Multimodal Interaction Modeling of Child Forensic Interviewing

Victor Ardulov[1], Madelyn Mendlen[1], Manoj Kumar[1],
Neha Anand[1], Shanna Williams[2], Thomas Lyon[2], Shrikanth Narayanan[1]
[1]Signal Analysis and Interpretation Laboratory
[2]Gould School of Law
University of Southern California
ardulov@usc.edu

ABSTRACT

Constructing computational models of interactions during Forensic Interviews (FI) with children presents a unique challenge in being able to maximize complete and accurate information disclosure, while minimizing emotional trauma experienced by the child. Leveraging multiple channels of observational signals, dynamical system modeling is employed to track and identify patterns in the influence interviewers' linguistic and paralinguistic behavior has on children's verbal recall productivity. Specifically, linear mixed effects modeling and dynamical mode decomposition allow for robust analysis of acoustic-prosodic features, aligned with lexical features at turn-level utterances. By varying the window length, the model parameters evaluate both interviewer and child behaviors at different temporal resolutions, thus capturing both rapport-building and disclosure phases of FI. Making use of a recently proposed definition of productivity, the dynamic systems modeling provides insight into the characteristics of interaction that are most relevant to effectively eliciting narrative and task-relevant information from a child.

KEYWORDS

Forensic interviewing, Interaction Modeling, Dynamic Mode Decomposition

ACM Reference Format:
Victor Ardulov, Madelyn Mendlen, Manoj Kumar, Neha Anand, Shanna Williams, Thomas Lyon, Shrikanth Narayanan. 2018. Multimodal Interaction Modeling of Child Forensic Interviewing. In 2018 Int'l. Conf. on Multimodal Interaction (ICMI 2018), Oct. 16-20, 2018, Boulder, CO, USA. ACM, NY, NY, USA, 8 pages. https://doi.org/10.1145/3242969.3243006

1 INTRODUCTION

When a child is a suspected victim or witness to a crime, legal-experts will conduct forensic interviews (FI) in order to elicit testimony from the child. The high-stakes nature and potential risks of the process requires that interviewers are prepared to effectively evoke substantive and accurate details concerning the crime. Furthermore, in an effort to not emotionally impact the child during recall of traumatic events, interviewers must be able to leverage affective markers through various modalities to guide their decisions as they navigate through the semi-structured interaction. Computational models would enable interviewers to identify patterns in the progression of interviews enabling methodological analysis and development of robust strategies and interviewing procedures. This study presents applications of dynamical systems modeling and system identification (SID) as a novel method to analyze the time-evolution of interaction dynamics across various modalities of FI interactions.

FI is a semi-structured interaction in which an interviewer must first establish rapport and then elicit a narrative from the child recanting their possible experience and recollection of a crime. The child's ability to cohesively structure a testimony is often a function of contextual factors such as their age, linguistic development, extent of trauma experienced and cognitive development [9, 21, 38]. In order to overcome the variability found in individuals and develop optimal interview strategies, we suggest implementing computational models which can predict and identify dynamics underlying both temporally within interviews, and across conditions.

Utilizing time-series analysis methods, we further develop recent proposals towards quantification of "verbal productivity" of child utterances within context of the interview "agenda" and the child's responsiveness to inquiries [3]. The current study presents a methodology for taking a child's per-turn productivity and affective lexical features as observations of a dynamical system and interviewer acoustic measurements and lexical features as a control input. By utilizing the Dynamic Mode Decomposition with control (DMDc) framework, we build predictive models of child responses, as well as, study broader behavioral patterns observed in the derived dynamics.

2 BACKGROUND

Children are exceptionally vulnerable to abuse and maltreatment, which has been shown to have long-term impacts and threats to their mental health and well-being. For example, child sexual abuse is known to manifest as various psychological disorders, including PTSD, anxiety, depression and attention problems [10, 12, 15]. Often in these cases, the primary perpetrator is a child's legal guardian or care-giver, placing the child in a dichotomy, where they may be tempted to protect their abuser [20, 32]. The high stakes emphasize the need for substantive and accurate information, as releasing a child to their previous care situation may put them at further risk. Further still, a child's cognitive abilities, language development

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243006

and age contribute to their capacity to completely and consistently recount the incidents in courtroom in the presence of a stranger (attorney) and possibly, the defendant. These factors in conjunction with their susceptibility to coercion and intimidation, make their testimonies incredibly delicate and predisposed to dismissal[23].

In an effort to protect children from coercion and intimidation, forensic interviews are administered outside of the courtroom setting. These interviews are held in isolation from the defendant and often lawyers, with the aim to create conditions where the child feels comfortable disclosing sensitive information pertaining to the case. Interviewers conduct the interaction in two phases: rapport building and disclosure. During rapport building the interviewer must establish a sense of trust with the child, and delineate truth from narrative building. Rapport building is followed by disclosure, during which the interviewer attempts to maximize the amount of substantive information the child will share, while also minimizing the emotional distress and retraumatization.

Considerable efforts have been made in the past decade studying structured protocols towards improving interview quality [19]. These studies have focused largely on the effects of categories of question prompts, deriving anecdotal evidence of effectiveness. As an example, focused-question prompts will produce fewer erroneous responses when compared to open-ended prompts [18]. Similarly, yes/no questions have been generally shown to have a misleading nature [6]. However, coding question types has historically been a manual process that can potentially introduce subjectivity among multiple annotators. In contrast, the current work is free of question and prompt codes, and will investigate more subtle aspects of the interviewer's linguistic and paralinguistic influence on the child behavior during interviews.

Dynamic Mode Decomposition (DMD) has been shown to effectively construct interpretable and robust dynamical models from time-series signals [36]. Developed as a method to numerically model dynamical features of flow data it has been shown to be a discrete approximation of the Koopman operator [34]. Importantly, the eigenvalues of the extracted dynamics matrix can be interpreted as the dominant spectral dynamics. DMD with Control (DMDc) extends the methodology to also incorporate an input control signal [31]. Framing domain relevant behavioral observations of the child as the observations of a system, and those of the interviewer as the input, the DMDc conceptualization models the dyadic interactions of FI. Our work will extend on DMDc, introducing an Online and Windowed DMDc building on the work presented in [41], which will allow us to also track the evolution of the dynamics over time and as more information becomes available.

Speech and language data are rich in the features need to construct affective state estimates. Particularly, affective states are captured in both the vocabulary and prosody of the interlocutor as they engage in conversations [7]. The dynamics of these affective states, and subsequently their expressions, can signal important behavioral indicators such as trust [13]. For example conversational synchrony as expressed in the entrainment of conversational partners has been shown to signal trust in economic exchanges [22, 37]. Following suit previous work [7, 14, 25], this study uses prosodic features of speech and psycho-linguistic norms to capture and model representation of emotion. A more detailed outline of the features used can be found in Section 3.1.

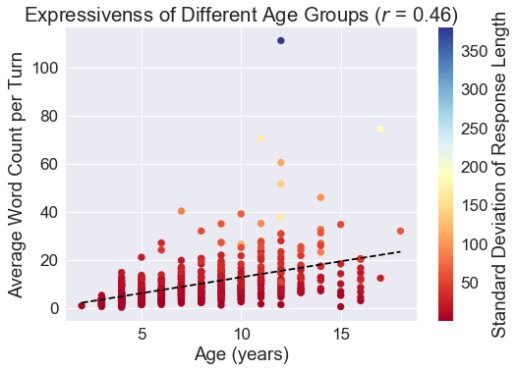

Figure 1: Expressiveness (in terms of average turn-level word count) of children across age groups collected over 527 Forensic Interview Transcripts

Behavioral signal processing (BSP) is a framework for computationally characterizing an individual's psychological state and traits, to produce a quantifiable understanding of their interaction and behavioral dynamics [27]. Operating within this framework, our models fuse extracted acoustic-prosodic (intonation, loudness) and lexical (psycho-linguistic norms) features to construct a multimodal time-series signal representing behavioral states for each utterance. In order to take advantage of this paradigm and gain immediately relevant insights, our models will be utilized to model verbal recall productivity. Most previous FI studies use word count [2, 8] or subjective measures such as the number of informative details [1], or other qualitative measures such as 'richness' [17] as representations of verbal productivity. However, the low variance within each child's per-turn word count (Figure 1) suggests that word count is reflective of an individual's language use and ability, rather than an indicator of expressiveness or narrative structuring. In contrast, Section 3.5 presents recent advances in computational verbal productivity representation which builds on topic modeling techniques allowing for a BSP formulation free of subjective notions of verbal productivity.

3 METHODS

3.1 Data preparation

In this paper, a total of 200 forensic interviews are analyzed. The children are separated into 2 groups according to their age as *Early Childhood (EC)* (4-6 years; μ=4.95, σ=0.73) and *Late Childhood (LC)* (10-12years; μ=10.92, σ=0.82).

The interviews were manually transcribed by trained research assistants at the utterance-level (Utterances/session: μ=176.74, σ=88.74), which are directly used for lexical feature extraction. In order to extract acoustic features audio and text for the entire session are automatically forced-aligned as described below.

Considering the long duration of sessions (μ=39.63 min, σ=13.46 min) and the difficulty in building automatic speech recognition (ASR) systems for child speech, straightforward forced-alignment often fails to provide accurate alignments. An iterative version of

forced-alignment is implemented by building atop of Gentle [1] [2] [26]. The original transcripts are matched with the output of the alignments to identify specific regions of audio associated with each utterance. These regions are then processed for prosodic feature computation. An utterance is considered successfully aligned if 5 consecutive words are successfully aligned at both the beginning and end of utterance.

3.2 Lexical Features

Lexical norms provide dimensional descriptions of affective and other psycholinguistic constructs from spoken language [24, 40]. Classical application of lexical norms were restricted to emotion (arousal, valence and dominance), however diverse affective norms such as age of acquisition, gender-ladenness, concreteness etc. have been applied in behavioral modeling tasks such as therapist empathy prediction during psychotherapy interactions [11] and modeling psychologist language use during autism diagnosis sessions [16]. In this work, norms representing three affective behaviors are selected which are hypothesized to play an important role in the emotional state of the child and thereby influencing verbal productivity. Namely the features examined are:

- *Arousal* represents the degree of physiological excitement associated with a wide range of emotions like anger, fear, enthusiasm, etc.
- *Valence* represents the polarity of emotion (positive:happy, negative:sad).
- *Pleasantness* measures the degree of friendliness or agreeableness, and as a result is highly correlated with the valence of a word.

Durations of strong arousal are expected to be encountered while children are recalling emotionally charged incidents during FI. During these highly aroused recalling valence and pleasantness act as indicators of whether the child is upset or not. We remove pleasantness from the linear mixed effects (LMEs) analyses to prevent multi-collinearity effects. Our choice of pleasantness is motivated by the fact that most discrete emotion can be captured by their relative valence and arousal [35].

The Emotiword tool [24] is used to identify words with emotional content. In the LME models the median value across an utterance is captured, while an *utterance intensity-inverse session intensity* session is computed for the time-series analysis. Stopwords are ignored in both cases. It should be noted that the Emotiword affective values range between -1 and 1, where a lower value represents lower arousal and negative valence.

3.3 Utterance Intensity-Inverse Session Intensity

Extending the concepts of term frequency-inverse document frequency (tf-idf), let us define utterance intensity-inverse session intensity (ui-isi) as a method for computing a logarithmically smoothed count vector for utterance. The motivation is to scale the intensity of each affective dimension (valence, arousal, and pleasantness) relative to the average intensity per utterance and per speaker.

[1] https://github.com/lowerquality/gentle

[2] Our specific implementation is open for further development and input from the community at: https://github.com/nsheth12/canetis

Given an emotion dictionary $\mathcal{E}$ such that

$$\forall w \in \mathcal{E},\ \exists e_w = \{\text{valence}_w, \text{arousal}_w, \ldots\}$$

expresses the mapping of word w onto its psycho-linguistic norms. η is then defined as:

$$\eta(w) = \begin{cases} \langle \text{valence}_w, \text{arousal}_w, \text{pleasantness}_w \rangle & w \in \mathcal{E} \\ \langle 0, 0, 0, \rangle & \text{else} \end{cases}$$

Then given $S = [s_0, s_1, \ldots, s_N]$ representing all of the utterances of only one of the speakers in a specific interview define

$$\nu_S = \frac{1}{m} \sum_{\forall s \in S} \sum_{\forall w \in s} |\eta(w)|$$

where m represents the count number of words for which $\eta(w) \neq \langle 0, 0, 0, \rangle$. Then for a given utterance $s \in S$, the ui-isi is defined as:

$$\xi(s) = \frac{\log(N)}{\log(\nu_S)} \sum_{w \in s} \eta(w)$$

in this case the $\log(\cdot)$ and division is performed element-wise across the vector.

3.4 Prosodic Features

Prosody refers to the rhythm, rate and intonation associated with speech and is concerned with how something spoken, rather than what was spoken. Speech prosody has been hypothesized as primary feature for understanding emotion from speech [29], and prosodic features have been used extensively to study affect from speech [33].

The fundamental frequency (intonation) and intensity (loudness) are used as the features representative of the vocal prosody. Log-pitch and intensity contours are extracted at frame-level using Praat [5], and mean normalized per speaker and per session. The median and standard deviation for each feature is computed across and utterance resulting in 4 prosodic features.

3.5 Verbal Productivity

Verbal Productivity is a measure which acts as a proxy to represent how much a child's utterance reveals relevant information during an interview [2, 30, 39]. To begin we introduce the "agenda" $\mathcal{A}$ which is a vector representation of words (excluding stop words) most frequently referenced by an interviewer. Each of the child's responses is mapped to $\mathbf{r}_t$ using term-frequency on the same dictionary as the agenda. The "agenda" score is computed as

$$g_t = \mathbf{r}_t \cdot \mathcal{A}$$

and is justified by noticing that the words most commonly repeated by the interviewer are also likely to be words that are very relevant to the over-arching topic of the interview. As such, the score captures a child's response being relevant to the topics at play.

By recognizing that the agenda is not always revealed to the child at the beginning of the interview, another mechanism is constructed to capture the child's responsiveness to prompted questions. To accomplish this a "rolling"-agenda vector is constructed for each time step:

$$a_t = \phi_t + \gamma a_{t-1}$$

where ϕ_t refers to the agenda features observed at time-step t and γ is decay term for historical data. The responsiveness is then measured by computing:

$$\rho_t = \mathbf{r}_t \cdot a_t$$

Finally, to capture a balance between the global verbal productivity g_t and the more localized version ρ_t, π_t^* is computed:

$$\pi_t^* = \beta \frac{\rho_t}{||a_t||} + (1-\beta)\frac{g_t}{||\mathcal{A}||}$$

where $\beta \in [0, 1]$ allows us to flexibly leverage the complementary importance of the skills being captured by each sub-metric.

3.6 Linear Mixed Effects Models

To begin, Linear Mixed Effects (LME) Models are utilized to examine overall relationship between the extracted features of both, child and interviewer, and verbal productivity. LME models fit the present problem due to their ability to model individual sessions. In two different model constructions, the fixed effects are set as either the median lexical affective features or the acoustic prosodic features, and the session IDs are set as the random effect.

Given a set of interviews Ψ, define c_t^{ψ} as the features associated with child and u_t^{ψ} as either the lexical or acoustic features observed for the interviewer at turn step t for interview $\psi \in \Psi$. With:

$$\Phi_t^{\psi} = [c_{t-1}^{\psi}\ u_{t-1}^{\psi}]$$

capturing the features of the previous child and interviewer features, the LME formulations then follows as:

$$\hat{\pi}_t^{\psi} = \bar{\pi} + \bar{\pi}^{\psi} + \mathcal{W}\Phi_t + \epsilon_t^{\psi}$$

where $\bar{\pi}$ represent the mean productivity for all sessions and $\bar{\pi}^{\psi}$ is the mean productivity of session ψ. Then $\mathcal{W}$ are the LME coefficients and ϵ is the residual defining a linear approximation which only considers the most recent preceding utterances.

3.7 Dynamical System Models

Following notation presented in Section 3.6, let us define a p-dimensional observation of a child as c_t and interviewer input, of dimension q, as u_t at given time-point $t \in [0, n]$. With this construct the dynamical model:

$$c_{t+1} = Ac_t + Bu_t$$

where A is referred to as the transition matrix, describing the autonomous evolution of the child's observations, and B henceforth referred to as the controller indicates the influence that the input signal has on the evolution of the observations.

When analyzed on per-session basis, the transition matrix and controller capture individual's interaction dynamics, illuminating the nature of how different children respond to input from their interviewers. By utilizing DMDc, the time-series observations of child's productivity, the interviewers prosodic features, and both interlocutor's lexical affective norms can be used to reconstruct the interaction dynamics.

The construction leverages the fact that for $C' = [c_{t+1}, c_t, \ldots, c_1]$, $C = [c_t, c_{t-1}, \ldots, c_0]$, and $\Upsilon = [u_t, u_{t-1}, \ldots, u_0]$:

$$C' = AC + B\Upsilon$$

Since the controller is not known a priori, the problem is reframed as:

$$C' = [A\,B]\begin{bmatrix} C \\ \Upsilon \end{bmatrix} = G\Omega$$

To solve for $G = [A\,B]$, it is required that rank$|\Omega| = p + q$. With this assumption upheld, it follows that:

$$G = C'\Omega^{\dagger}$$

where $(\cdot)^{\dagger}$ refers to the Moore-Penrose inverse.

This analysis assumes that the snapshots for the entire interaction are available. However FI is typically broken up into at least 2 phases - rapport building, and disclosure (Section 2), during which times it is reasonable to expect changes in the dynamics. In order to capture this progression DMDc is extended by the Online and Windowed DMD paradigms presented in [41]. This framework allows us to update our interaction dynamics model as more data becomes available. However, the methods are still contingent on the rank requirement of Ω_t in order to produce the initial fit of the data.

The experiments in dynamic interaction modeling presented here will first examine the derived dynamics of using full DMDc, and evaluate predictive components in the child's state, and the interviewers input. The eigenvalues of the extracted dynamics will be evaluated to explore similarity and identify behavioral patterns with control theoretic conceptualizations. Next, we will experimentally evaluate the ability of online algorithms to capture changes in the dynamics over time, and compare them to a full DMDc as an oracle.

4 RESULTS AND DISCUSSION

4.1 Statistical Analysis

The intensity variations (as measured using the standard deviation of speech intensity within an utterance) of both speakers are found to be statistically correlated with the productivity. Large variation in the child's prosodic intensity is suggestive of productivity. In contrast the intensity variations in the interviewer's questions (*ad_in_st*) are negatively associated with productivity, suggesting that interviewers speaking loudness is consistent prior to a productive response from the child. However this observation varies with age, where children in the LC group do not have a significant correlation ($p \approx 0.12$) with adult intensity. These results support the notion that adult paralinguistic behavior observed in their speech, plays a greater role for younger children. One explanation is that LC children will typically have more developed language ability, and require less indicators from the voice of the interviewer.

Arousal as measured through the psycho-linguistic norm from both child and interviewer is significantly associated with productivity. Some of the emotions that encompass high arousal especially anger and fear have been reported during forensic interviews [4, 28], hence this result shows that entities salient to the alleged crime including names, places and objects (which were shown to feature in top productive responses [3]) tend to be associated with, and possibly evoke strong emotions in the child during FI. The corresponding valence normative from the interviewer questions was negatively associated, although not significant ($p \approx 0.14$). Surprisingly, valence from the child's preceding response was positively associated

Table 1: LME analysis for studying effect of (a) prosodic features, and (b) lexical features on verbal productivity. t-statistics for statistically significant features are reported here. $(p < 0.05)^*$, $(p < 0.01)^{**}$

(a) Prosodic Features

	All (df=895)	*EC* (df=497)	*LC* (df=389)
ad_in_st	-2.66**	-2.77**	-1.55
ch_in_st	2.36*	1.60	2.30*

(b) Lexical Features

	All (df=35343)	*EC* (df=19162)	*LC* (df=15726)
ad_ar	3.21**	3.07**	2.67**
ch_ar	2.41*	1.75*	2.07*
ch_va	2.09*	1.41	2.06*

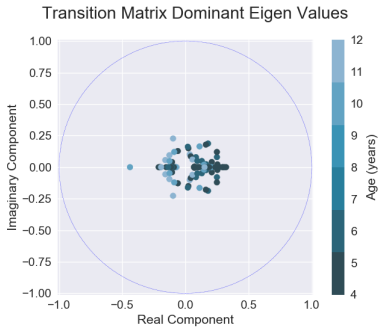

Figure 2: Dominant Eigenvalues associated with transition matrix of child state evolution. This is indicative of the dynamics that are dominantly present in the observation time-series

with productivity, suggesting that narratives by children expressing positive but strong emotion tend to facilitate productivity in the following responses.

4.2 Dynamic Modes Analysis

Performing DMDc and evaluating the eigenvalues of the transition matrix helps identify the dominant dynamical characteristics associated with the child, as well as visualizing common behavioral types across children. In Figure 2 it can be observed that many children in the EC age group have eigenvalues that lie closer to +1 consistent with slow decaying and dampened oscillation. Meanwhile older children generally lie on the other side of the unit circle implying faster decaying exponentials. This can imply that children who are older generally have behavior that is more sporadic and closer to impulse characteristics, while younger children have behaviors that are more constant.

Since the controller, B, is not square in this case, eigenvalue decomposition cannot be performed. Instead singular values can be interpreted in a similar fashion. Figure 3 demonstrate the distributions of singular values rounded to their closest integer values.

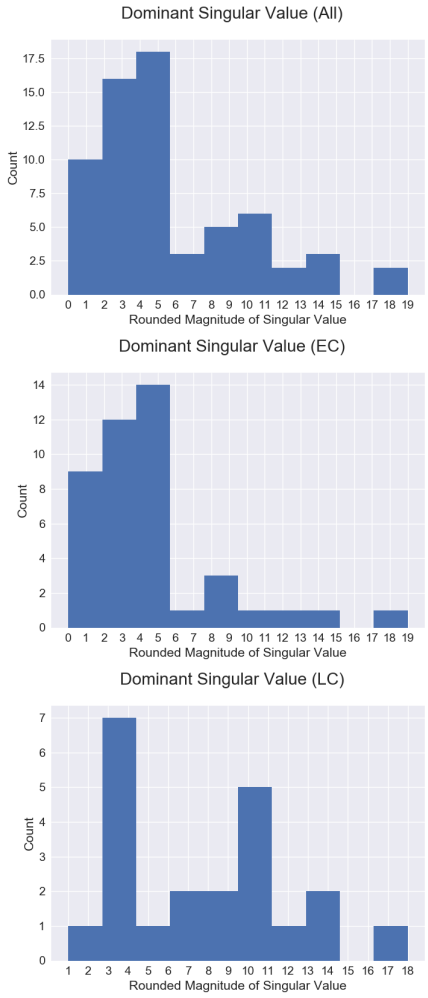

Figure 3: Distribution of dominant singular values. *All* representing all age groups together, *EC* and *LC* representing "Early Childhood" and "Late Childhood" respectively. Larger dominant singular values indicate larger contribution to output from the control input.

Large singular values correspond to large gain applied on the inputs. Children in EC generally have smaller singular values, implying that they respond slower to the inputs of the interviewer than the children in the LC group, whose singular values are generally more spread out, implying a varying degree of responses across the group.

Similar to LME, DMDc enables the analysis of model parameters to understand dominant contributors in the interactions. Figure 4 shows that in general, the child's preceding productivity is the most indicative component in predicting their future state. Similarly, the the preceding "rolling agenda" observations are most indicative of the child's productivity. This is expected since the productivity is a measure of the child's alignment and responsiveness to the interviewer's inquiries about agenda specific topics. Of particular interest though is the observation of the remaining features,

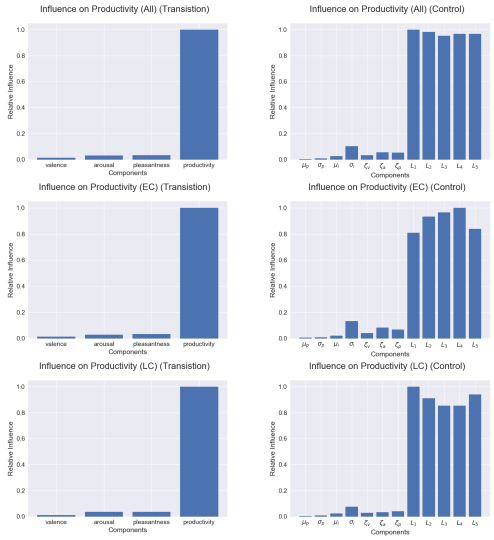

Figure 4: Relative scale of influence on productivity across age groups and features. Average absolute value of row used to calculate contribution to productivity. Contributing components from the derived "transition" (*left*) showing how the child's own state contributes to their future state, and "controller" (*right*) indicative of which components of interviewer speech contributes to the child's productivity the most. μ_i, σ_i correspond to the mean and standard deviation of the therapist prosodic intensity, while μ_p, σ_p correspond to the pitch. $\zeta_v, \zeta_a, \zeta_p$ correspond to the valence, arousal, and pleasantness of the interviewers utterance. $[L_1 - L_5]$ correspond to the 5 most important "agenda" and their appearance in the previous utterances.

the variability of the interviewers prosodic intensity is the next strongest indicator of productivity in the child's response. This supports the results found in Table 1 which indicated a statistically significant correlation between the two variables.

Figure 4 also suggests that affective lexical features make up a larger part of the interviewer's influence in EC children than it does in LC children. This suggests that the content of the words being said and questions being asked of the child become more important the child gets older, and presumably will have better developed language skills.

4.3 Evolving Dynamics

Online and Windowed DMDc algorithms allow for continually update dynamics models as more data becomes available. This is a highly useful tool when evaluating systems, especially when the dynamics observed could be time-varying. The Online DMDc algorithm efficiently updates the transition and controller matrix at every time step taking into consideration the entire past. In contrast, Windowed DMDc performs a similar update when new data is made available, but will only consider the data from a predefined sized window.

To evaluate the effectiveness of the Online and Windowed DMDc algorithms we first compute an "Oracle", which is a full DMDc performed on the entire time-series as to minimize the error globally. Then after performing an initial fit both updating DMDc algorithms and the "Oracle" are used to predict each next time step, after which the update step occurs.

Table 2: Comparing Online DMDc and Windowed DMDc. Values represent relative error when compared to an "Oracle"

Age Group	Online	Windowed
All	0.17	1.04
EC	0.19	1.47
LC	0.13	0.27

Table 2 shows that the Online DMDc algorithm performs better than the Windowed DMDc algorithm when compared to the oracle. This implies that keeping track of the entire history is generally better than only keeping track of local window, suggesting that behavior at the beginning of an interview will continue to remain informative and indicative of behavior in the future. The Windowed "forgets" past experiences; over time the dynamics it derives will change more rapidly than those computed using the Online DMDc algorithm. The smaller errors produced by the Online DMDc, suggest that the behavior observed during rapport building can be highly predictive of the behavior observed in the disclosure phase. The difference in errors between the Online and Windowed versions of DMDc drastically reduce for children in LC, suggesting that the effect of early parts of interview (rapport-building) on the productivity reduce in significance when compared to children in EC. In other words, rapport-building plays a dominant role in influencing session dynamics for younger children.

5 CONCLUSIONS

This work presented a number of different frameworks for the analysis of influence and dynamics of productivity in FI. In LME Models we were able to identify statistically significant indicators of productivity using first-order linear approximations. These indicators were verified and expanded upon when exploring broader nonlinear modeling techniques across a battery of DMDc algorithms.

DMDc provided meaningful visualizations and insights into the intrinsic dynamics and influences predictive of child behavior. Furthermore, DMDc demonstrated the temporal relationships present between rapport building and disclosure phases of interview.

To the best of our knowledge this is the first work to computationally model the interaction dynamics of FI and we believe this is a strong starting point towards developing BSP driven interview strategies.

6 FUTURE WORK

The methods we presented are highly dependent on the ability of prosodic feature extraction, which currently leaves room for improvement of child voice detection and utterance acoustic-lexical alignment. As a result when performing the time-series dynamic system modeling it was not currently possible to incorporate child acoustic-prosodic features. In the future, as voice recognition technologies improve a richer feature set with a more diverse variety of

prosodic features will be able to drive richer models of interaction dynamics.

ACKNOWLEDGEMENTS

We would like to thank the Child Forensic Interviewing Lab at the Gould School of Law for their efforts in collecting and sharing their data on FI with us.

REFERENCES

[1] Elizabeth C Ahern, Samantha J Andrews, Stacia N Stolzenberg, and Thomas D Lyon. [n. d.]. The productivity of wh-prompts in child forensic interviews. *Journal of interpersonal violence* ([n. d.]), 0886260515621084.

[2] Elizabeth C Ahern, Stacia N Stolzenberg, and Thomas D Lyon. 2015. Do prosecutors use interview instructions or build rapport with child witnesses? *Behavioral sciences & the law* 33, 4 (2015), 476–492.

[3] V. Ardulov, M. Kumar, S. Williams, T. Lyon, and S. Narayanan. 2018. Measuring Conversational Productivity in Child Forensic Interviews. *ArXiv e-prints* (June 2018). arXiv:cs.CL/1806.03357

[4] Lucy Berliner and Jon R Conte. 1995. The effects of disclosure and intervention on sexually abused children. *Child abuse & neglect* 19, 3 (1995), 371–384.

[5] Paul Boersma. 2006. Praat: doing phonetics by computer. *http://www. praat. org/* (2006).

[6] Michael S Brady, Debra A Poole, Amye R Warren, and Heather R Jones. 1999. Young children's responses to yes-no questions: Patterns and problems. *Applied Developmental Science* 3, 1 (1999), 47–57.

[7] Rafael Calvo, Sidney D'Mello, Jonathan Gratch, Arvid Kappas, Chi-Chun Lee, Jangwon Kim, Angeliki Metallinou, Carlos Busso, Sungbok Lee, and Shrikanth S. Narayanan. 2015. Speech in Affective Computing. *The Oxford Handbook of Affective Computing* August 2018 (2015), 1–23. https://doi.org/10.1093/oxfordhb/9780199942237.013.021

[8] Roger Collins, Robyn Lincoln, and Mark G Frank. 2002. The effect of rapport in forensic interviewing. *Psychiatry, psychology and law* 9, 1 (2002), 69–78.

[9] Lindsay E. Cronch, Jodi L. Viljoen, and David J. Hansen. 2006. Forensic interviewing in child sexual abuse cases: Current techniques and future directions. *Aggression and Violent Behavior* 11, 3 (2006), 195–207. https://doi.org/10.1016/j.avb.2005.07.009

[10] David M Fergusson, L John Horwood, and Michael T Lynskey. 1996. Childhood sexual abuse and psychiatric disorder in young adulthood: II. Psychiatric outcomes of childhood sexual abuse. *Journal of the American Academy of Child & Adolescent Psychiatry* 35, 10 (1996), 1365–1374.

[11] James Gibson, Nikolaos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Proc. Interspeech, Dresden, Germany*. 1947–1951.

[12] Tanja Hillberg, Catherine Hamilton-Giachritsis, and Louise Dixon. 2011. Review of meta-analyses on the association between child sexual abuse and adult mental health difficulties: A systematic approach. *Trauma, Violence, & Abuse* 12, 1 (2011), 38–49.

[13] Tomoharu Iwata and Shinji Watanabe. 2013. Influence relation estimation based on lexical entrainment in conversation. *Speech Communication* 55, 2 (2013), 329–339. https://doi.org/10.1016/j.specom.2012.08.012

[14] Jeffrey H. Kahn, Renée M. Tobin, Audra E. Massey, and Jennifer A. Anderson. 2007. Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology* 120, 2 (2007), 263–286. https://doi.org/10.2307/20445398

[15] Julie B Kaplow, Erin Hall, Karestan C Koenen, Kenneth A Dodge, and Lisa Amaya-Jackson. 2008. Dissociation predicts later attention problems in sexually abused children. *Child Abuse & Neglect* 32, 2 (2008), 261–275.

[16] Manoj Kumar, Rahul Gupta, Daniel Bone, Nikolaos Malandrakis, Somer Bishop, and Shrikanth S. Narayanan. 2016. Objective Language Feature Analysis in Children with Neurodevelopmental Disorders During Autism Assessment. In *Proc. Interspeech, San Francisco, USA*. 2721–2725.

[17] Michael E Lamb. 1996. Effects of investigative utterance types on Israeli children's responses. *International Journal of Behavioral Development* 19, 3 (1996), 627–638.

[18] Michael E Lamb and Angele Fauchier. 2001. The effects of question type on self-contradictions by children in the course of forensic interviews. *Applied cognitive psychology* 15, 5 (2001), 483–491.

[19] Michael E Lamb, Yael Orbach, Irit Hershkowitz, Phillip W Esplin, and Dvora Horowitz. 2007. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child abuse & neglect* 31, 11-12 (2007), 1201–1231.

[20] Michael E Lamb, Kathleen J Sternberg, Yael Orbach, Irit Hershkowitz, and Dvora Horowitz. 2003. Differences between accounts provided by witnesses and alleged victims of child sexual abuse. *Child Abuse & Neglect* 27, 9 (2003), 1019–1031.

[21] Chelsea Leach, Martine B. Powell, Stefanie J. Sharman, and Jeromy Anglim. 2017. The Relationship Between Children's Age and Disclosures of Sexual Abuse During Forensic Interviews. *Child Maltreatment* 22, 1 (2017), 79–88. https://doi.org/10.1177/1077559516675723

[22] Noah Liebman and Darren Gergle. 2016. Capturing Turn-by-Turn Lexical Similarity in Text-Based Communication. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16* (2016), 552–558. https://doi.org/10.1145/2818048.2820062

[23] Thomas D. Lyon, Stacia N. Stolzenberg, and Kelly McWilliams. 2017. Wrongful Acquittals of Sexual Abuse. *Journal of Interpersonal Violence* 32, 6 (2017), 805–825. https://doi.org/10.1177/0886260516657355

[24] Nikolaos Malandrakis and Shrikanth S. Narayanan. 2015. Therapy language analysis using automatically generated psycholinguistic norms. In *Proc. Interspeech, Dresden, Germany*. 1952–1956.

[25] Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition & emotion* 23, 2 (2009), 209–237. https://doi.org/10.1080/02699930802204677

[26] Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *Fifth International Conference on Spoken Language Processing*.

[27] Shrikanth Narayanan and Panayiotis G. Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233. https://doi.org/10.1109/JPROC.2012.2236291 arXiv:NIHMS150003

[28] Mary L Paine and David J Hansen. 2002. Factors influencing children to self-disclose sexual abuse. *Clinical Psychology Review* 22, 2 (2002), 271 – 295.

[29] Branka Zei Pollermann. 2002. A place for prosody in a unified model of cognition and emotion. In *Speech Prosody 2002, International Conference*.

[30] Eleanor A Price, Elizabeth C Ahern, and Michael E Lamb. 2016. Rapport-building in investigative interviews of alleged child sexual abuse victims. *Applied Cognitive Psychology* 30, 5 (2016), 743–749.

[31] Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. 2014. Dynamic mode decomposition with control. 15, 1 (2014), 142–161. https://doi.org/10.1137/15M1013857 arXiv:1409.6358

[32] Lorraine Radford, Susana Corral, Christine Bradley, Helen Fisher, Claire Bassett, Nick Howat, and Stephan Collishaw. 2011. Child abuse and neglect in the UK today. (2011).

[33] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 3–9.

[34] Clarence W. Rowley, Igor Mezi, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. 2009. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics* 641, Rowley 2005 (2009), 115–127. https://doi.org/10.1017/S0022112009992059

[35] James A Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[36] Peter J. Schmid. 2010. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics* 656 (2010), 5–28. https://doi.org/10.1017/S0022112010001217 arXiv:arXiv:1312.0041v1

[37] Lauren E. Scissors, Alastair J. Gill, Kathleen Geraghty, and Darren Gergle. 2009. In CMC we trust: the role of similarity. *27th international conference on Human factors in computing systems - CHI 09* (2009), 527–536. https://doi.org/10.1145/1518701.1518783

[38] DA Segovia and AM Crossman. 2012. Cognition and the Child Witness: Understanding the Impact of Cognitive Development in Forensic Contexts. *Current topics in children's learning and cognition* (2012), 83–104. https://doi.org/10.5772/53938

[39] Victoria Talwar, Kyle Hubbard, Christine Saykaly, Kang Lee, RCL Lindsay, and Nicholas Bala. 2018. Does parental coaching affect children's false reports? Comparing verbal markers of deception. *Behavioral sciences & the law* 36, 1 (2018), 84–97.

[40] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[41] Hao Zhang, Clarence W. Rowley, Eric A. Deem, and Louis N. Cattafesta. 2017. Online dynamic mode decomposition for time-varying systems. (2017), 1–22. arXiv:1707.02876 http://arxiv.org/abs/1707.02876