody (rhythm and intonation of speech) and intensity (loudness) since they are primary features for understanding emotion and affect from speech [3, 12, 13, 14]. Logpitch and intensity contours are extracted at frame-level using Praat [15], and mean normalized per speaker and per session. Each signal’s median and standard is computed per-utterance resulting in 4 acoustic features.

3.2. Dynamic Mode Decomposition with Control

Having compiled turn-level features into a time-series signals, we are interested in investigating whether a dynamical system describing this data would be descriptive. To achieve this we utilize the methods of Dynamic Mode Decomposition with Control (DMDc) [16].

Taking the perspective that $y_{t+1} = Ay_t + Bx_t$, where $A$ is the transition matrix, describing the autonomous evolution of the child’s observations, and $B$ is the controller indicating the influence that the input signal has on the evolution of the observations. Then we can extend this system to describe entire system by taking $Y_{[t:t+1]} = [y_{t+1}, y_t, \ldots, y_1]$, $Y_{[0:t]} = [y_t, y_{t-1}, \ldots, y_0]$, and $X_{[0:t]} = [x_t, x_{t-1}, \ldots, x_0]$, where $Y$ and $X$ represent the signals produced by the child and interviewer respectively: $Y_{[t+1]} = AY_{[0:t]} + BX_{[0:t]}$

Solving for $A$ and $B$ we re-organize our equation as $X_{[1:t+1]} = [A B] \begin{bmatrix} X_{[0:t]} \\ Y_{[0:t]} \end{bmatrix}$ and find that $[A B] = X_{[1:t+1]} X_{[0:t]}^{-1}$ where $(\cdot)^+$ refers to the Moore-Penrose inverse.

1Due to the sensitive nature and privacy concerns, the data for these studies are not available to the public.
2Our specific implementation is open for further development and input from the community at: https://github.com/nsheth12/canetis
The behavior of the dynamical model is studied via the eigenvalues, \( \lambda \in \mathbb{C} \), which are interpreted as the system’s \( Z \)-transform open-loop poles, representing the dynamics of the system [17]. In brief, given an eigenvalue, its magnitude describes the dampening or exponential behavior of a system as time \( t \to \infty \), thus the larger magnitudes more “dominant” behaviors. Similarly, any imaginary component of an eigenvalue creates an angle, \( \omega \), with relation to the real axis describing the frequency of oscillatory behavior. Evaluating the eigenvalues of the transition matrix, it is possible to describe the child’s behavior as a function of their previous behavior. In this instance, since the controller will not be square eigenvalue decomposition is not possible, however analysis of the singular values is possible.

3.3. Results

Evaluating the single most dominant mode of the DMDc computed transition matrix identifies the dynamics associated with a given child, as well as visualizing common behavioral types across children. Figure 2 shows that there is a trend among many children in the EC age group have eigenvalues that have a larger magnitude and fall on the positive end of the number line. Points closer to the unit circle represent slow decay and dampened oscillation. Meanwhile older children generally lie on the negative side of the real axis implying high frequencies of oscillatory behavior, but generally, also having smaller eigenvalue magnitudes representing faster decay rates.

4. CASE STUDY 2: COGNITIVE THERAPY

Cognitive Behavioral Therapy (CBT) is a therapeutic process during which a therapist helps a client overcome personal, emotional, and behavioral problems via conversation by navigating through a series of exercises. In order to support high-quality care and positive clinical outcomes, researchers and clinicians developed the Cognitive Therapy Rating Scale (CTRS) to assess a therapists’ adherence to core components of CBT [18]. During periods of evaluation and training, the use of CTRS provides therapists with targeted feedback to improve their CBT delivery across 10 sub-scores: agenda, application of technique, collaboration, feedback, guided discovery, homework, interpersonal, key cognition behavior, pacing and timing, strategy for change, and understanding. This feedback improves consistency of clinician competence in the therapy they deliver to clients [19]. To increase the availability of quality care, previous works in automated psychotherapy assessment and feedback have shown success of data-driven evaluation of therapist competence [20, 21, 22, 10]. These works utilize acoustic and language elements of speech towards predicting behaviors and outcomes of the clients and the therapists’ utilization of therapeutic skills. Alternatively, in the present work we will extend the concepts developed for CFI to demonstrate how local dynamics might be used to identify positive therapeutic skills.

4.1. Data

The data consists of 292 CBT session transcripts annotated by trained coders for each CTRS sub-score, which are then accumulated intoa totalCTRS scores. Individual sub-scores are scored on a scale between 0 and 6 where 4 or higher is considered demonstrating adherence, and subsequently a total CTRS score of 40 is the threshold necessary overall for session to be considered as adherent [23]. The scores were converted into binary rating where 0 and 1 represented non-adherent and adherent respectively.

Talk turns of each transcript are converted into embeddings using the DistilBERT model [24] which are considered the inputs and observables of the dynamical system.

4.2. Methods

To identify therapeutic moments that lead to high-adherence within the session, in contrast to §3.2, local dynamics are assessed by compiling turn level vectors and computing the transition and controller eigenvalues over short windows of interaction, \( X_{[t:t+w]} \) and \( Y_{[t:t+w]} \) as demonstrated in Figure 3. These eigenvalues are passed to an ML model to classify whether a window’s dynamics belonged to an adherent session or not. While a number of various models were investigated, Gaussian Naive Bayes (GNB) proved to be the most effective.

4.2.1. Aggregating Global Scores from Local Predictions

Since the belief is that adherent dynamics would occur in non-adherent session but with less frequency, the scores computed by the GNB model are aggregated over the entire session to predict the session level sub-scores. Our analysis tested
Fig. 3. Each talk turn is converted into their respective vector representation via the DistilBERT transformer model, and the used to construct $X$ and $Y$ matrices, which are in turn window conditioned on the window size $w$.

$\text{sum}$ and $\text{average}$ as accumulation methods where $\text{sum}$ simply adds up all of the predicted instances of adherence over each extracted window for a session, while $\text{average}$ divides this value by the number windows in the session. Then a Logistic Regression (LR) models use the training split scores to fit a linear classifier to predict session-level adherence for a given accumulated score.

4.3. Results

Across stratified 5-fold cross-validation, the GNB models performing best on local predictions are compared to statistical bootstrap scores, finding that aggregated scores from most of the models trained with $w=8$ outperform the $2\sigma$ (95-th percentile) baselines but ultimately fall short of the $3\sigma$ (99.7-th percentile). Two of the sub-scores: homework ($f_1 = 0.6191$) and interpersonal ($f_1 = 0.5995$) do not perform better than the $2\sigma$ baselines 0.6256 and 0.6269 respectively. Pacing and timing ($f_1 = 0.6868$) beats the $3\sigma$ baseline ($f_1 = 0.6848$), while agenda ($f_1 = 0.6835$) and total CTRS ($f_1 = 0.6767$) miss with-in error bounds. Within the CTRS coding manual, pacing and timing evaluates the use of time via the induced structure and control the therapist exercises over the session. It is likely the case that setting an agenda at the beginning helps with the pacing and timing, which in turn strongly correlates with the overall CTRS competence. The result highlights the implication that the local dynamics extracted using windowed DMDc capture a meaningful interpretation of the interaction between the therapist and client.

5. DISCUSSION

Overall the results suggest the value in extracting these dynamic modes as they adequately capture the desired behaviors of the therapist in assessing and providing feedback. When considering how these systems might be utilized for real-time feedback Figure 4 shows the resulting models can used to track the relative trajectory of the session as it is happening. With these models it would be possible to evaluate and even plan the future control signals.

6. CONCLUSIONS

In both instances, our systems-theoretic analysis of goal-oriented conversations extracted meaningful representations of interlocutor dynamics. Grounding these interactions in the perspective of a control-affine dynamical system creates a pathway to apply well-established composition and optimization techniques such Multi-Model Adaptive Estimation and Control and Q-Learning [25, 26, 27]. Accounting for various modes of interaction, therapists and interviewers can find optimal interaction policies personalized to each of the client’s or child’s experience.

In CFI, by fitting individualized models to each session our models highlight the differences and similarities across children that are made clear. Prior analyses on simulated forensic interviews show linguistic indicators of readiness to disclose and truthfulness of statements regarding a minor transgression [28, 29]. Applying the present work to these could derive robust and adaptive strategies to elicit relevant information during testimony with less trauma to the child.

For CBT, the local dynamics enable insight into interaction patterns associated with adherent therapeutic strategies. Applying optimizations to therapists’ policies would allow them to strengthen adherent moments and support therapists to make impactful decisions throughout the session. To further unify the deep-learning and systems theoretic perspective, it would be possible to finetune BERT-like conversational models to consider the Koopman operator which in turn can be used to further improve the semantic and predictive information of these embeddings [30].
7. REFERENCES


