Summer July 18, 2021

# 89. Causal indicators for assessing the truthfulness of child speech in forensic interviews.

Zane Durante, *University of Southern California*
Victor Ardulov, *University of Southern California*
Manoj Kumar, *University of Southern California*
Jennifer Gongola, *University of Southern California*
Thomas D. Lyon, *University of Southern California Law School*, et al.

Computer Speech and Language 71 (2022) 101263

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# Causal indicators for assessing the truthfulness of child speech in forensic interviews

Zane Durante *, Victor Ardulov *, Manoj Kumar, Jennifer Gongola, Thomas Lyon, Shrikanth Narayanan

*University of Southern California Los Angeles, CA, USA*

ARTICLE INFO

*Keywords:*
Automated deception detection
Narrative truth induction
Child forensic interviewing
Granger causal analysis

ABSTRACT

When interviewing a child who may have witnessed a crime, the interviewer must ask carefully directed questions in order to elicit a truthful statement from the child. The presented work uses Granger causal analysis to examine and represent child–interviewer interaction dynamics over such an interview. Our work demonstrates that Granger Causal analysis of psycholinguistic and acoustic signals from speech yields significant predictors of whether a child is telling the truth, as well as whether a child will disclose witnessing a transgression later in the interview. By incorporating cross-modal Granger causal features extracted from audio and transcripts of forensic interviews, we are able to substantially outperform conventional deception detection methods and a number of simulated baselines. Our results suggest that a child's use of concreteness and imageability in their language are strong psycholinguistic indicators of truth-telling and that the coordination of child and interviewer speech signals is much more informative than the specific language used throughout the interview.

## 1. Introduction

In legal proceedings and investigations involving children suspected of being the victim or witness to a crime, Child Forensic Interviews (CFIs) are administered to elicit testimony from a child in a controlled environment. Because of their particularly vulnerable developmental state, children can be perceived to produce unreliable testimony (Lyon et al., 2017). Furthermore, children can be coached and coerced to admit or omit falsely when interviewed (Talwar et al., 2018). To address these issues, legal experts have developed CFI, a structured conversation conducted with a child by a trained legal professional. CFI begins with a process called *rapport building*, in which the interviewer asks the child benign open-ended questions in order to make the child comfortable answering questions in a narrative form. Afterwards, the interviewer elicits testimony during the *recall* section by directing open-ended questions towards the topic of interest. This procedure minimizes manipulation from the child's testimony in order to allow for it to be admissible in a court.

Due to the high stakes involved, legal scholars and child psychologists are invested in finding factors that indicate whether a child is prepared to disclose information and whether the information disclosed by the child is truthful or deceptive, either by declaration or omission. However, in a real-life court setting it is inappropriate, and often impossible, to determine if the statements made were honest or deceitful. In order to study and refine interviewing strategies, legal scholars and researchers have constructed the broken toy paradigm, which experimentally predetermines whether or not the child experiences a transgression, which is a toy breaking (Lyon et al., 2014).

* Corresponding authors.
*E-mail addresses:* durante@usc.edu (Z. Durante), ardulov@usc.edu (V. Ardulov).

https://doi.org/10.1016/j.csl.2021.101263
Received 7 April 2021; Received in revised form 25 June 2021; Accepted 5 July 2021
Available online 19 July 2021

A meta-analysis of studies examining adults' ability to detect children's lies found an overall accuracy rate of 54% (Gongola et al., 2017). When examining adults' ability to distinguish between true and false non-disclosure of toy breakage in response to recall questions, Domagalski et al. (2020) found an accuracy rate of 51%. Computational approaches to detect deception in broken toy interview transcripts have leveraged language models and paralinguistic features using machine learning to gain insights into how children's language use differs depending on whether or not they are telling the truth (Yancheva and Rudzicz, 2013; Ardulov et al., 2020). These studies have focused on static and bag-of-words features in order to represent the child's and interviewer's language use.

The underpinning hypothesis of our study is that the ways in which children adapt their behavior in response to an interviewer's is a more informative signal of deception than the behavior itself. In other words, metrics that capture the ways in which children regulate their behavior in response to an interviewer will be better indicators of deception than previously studied aggregated session-level representations. Our work builds upon prior results by introducing acoustic features and Granger causality analysis to capture information about interactions between the child and the interviewer. These Granger causal features capture the relationship of a child's responses to an adult's questions and statements through time. Our results demonstrate that these features are not only powerful predictors of truthfulness and disclosure but can also characterize how child–interviewer interactions differ when a child is lying or telling the truth. These results can inform interviewers to create new strategies for adapting their speech and language to better identify truthful statements or determine if a child is prepared to disclose that a transgression occurred.

## 2. Background

### 2.1. Forensic interviews with children

In legal and investigative proceedings involving children, they are often the sole victim or witness to a crime (Lamb et al., 2003; Radford et al., 2011). The same developmental attributes which make a child vulnerable or complicit to maltreatment and abuse also make the child's testimony susceptible to manipulation and dismissal in court (Lyon et al., 2017). To combat this, the CFI protocol is administered by a trained professional to elicit reliable testimony. These interviews are designed to minimize retraumatization and maximize information retrieval without the use of coercion or leading questions. Studies have shown that interviewers are able to obtain reliable information during these interviews (Brown and Lamb, 2015).

The CFI relies on a two phase approach: rapport building and incident recall. During *rapport building*, the interviewer will ask about innocuous topics to get the child comfortable narrating and responding to open-ended questions. Once the interviewer feels that the child is in a state where they are reliably responding, they will begin the recall section of the interview and ask questions more directly pertinent to the investigation. Since the questions are open-ended, the child is not pressured to disclose specific details, and thus the testimony can be legally admissible in future proceedings.

Researchers have studied notions of success and verbal productivity in CFI administered in real-life court conditions with both manually and computationally generated labels (Lamb, 1996; Ahern et al., 2015; Price et al., 2016; Talwar et al., 2018; Ardulov et al., 2018). However, since these methods are focused on identifying moments of disclosure and quantifying the amount of information disclosed, they provide an incomplete measure of interview quality. Furthermore, it is often impossible to evaluate the truthfulness of these statements. To address these issues, the broken toy paradigm was introduced as an experiment designed to see how children respond to the CFI protocol when asked about minor transgressions (Lyon et al., 2008). Details of the protocol are described in Section 3.

### 2.2. Deception detection

When presented with statements made by children during forensic interviews, adults performed just slightly better than random guessing, yielding an average accuracy of 54% (Gongola et al., 2017). This work further alludes to the fact that relevant professional backgrounds did not significantly impact the ability of the adult to predict whether a child was telling the truth, and that adults performed better (59% accurate) when identifying true non-disclosure statements compared to false non-disclosure (49% accurate). More recent studies examined the effects of specific interview instructions and their impact on adults' abilities to detect deception in interviews (Gongola et al., 2018, 2020). These studies showed that adults tend to be biased towards believing that a child's statements are truthful (Gongola et al., 2020), and that additional specific interview instructions only slightly improved adult performance (Gongola et al., 2018).

Automated deception detection in courtroom settings has been largely confined to adult subjects, utilizing multi-modal streams of features including video, audio, and text (Mihalcea and Strapparava, 2009; Pérez-Rosas et al., 2015; Mathur and Matarić, 2020). Previous work on automated deception detection in child broken toy experiments used syntactic linguistic features (Yancheva and Rudzicz, 2013) and bag-of-word representations of vocabulary supplemented with psycholinguistic norms (Ardulov et al., 2020).

In contrast, the work presented here utilizes novel acoustic features and considers the coordination and relative causal dependencies between child and interviewer turn-level features to better understand the dynamics of child truth-telling within the interview. Additionally, a new task is introduced which evaluates the rapport-building phase to see if there are causal relationships that indicate whether or not a child will disclose later during recall, under the assumption that a transgression occurred.

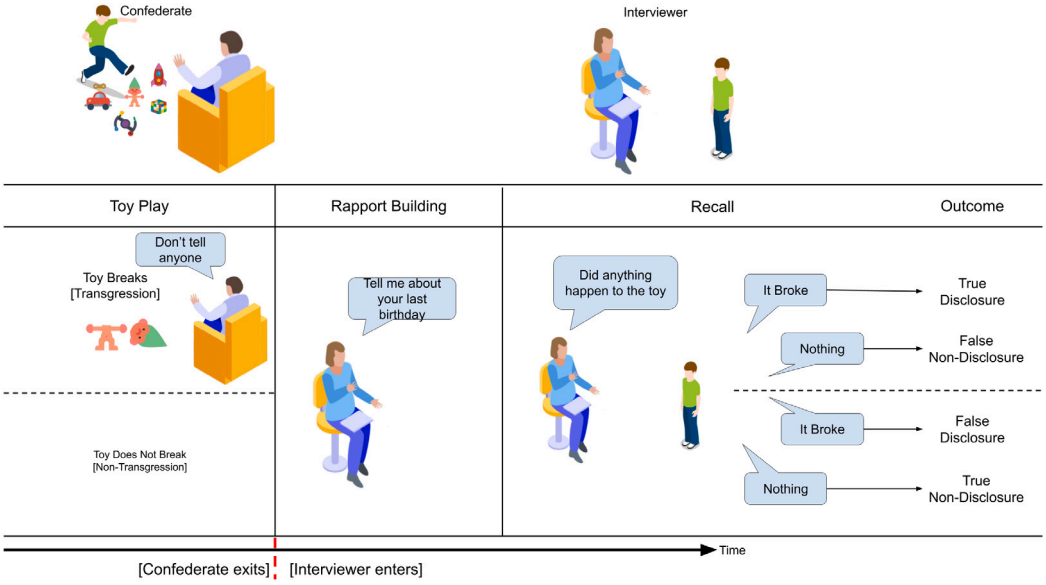


**Fig. 1.** An overview of possible interview outcomes depending on transgression (toy-breaking) and disclosure conditions.

**Table 1**
Number of session transcripts in the dataset for each transgression and disclosure condition.

| | Transgression | No transgression |
|---|---|---|
| Disclosure | 40 | 2 |
| Non-Disclosure | 109 | 49 |

## 3. Data

### 3.1. Dataset

The data consist of 200 interactions between a child and two adult experimenters: a *confederate*, and an *interviewer*. Each session is associated with a unique child, while the same adults may appear in multiple sessions. Specific demographic information about each child and experimenter is unknown, however, generally the children included in the study are of elementary school age and the experimenters are legal and psychology domain experts trained in the CFI protocol.

A session begins with the *confederate* and the child playing in a room full of toys. The confederate is informed prior to the session as to whether or not one of the toys is designed to break during play. If a toy breaks, then a *transgression* has occurred, and the confederate tells the child that an interviewer will enter the room to ask some questions. The confederate then asks the child to promise not to tell the interviewer about the toy breaking because they "could get in trouble". The confederate exits the room, and the interviewer enters.

Blind as to whether or not a transgression occurred, the interviewer follows a modified CFI protocol beginning with the *rapport building* phase, adhering to a semi-structured script that does not discuss the toys. After rapport building, the interviewer enters the *recall* phase, and asks the child to name each toy and describe what happened with it. The interviewer only asks the child about each toy exactly once and only repeats if the child indicates that they did not understand. The interaction is then transcribed and annotated as to whether the child indicated that one of the toys broke, which is referred to as a *disclosure*. The frequency of each transgression and disclosure condition in our dataset is shown in Table 1, while Fig. 1 illustrates the different stages of the interview and all of the possible outcomes.

### 3.2. Psycholinguistic norms

Psycholinguistic norms are a numerical representation of a word's general perceived alignment with certain affective and cognitive measures. Malandrakis et al. (2011) developed EmotiWord, a dictionary that maps words to psycholinguistic norms constructed via crowd-sourced perception. Each psycholinguistic norm lies on a bounded and continuous domain of $[-1, 1]$. For example, a word's pleasantness measures its degree of pleasant feelings and "cookies" has a pleasantness of $+1$, while "bedbug" has a pleasantness of $-1$.

More specifically, this study uses the psycholinguistic norms of *valence*, *arousal*, *pleasantness*, and *age of acquisition*, as these affective and cognitive signals have been shown to be predictive of child deception in Ardulov et al. (2020). Child interviewers have also demonstrated that children utilize vague language as an attempt to avoid admitting to a transgression (Clemens et al., 2010; Gongola et al., 2021), so *concreteness* and *imageability*, which correspond to a word's descriptiveness and clarity, are also used as psycholinguistic indicators of deception.

Psycholinguistic norms for each word spoken by the child and the interviewer (excluding backchannels) are averaged along each conversational turn. The constructed time-series signal captures an observable representation of how the child responds to the language of the interviewer. In the event that a turn has no words, such as in cases of non-verbal or exclusively back-channeled responses, the value is coded as neutral and represented by 0.

### 3.3. Acoustic features

Annotated time boundaries for the rapport building and recall sections of the interview are used to generate segments to extract features from the interview audio. Each segment is then processed using an off-the-shelf model[1] to align the audio with the text transcriptions using *forced alignment* (Moreno et al., 1998), in which each word in the transcript is aligned with a timestamp irrespective of the confidence of the alignment. Turn-level features for *speaking rate* and *latency* are then calculated for both the child and the interviewer.

## 4. Methods

### 4.1. Classification tasks

For each binary classification task described below, a set of models are trained to classify a given interview session. Specifically, Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), and Linear Support Vector Machine (L-SVM) models are trained using stratified 5-fold cross-validation (CV). Results and baselines are reported as the average CV F1-score, accuracy (Acc.), precision (Prec.), and false negative rate (FNR). For the DT and RF models, the splitting criterion was tuned, while for the L-SVM the penalty parameter and the use of class balanced loss penalties were tested. In our analyses, we report the performance and associated hyperparameters with the highest average CV F1-score. Model implementations were used from *scikit-learn*.[2]

#### 4.1.1. Truth-telling task

The **truth-telling task** is an evaluation of the interactions that resulted in non-disclosure. A true non-disclosure corresponds to when the toy did not break and the child did not disclose ($n = 49$). In contrast, a false non-disclosure is an instance in which the toy breaks, but the child does not disclose that the toy broke ($n = 109$). Turn-level features from both the rapport building and free recall sections of the interview are used for the analysis. The truthful (true non-disclosure) and deceptive (false non-disclosure) interviews are labeled as the positive and negative classes, respectively.

Due to the relatively poor performance of adults on child deception detection, which is only 54% accurate on average (Gongola et al., 2017), two baselines are constructed using bootstrapping: a simulated human baseline, $h^2$, and a probabilistic sampling from the training distribution, $\sigma_t$. After 10,000 simulations, the baselines are set at two standard deviations above the mean, corresponding to the $97.5^{th}$ percentile of the 10,000 simulations. See Table 2.

#### 4.1.2. Disclosure task

To discover indicators during rapport building that may signal to an interviewer that the child is ready to disclose, the **disclosure task** evaluates the turn-level features extracted from the rapport building phase to predict if a child will disclose during the recall phase. Here, we only consider the case where a transgression occurred; thus, the two outcomes of interest are a true disclosure ($n = 40$) and a false non-disclosure ($n = 109$). Similar to the truth-telling task, the truthful (true-disclosure) and deceptive (false non-disclosure) interviews are labeled as the positive and negative classes, respectively.

Since no known human baseline exists for this task, we use only the probabilistic sampling baseline, similar to the one described for the truth-telling task. For the disclosure task, the baseline is denoted as $\sigma_d$ in Table 2.

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

[2] https://github.com/scikit-learn/scikit-learn

**Table 2**
Human and probabilistic sampling baselines for both tasks. F1-score, accuracy, precision, and false negative rate (FNR) are considered for each baseline. $\sigma_t$ and $\sigma_d$ values represent thresholds of significant results for the truth-telling and disclosure tasks, respectively. $h^0$, $h^1$, and $h^2$ represent the 50th, 84th, and 97.5th percentiles of simulated human performance for the truth-telling task, respectively.

| | F1 | Acc. | Prec. | FNR |
|---|---|---|---|---|
| $\sigma_d$ | 0.462 | 0.653 | 0.482 | 0.257 |
| $\sigma_t$ | 0.475 | 0.633 | 0.493 | 0.277 |
| $h^0$ | 0.448 | 0.542 | 0.388 | 0.312 |
| $h^1$ | 0.505 | 0.591 | 0.440 | 0.264 |
| $h^2$ | 0.561 | 0.639 | 0.494 | 0.217 |

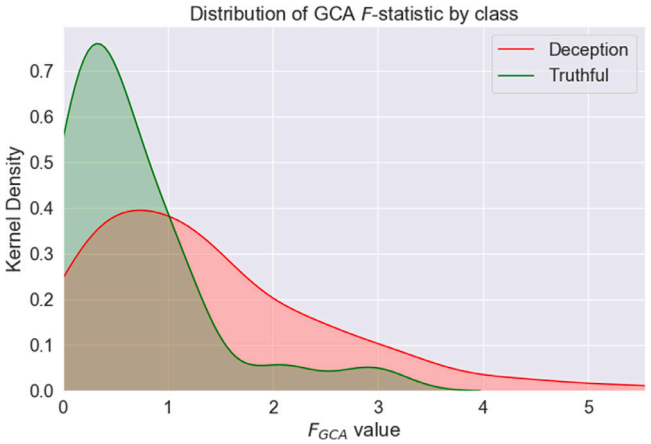


**Fig. 2.** The distribution for $F_{GCA}$ comparing child imageability in response to adult imageability for the truth-telling task. Kernel density smoothly approximates the probability density of the two distributions. As shown in Fig. 3, the imageability–imageability $F_{GCA}$ is the most predictive causality score for the truth-telling task.

### 4.2. Representing interaction dynamics

To capture information surrounding the influence of the interviewer on the child's behavior, we utilize a measure of dynamic coordination known as Granger causality analysis (GCA). Although GCA was originally developed for econometric forecasting (Granger, 1969), it has been shown to significantly measure coordination and synchrony between interlocutors (Kalimeri et al., 2011, 2012; Bone et al., 2014). Thus, the strength of temporal causality between the interviewer's and child's speech signals can be interpreted as a measure of coordination between the two interlocutors.

Explicitly, for two given signals $Y_{[0:T]} = \{y_0, y_1, \dots, y_T\}$ and $X_{[0:T]} = \{x_0, x_1, \dots, x_T\}$, $X$ is said to "*Granger cause*" $Y$ if the auto-regressive error, $\epsilon_{\alpha,t}$, is significantly larger than the error of the influence model, $\epsilon_{\beta,t}$, according to an F-test. GCA implies that the inclusion of signal $X$ in Eq. (1) with a lag $L$ better explains the observation of $Y$ than the auto-regressive model shown in Eq. (2):

$$y_t = \vec{a} \cdot Y_{[0:t]} + \vec{b} \cdot X_{[0:t-L]} + \epsilon_{\beta,t} \tag{1}$$

$$y_t = \vec{a} \cdot Y_{[0:t]} + \epsilon_{\alpha,t} \tag{2}$$

The GCA yields an $F$-statistic, $F_{GCA}$, and a corresponding $p$ value which can be interpreted as the strength of the influence and a measure of how likely the relationships is to occur by chance.

Given an interview session, a pair of turn-level child and adult speech signals, $(X_{[0:T]}, Y_{[0:T]})$ respectively, are extracted. A causality score, $F_{GCA}$, is computed for each combination of available speech signals (both acoustic and psycholinguistic) by applying the GCA with a maximum lag of 5.

To determine which causality scores are most predictive, the $F_{GCA}$ distributions are evaluated using a one-way analysis of variance (ANOVA). An example for how the distribution of causality scores varies based on the underlying transgression and disclosure conditions can be seen in Fig. 2, which demonstrates the distribution of the child–adult imageability–imageability $F_{GCA}$ for the truth-telling task.

These ANOVA results were used for feature selection by constructing 10 feature sets per task, consisting of the causality scores for which the ANOVA significance yielded $p < P_{max}$, where $P_{max} \in [0.05, 0.10, 0.15, \dots, 0.50]$. A set of models were trained on each feature set, and the feature sets using ANOVA significance thresholds of $p < 0.15$ and $p < 0.2$ yielded the models with the highest cross-validation F1 score for the truth-telling and disclosure tasks, respectively. These feature sets are visualized in Figs. 3 and 4.

For certain ANOVA thresholds and speech signal combinations, GCA yields multiple lag values that meet the inclusion criteria. Thus, our study also compared using only the lag with smallest $p$-value (*single-lag*) with the inclusion of all lags that meet the

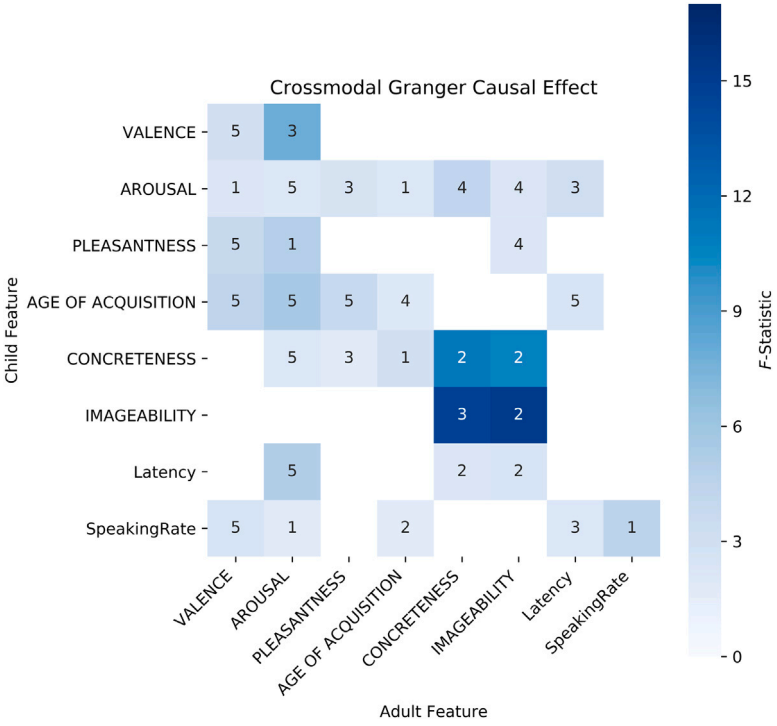


**Fig. 3.** A visualization of the feature set that obtained the model with the highest F1 score for the truth-telling task. The $F$-statistic measures which combinations of signals are most differentiating for the truth-telling task according to a one-way ANOVA. The annotation indicates which lag had the strongest measured causal effect. All $F$-statistics reported with a $p < 0.15$. The distribution of the strongest predictor, imageability–imageability, can be seen in Fig. 2.

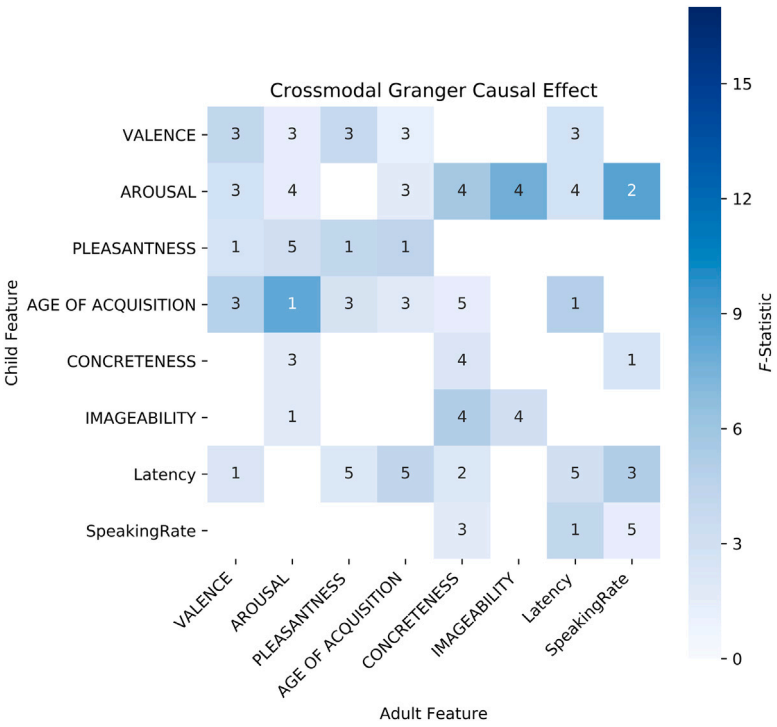


**Fig. 4.** A visualization of the feature set that obtained the model with the highest F1 score for the disclosure task. The $F$-statistic measures which combinations of signals are most differentiating for the disclosure task according to a one-way ANOVA. The annotation indicates which lag had the strongest measured causal effect. All $F$-statistics reported with a $p < 0.20$.

**Table 3**
Truth-telling task performance using session-level aggregated features and a feature significance threshold of $p < 0.30$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively. The best performing L-SVM hyperparameters, shown in the table above, are $C = 1$ along with balanced class weights. The decision tree classifier used entropy as the splitting criteria, while the random forest classifier used Gini impurity.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.540[a] | 0.662[b] | **0.587**[b] | 0.268[a] | 0.525 | 0.737 |
| RF | 0.324 | 0.633 | 0.517[b] | 0.328 | 0.250 | **0.840** |
| GNB | 0.378 | 0.536 | 0.370 | 0.354 | 0.393 | 0.613 |
| L-SVM | **0.640**[b] | **0.719**[b] | 0.584[b] | **0.165**[b] | **0.725** | 0.718 |

[a]Indicates performance better than randomized bootstrap $\sigma_t$.
[b]Indicates performance better than human simulation $h_2$.

**Table 4**
Disclosure task performance using session-level aggregated features. Using feature significance of $p < 0.20$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively. The best performing L-SVM hyperparameters, shown in the table above, are $C = 0.1$ along with balanced class weights. The decision tree classifier used entropy as the splitting criteria, while the random forest classifier used Gini impurity.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.452 | 0.641 | 0.451 | 0.259 | 0.471 | 0.725 |
| RF | 0.229 | 0.633 | 0.327 | 0.316 | 0.186 | **0.849** |
| GNB | 0.480[a] | 0.603 | 0.434 | 0.256 | 0.557 | 0.622 |
| L-SVM | **0.609**[a] | **0.703**[a] | **0.531**[a] | **0.161**[a] | **0.719** | 0.697 |

[a]Indicates performance better than randomized bootstrap $\sigma_d$.

**Table 5**
Truth-telling task performance using Granger causal features and a feature significance threshold of $p < 0.15$. * indicates performance better than randomized bootstrap $\sigma_t$. ** indicates performance better than human simulation $h_2$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively. The best performing L-SVM hyperparameters, shown in the table above, are $C = 0.01$ along with balanced class weights. The decision tree classifier used entropy as the splitting criteria, while the random forest classifier used Gini impurity.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.408 | 0.582 | 0.387 | 0.306 | 0.439 | 0.657 |
| RF | 0.391 | 0.670** | 0.521** | 0.294 | 0.329 | **0.851** |
| GNB | **0.749**** | **0.797**** | **0.684**** | 0.107** | 0.836 | 0.775 |
| L-SVM | 0.713** | 0.767** | 0.614** | 0.084** | **0.857** | 0.716 |

$p < P_{max}$ criteria (*multi-lag*). This comparison can be seen in Appendix A, but generally, models trained on the single-lag feature sets outperformed models trained on the multi-lag feature sets. This is likely due to the substantial overlap between $F_{GCA}$ features that are created from the same speech signals and only differ by lag value.

### 4.3. Static baseline models

In order to determine the effectiveness of using Granger causality to model the child interview, we train and evaluate a set of baseline models (GNB, DT, RF, L-SVM) on the same interview cross-validation set using session level averages of the child and adult acoustic and psycholinguistic features, similar to the approach in Ardulov et al. (2020). A one-way ANOVA is also used for feature selection, where 10 feature sets are created per task, consisting of the session-level features that yielded $p < P_{max}$, where $P_{max} \in [0.05, 0.10, 0.15, \ldots, .50]$. Tables 3 and 4 show the cross-validation performance of models trained on the session-level feature sets that produced the highest F1 cross-validation scores for the truth-telling and disclosure tasks, respectively.

## 5. Results

### 5.1. Truth-telling task results

Previous work used psycholinguistic norms and static bag-of-words approaches to achieve a maximum F1 score of 0.556 on the truth-telling task (Ardulov et al., 2020). However, the GNB model using a feature significance threshold of $p < 0.15$ outperforms all baselines and achieved an F1 score of 0.749 for the truth-telling task as seen in Table 5. The high positive class accuracy (0.836) of the GNB model is particularly remarkable, considering the truthful class is only 31% of the data in the truth-telling task. Both the GNB and L-SVM models outperformed all probabilistic, static, and simulated human baselines.

**Table 6**
Disclosure task performance using Granger causal features and a feature significance of $p < 0.10$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively. The best performing L-SVM hyperparameters, shown in the table above, are $C = 0.001$ along with balanced class weights. The decision tree classifier used Gini impurity as the splitting criteria, while the random forest classifier used entropy.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.400 | 0.580 | 0.396 | 0.288 | 0.443 | 0.651 |
| RF | 0.395 | 0.725[a] | **0.733**[a] | 0.269 | 0.286 | **0.938** |
| GNB | 0.614[a] | 0.725[a] | 0.559[a] | 0.164[a] | 0.686 | 0.743 |
| L-SVM | **0.648***  | **0.736**[a] | 0.585[a] | **0.129**[a] | **0.757** | 0.729 |

[a]Indicates performance better than randomized bootstrap $\sigma_d$.

### 5.2. Disclosure task results

Similar to the truth-telling task, the GNB and L-SVM models outperformed both the probabilistic and static baselines when using a feature significance threshold of $p < 0.10$. Results for each classifier can be seen in Table 6.

### 5.3. Discussion

As seen in Fig. 3, the best single predictor for whether a child is telling the truth is the imageability–imageability $F_{GCA}$. Notably, every combination of concreteness and imageability is shown to be a significant predictor. This indicates that there exists an important relationship between children's and interviewers' use of explicit, vivid language that evokes a clear mental image. Fig. 2 explores this relationship, and shows the distribution of the imageability–imageability $F_{GCA}$ feature between truthful and deceptive children. The imageability of deceptive children's language is more dependent on the imageability of the interviewer's language than truthful children's.

One interpretation of this result is that children who are planning on omitting that a transgression occurred more carefully choose their language based on the interviewer's. Thus, the child becomes more or less vague in their descriptions based on what level of specificity the interviewer is using. In contrast, a child that gives an honest non-disclosure will not modify their concreteness or imageability depending on the speech of the interviewer. This relationship suggests that novel interview protocols that require interviewers to modulate the levels of imageability in their language may more reliably track and differentiate between truthful and deceptive speech patterns.

Referencing Fig. 4, it is seen that in the disclosure task there is no relationship as strong as the ones observed for imageability and concreteness for predicting truthfulness. This observation speaks to the challenge presented to an interviewer in determining whether or not rapport has been effectively established allowing the child to feel comfortable disclosing a transgression. The results in Table 6 show that there is valuable and detectable information in the coordination of these signals. By examining the average cross-fold coefficients learned by the LSVM mode in Table 7, we identify that high coordination between a child's language complexity (age of acquisition) and the adult's arousal indicates strongly that a child is prepared to disclose. In contrast, if a child is adjusting their latency with the speaking rate of the adult, it suggests that the child is less likely to disclose. By creating methods to test a child's coordination across these behavioral signals and informing interviewers into these relationships, it may be easier to determine when an interviewer should transition to the recall phase to attempt to elicit a disclosure.

## 6. Conclusion

Our work improves upon previous approaches in automated child deception detection by introducing Granger causal features to capture child–adult interaction dynamics, and we show that these features are significantly better indicators than static aggregation methods. This suggests that a child's response to interviewers is more informative than the specific language they use throughout the interview. Additionally, we demonstrate the effectiveness of dynamic cross-modal modeling and introduce audio features for the first joint computational speech and language analysis of child deception. Furthermore, significant performance in both the truth-telling and disclosure tasks suggests the existence of speech cues that can be used to inform interviewers of whether a child is prepared to disclose and whether their statements during disclosure are truthful. These insights can be used to evaluate CFI strategies and inform improvements to existing protocols.

In the future, dynamical system models that incorporate interlocutor interaction and auto-regressive behavior may provide further insights and improved classification accuracy. Building such models would also allow for the simulation of interviews and the optimization of interview strategies within a computational framework.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Table 7**
Average cross-fold feature importances assigned by the L-SVM models for *disclosure*.

| Adult feature | Child feature | Feature weight |
|---|---|---|
| Arousal | Age Of Acquisition | 0.0426 |
| Speaking Rate | Arousal | 0.0238 |
| Latency | Speaking Rate | 0.0229 |
| Imagability | Arousal | 0.0220 |
| Concreteness | Imagability | 0.0209 |
| Latency | Age Of Acquisition | 0.0204 |
| Valence | Arousal | 0.0181 |
| Latency | Latency | 0.0155 |
| Concreteness | Arousal | 0.0119 |
| Age Of Acquisition | Latency | 0.0116 |
| Imagability | Imagability | 0.0100 |
| Valence | Valence | −0.0231 |
| Pleasantness | Valence | −0.0248 |
| Latency | Valence | −0.0262 |
| Pleasantness | Pleasantness | −0.0267 |
| Age Of Acquisition | Pleasantness | −0.0277 |
| Valence | Age Of Acquisition | −0.0335 |
| Arousal | Pleasantness | −0.0351 |
| Speaking Rate | Latency | −0.0369 |

**Table A.8**
Truth-telling task performance using Granger causal features from the multi-lag feature set and a feature significance threshold of $p < 0.30$. * indicates performance better than randomized bootstrap $\sigma_t$. ** indicates performance better than human simulation $h_2$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.433 | 0.546 | 0.421 | 0.345 | 0.471 | 0.589 |
| RF | 0.332 | 0.659* | 0.430 | 0.307 | 0.271 | **0.865** |
| GNB | **0.657*** | **0.729*** | **0.599*** | 0.153* | **0.746** | 0.716 |
| L-SVM | 0.627* | 0.719* | 0.564* | **0.129*** | 0.743 | 0.701 |

**Table A.9**
Disclosure task performance using Granger causal features from the multi-lag feature set and a feature significance threshold of $p < 0.10$. Bold values indicate the best performance in their respective columns. Pos. Acc. and Neg. Acc. indicate the accuracy for the positive and negative classes, respectively.

| Model | F1 | Acc. | Prec. | FNR | Pos. Acc. | Neg. Acc. |
|---|---|---|---|---|---|---|
| DT | 0.458 | 0.654[a] | 0.452 | 0.247[a] | 0.481 | 0.743 |
| RF | 0.150 | 0.683[a] | 0.500[a] | 0.312 | 0.090 | **0.969** |
| GNB | 0.579[a] | 0.663[a] | 0.494[a] | 0.170[a] | **0.719** | 0.637 |
| L-SVM | **0.624**[a] | **0.737**[a] | **0.600**[a] | **0.154**[a] | 0.695 | 0.759 |

[a]Indicates performance better than randomized bootstrap $\sigma_d$.

## Acknowledgments

We would like to extend our gratitude to the members of the Child Forensic Interviewing lab for their contributions, including recording and transcribing the data. We would also like to thank the participants of the study.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) of the National Institutes of Health, USA under award number R01HD087685. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Appendix A. Single-lag and multi-lag GCA features

During GCA analysis, for a given speech signal pair, multiple lag values may sufficiently separate the positive and negative classes such that $p < P_{max}$. However, using multiple significant lag values resulted in reduced model performance compared to only using the feature that has the lowest $p$-value. Models trained on the multi-lag feature sets obtained the highest F1 score when using ANOVA significance thresholds of $p < 0.30$ and $p < 0.20$ for the truth-telling and disclosure tasks, respectively. Tables A.8 and A.9 show the performance of models using these feature sets.

## References

Ahern, E.C., Stolzenberg, S.N., Lyon, T.D., 2015. Do prosecutors use interview instructions or build rapport with child witnesses? Behav. Sci. Law 33 (4), 476–492.

Ardulov, V., Durante, Z., Williams, S., Lyon, T.D., Narayanan, S., 2020. Identifying truthful language in child interviews. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8074–8078.

Ardulov, V., Mendlen, M., Kumar, M., Anand, N., Williams, S., Lyon, T.D., Narayanan, S., 2018. Multimodal interaction modeling of child forensic interviewing. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, pp. 179–185.

Bone, D., Lee, C.C., Potamianos, A., Narayanan, S., 2014. An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model. In: Fifteenth Annual Conference of the International Speech Communication Association.

Brown, D.A., Lamb, M.E., 2015. Can children be useful witnesses? It depends how they are questioned. Child Dev. Perspect. 9 (4), 250–255.

Clemens, F., Granhag, P.A., Strömwall, L.A., Vrij, A., Landström, S., Hjelmsäter, E.R.a., Hartwig, M., 2010. Skulking around the dinosaur: Eliciting cues to children's deception via strategic disclosure of evidence. Appl. Cogn. Psychol. 24 (7), 925–940.

Domagalski, K., Gongola, J., Lyon, T., Clark, S., Quas, J., 2020. Detecting children's true and false denials of wrongdoing: Effects of question type and base rate knowledge. Behav. Sci. Law 38, 612–629.

Gongola, J., Quas, J., Clark, S.E., Lyon, T.D., 2020. Adults' difficulties in identifying concealment among children interviewed with the putative confession instructions.

Gongola, J., Quas, J.A., Clark, S.E., Lyon, T.D., 2021. Adults' difficulties in identifying concealment among children interviewed with the putative confession instructions. Applied Cognitive Psychology 35 (1), 18–25.

Gongola, J., Scurich, N., Lyon, T.D., 2018. Effects of the putative confession instruction on perceptions of children's true and false statements. Appl. Cogn. Psychol. (ISSN: 10990720) (May), 1–7.

Gongola, J., Scurich, N., Quas, J.A., 2017. Detecting deception in children: A meta-analysis. Law Hum. Behav. (ISSN: 01477307) 41 (1), 44–54.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 424–438.

Kalimeri, K., Lepri, B., Aran, O., Jayagopi, D.B., Gatica-Perez, D., Pianesi, F., 2012. Modeling dominance effects on nonverbal behaviors using granger causality. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 23–26.

Kalimeri, K., Lepri, B., Kim, T., Pianesi, F., Pentland, A.S., 2011. Automatic modeling of dominance effects using granger causality. In: International Workshop on Human Behavior Understanding. Springer, pp. 124–133.

Lamb, M.E., 1996. Effects of investigative utterance types on Israeli children's responses. Int. J. Behav. Dev. 19 (3), 627–638.

Lamb, M.E., Sternberg, K.J., Orbach, Y., Esplin, P.W., Stewart, H., Mitchell, S., 2003. Age differences in young children's responses to open-ended invitations in the course of forensic interviews. J. Consult. Clin. Psychol. 71 (5), 926.

Lyon, T.D., Malloy, L.C., Quas, J.A., Talwar, V.A., 2008. Coaching, truth induction, and Young maltreated children's false allegations and false denials. Child Dev. 79 (4), 914–929.

Lyon, T.D., Stolzenberg, S.N., McWilliams, K., 2017. Wrongful acquittals of sexual abuse. J. Interpers. Violence 32 (6), 805–825.

Lyon, T.D., Wandrey, L., Ahern, E., Licht, R., Sim, M.P.Y., Quas, J.A., 2014. Eliciting maltreated and nonmaltreated children's transgression disclosures: Narrative practice rapport building and a putative confession. Child Dev. 85 (4), 1756–1769.

Malandrakis, N., Potamianos, A., Iosif, E., Narayanan, S., 2011. Emotiword: Affective lexicon creation with application to interaction and multimedia data. In: MUSCLE.

Mathur, L., Matarić, M.J., 2020. Introducing representations of facial affect in automated multimodal deception detection. In: Proceedings of the 2020 International Conference on Multimodal Interaction. ICMI '20, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450375818, pp. 305–314.

Mihalcea, R., Strapparava, C., 2009. The Lie detector: Explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. ACLShort '09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 309–312.

Moreno, P.J., Joerg, C., Thong, J.M.V., Glickman, O., 1998. A recursive algorithm for the forced alignment of very long audio segments. In: Fifth International Conference on Spoken Language Processing.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M., 2015. Deception detection using real-life trial data. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, pp. 59–66.

Price, E.A., Ahern, E.C., Lamb, M.E., 2016. Rapport-building in investigative interviews of alleged child sexual abuse victims. Appl. Cogn. Psychol. 30 (5), 743–749.

Radford, L., Corral, S., Bradley, C., Fisher, H., Bassett, C., Howat, N., Collishaw, S., 2011. Child abuse and neglect in the UK today. London: NSPCC.

Talwar, V.A., Hubbard, K., Saykaly, C., Lee, K., Lindsay, R.C.L., Bala, N., 2018. Does parental coaching affect children's false reports? Comparing verbal markers of deception. Behav. Sci. Law 36 (1), 84–97.

Yancheva, M., Rudzicz, F., 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 944–953.