A Conceptual Model of Real World, High Stakes Deception Detection

Joseph Thomas
Oklahoma State University
Joseph.thomas@okstate.edu

Dr David Biros
Oklahoma State University
david.biros@okstate.edu

ABSTRACT
There is a strong need for more field studies in deception detection research utilizing objective measures. The study of deception detection is heavily skewed toward laboratory research utilizing university students in mock lie scenarios which are then subjectively scored by humans. This paper introduces a conceptual model that may aid in detecting deception during high stakes interview style communication, utilizing text analytics tools. The model attempts to examine the interactive nature of communication channels within a real world high stakes case study. More specifically it attempts to show that when the source of interactive communication originated from a richer channel in a field setting deception detection accuracy improves even when examined in a less-rich channel. This paper describes a new model and methodological approach utilizing existing text analytics tools that has the potential to positively impact a wide variety of law enforcement agencies and others who conduct high stakes interviews.

1. Introduction

Deception detection and its cues have been formally studied in numerous cross discipline studies and the consensus is that humans are poor detectors of deceit [3], [31], [44]. The subject itself is difficult to study because of the complex nature of human interactions and the dynamic interchange that takes place in a question response session. The amount of information that passes between two people, conscious and unconscious, that must be processed in order to identify deceit is incredible.

In fact, most studies on the topic use college students instructed to lie in laboratory experiments [17]. These mock lies have been critiqued because of lack of motivation; the participants did not chose to lie hence had little or no vested interest in whether or not they got caught [35]. A lack of personal involvement in the lie is another critique of the lab studies [30]. Furthermore, there is a lack of studies utilizing objective automated measures of deceptive cues. We acknowledge the value of these laboratory studies but feel there is a need for more research examining their results in real world high stakes (RWHS) environments utilizing objective measures to the greatest extent possible.

High stakes deceptive environments are those were the parties are personally invested in the exchange and highly motivated to deceive or detect deception. Previous research has found that cues to deception differ when the motivation differs from low to high [54]. However humans are poor detectors of deceit as well as poor judges of what cues are indicators of deception and are often affected by biases toward unreliable deceptive cues which hampers our ability to accurately detect deception [42].

Deception researchers have confirmed that people are “truth bias” or regularly assume others are truthful in interpersonal encounters [6], [9], [28], [36], [38]. In contrast, law enforcement (LE) personnel have what Meissner and Kassin [33] called “investigator bias” or are significantly more likely to respond “deceitful” rather than “truthful” in an suspect interview session [33]. Many LE personnel are trained in techniques and how to spot deceptive cues but even with advanced training an interviewer’s accuracy at detecting deception is not significantly increased [42], [1], [2]. It has also been pointed out that LE training on deception detection cues could be wrong [29], [32], [43]. What is needed are unbiased tools to aid investigators during interviews with potential deceivers in high stakes environments where judgments are often initially made based off very little data.

However, research on deception detection in high stakes environments, such as interviews during a criminal investigation, are lacking [37]. Also, the preponderance of deception detection research looks at yes/no or short answer lies. (e.g. polygraph). The yes/no response to structured questioning has its value but is not characteristic of suspect interview questioning where interpersonal dyadic communication is more typical. Another predominant method in deception detection research
looks at mean scores over entire interactions, in essence detecting general deceptive states not pinpointing lies. This course granularity of deceptive episodes is of little relevance to practitioners, what is needed here is to identify the needle (lie) in the haystack (interview). Our research questions are:

RQ 1: Can reliable text- and vocal-based cues to deceptive behavior be extracted from recordings of RWHS interviews?

RQ2: Can changes in cues to deceptive behavior be identified in transcripts of RWHS interviews?

RQ3: Can changes in cues to deceptive behavior be identified between transactions of RWHS interviews?

This article is a conceptual study of the information collection, measurement, and management processes of the cues extracted from real world high stakes environments. The research method will take the form of a case study and the communication channel of interest is audio, namely the simultaneous analysis of paralanguage characteristics of speech and lexical cues from text. Section II & III give a brief overview of deception detection and the audio channel as a promising source for cues. Section IV offers a justification for utilizing the case study methodology in this context. Section V discusses the theoretical, and measurement models used in this research. Sections VI describes the methodology and data set to be tested against as well as the procedural model used to process the data for analysis. Sections VII & VIII describe the details of the case study and possible future research. Finally, conclusions and implications are in Section IX.

The contributions we hope to gain by this study are: 1) What are the appropriate constructs for use in studying deception in vocal-based, high stakes environment; 2) Better understanding of how people deceive during interview communication methods in real world high stakes (RWHS) settings; 3) Determine which vocal-based cues distinguish truthful from deceptive messages in a RWHS; 4) Begin to understand if deception detection can be automated using information systems tools, techniques, and procedures; 5) Can the veracity of individual messages be accurately determined using vocal-based cues in a RWHS interview?

2. Deception detection

Deception is defined as the knowingly transmitting a messages to foster a false belief [5]. For as long as people have been lying, people have been trying to detect lies [22]. With the advancement of new methods and new technology humans are continuing to pursue an automated system that can accurately recognize deception.

Deception detection methods can be split into two categories, invasive and non-invasive. Of the invasive technologies currently available to help identify and measure deceit, the polygraph is the most well known. In a summary of laboratory tests, Vrij reports that the polygraph is about 82% accurate at identifying deceivers [42]. Although it is not admissible in a court of law, the polygraph can be useful in investigations for identifying potential suspects or to narrow down possible leads. However polygraph exams have several strong limiters namely a willing subject, an invasive exam, and the need for a trained examiner. The newest method to detect deception utilizes functional magnetic resonance imaging (fMRI) to map blood flow in the brain during structured questioning. Though initial findings are promising, MRI shares the same restrictions as the polygraph but to a greater extent, as well as their sheer size and cost to operate.

There is a need for a non-invasive, unobtrusive, deception detection technology that is easily automated. One such advancement which failed to stand up to academic rigor was voice stress analysis (VSA). VSA was introduced in the 1970s as a possible replacement of the polygraph. VSA is a technique that analyzes the voice pitch changes as a measure of arousal. The machines built around VSA are fundamentally designed to detect stress, not lies. The accuracy of voice stress analyzers is reported to range from chance to almost that of the polygraph [24], [41]. Despite its initial promise, the voice stress analyzer has failed to gain scientific acceptance [22], [25], [26].

As mentioned earlier, a vast majority of current deception detection studies are performed in a laboratory setting using college students as subjects [46], [17]. Several meta-analyses exist that attempt to summarize the large body of studies in deception and deception detection [54], [17]. A meta-analysis of 120 studies showed 101 used student subjects. Only four of these studies involved situations where the subjects were not given instructions to lie but chose to do so on their own [17]. There is evidence that behavior differs between those who choose to lie and those directed to lie by an experimenter [20]. Therefore, studies utilizing real-world subjects who chose to be or not to be deceptive may contribute more deeply to the understanding of deception as well as provide findings which are more generalizable to reality. Generalizability should also improve if instead of testing only a few
question/response interactions per subject in a lab setting, a more realistic environment is examined where many questions are asked with potentially many deception attempts during a prolonged interview.

From these meta-analyses and other literature several theories of deception have been developed to explain the cues to deception. One of the more promising theoretical foundations to examine deception detection from in the context of an interview style interaction is Interpersonal Deception Theory (IDT) by [5]. IDT says that deception is an interactive process between sender and receiver in a back and forth nature where each “reads” the other and adapts their behavior and speech to meet their goal, deception or detection. According to the authors of IDT, communication includes both strategic and nonstrategic behaviors. Strategic behavior refers to large-scale plans and intentions, not necessarily to specific routines or tactics. Non-strategic behaviors reflect unintentional, unconscious behavior also labeled leakage [18], [19]. It is these non-strategic behaviors that we are trying to measure and test for correlations to deceptive behavior.

According to this model, both the sender and receiver of deception bring to an interaction their expectations, goals, familiarity, etc. During the interaction the sender will begin his or her deceit with certain strategies but will modify those strategies throughout the interaction based on perceived deception success. The receiver, on the other hand, begins with some level of suspicion, which is modified throughout the interaction based on credibility judgments. Both parties will likely inadvertently reveal verbal and nonverbal cues of their psychological state. In the end, both sender and receiver will be able to evaluate their success at deceiving and detecting deceit, respectively.

IDT is a good theoretical match to the LE interview process where interviews are typically longer than lab studies will allow [40]. These longer interactions may intensify the impact of repeated questioning during the IDT interaction loop.

It is believed that deception is more difficult than telling the truth because our brains are taxed more when fabricating and maintaining cohesive facts and timelines then when telling the truth [45].

It is common practice in LE interviews to “dig deeper” during questioning which would required a more elaborate deception as questioning when on. This complexity increases cognitive load which may lead to identifiable changes in the behavior of the subject such as more frequent hesitations, and a decrease in frequency of illustrators. Figure 2 is an overlay of the LE interview process on the IDT model.
A theory that may help us better understand the simultaneous exchange of information across multiple channels is Media Richness Theory (MRT) by Daft and Lengel [14]. RMT comes from the management information systems (MIS) field and defines media richness as the ability of information communicated on the medium to reduce equivocality. It tries to explain that richer, more personal means of communication are generally more effective for communication of equivocal issues than leaner, less rich media. It suggests that the richer a medium the more information that is communicated. The levels of media channels in order or richness are Face to Face, Video, Audio, then Text. Even though more recent research has failed to support the full breadth of the original MRT [15], [16] it does lend support for the proposition that relying on any single channel may be more disadvantageous than relying on multiple channels, especially when those channels are closely related as with nonverbal audio and linguistics. Granted more cues should be able to be extracted from audio obviously but MRT suggests that lexical cues will be richer when pulled from audio vs hand written.

As suggested by Carlson et. al. [8] electronically encoded media may stand the best chance for automation in support of deception detection and by extension the best channel for LE interview tool development. Carlson et. al. further state that during this encoding that some of the variety of cues may be lost. They based their study on the theories that describe aspects of this change, namely media richness and media synchronicity. Our empirical study of real world data will examine whether the encoding from face-to-face to electronic media channels will in fact result in loss of cues when examining the audio channel only.

3. Promising channel

Almost every possible communication channel has been studied with regards to deception detection. Currently people communicate across multiple channels such as email, video, telephone, and face-to-face, often simultaneously. The nonverbal channel conveys an incredible amount of information, some of which can be used for deception detection [5], [42]. Linguistic-based cues also show promise as a deception detection diagnostic [50].

Audio recordings are commonly used in LE and high stakes settings for their ease of use in a variety of conditions and their admissibility in a court of law. Audio is a strong source of deceptive cues [17]. It is also a one-to-many communication channel namely text-based linguistic cues, vocal-based cues and content can be pulled from the single channel, audio [53], [5].

The relationship between vocal characteristics and nonverbal characteristics is an intricate one. These two communication modes are interdependent with temporal referencing and the interaction of personal and situational objectives [4]. We believe studying both these communication modes simultaneously will reveal useful information towards deception detection research. Because of the complexity of multi-channel communication it may be helpful to examine deceptive cues from the theoretical lens of management information systems research.

4. Case study methodology

Before describing the current model our use of a case study in this context needs to be explained. According to Yin [48] a case study is, “an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident”. In studying deception the context of a real world situation where the stakes are high is very pertinent to the study of detecting deception. Case study method allows investigators to retain the holistic and meaningful characteristics of real-life events. Exploratory studies attempt to answer how and why questions. These questions deal with operational associations that need to be traced over time, rather than statistical frequencies alone. A major strength of case studies is their ability to trace changes over time while not being limited to cross-sectional or static assessments [48].

The ideal event in which to use a case study is when relevant behaviors cannot be manipulated or to do so would be unethical (i.e. real world high stakes deception). The overlap of case studies and experimental design are commonly regarded as quasi-experimental where behaviors cannot be manipulated but the logic of experimental design can still be applied [11].

There are some prejudices against the case study method; the greatest is a concern for a lack of rigor. Too often case studies are attacked for being sloppy and not systematic in their procedures. To address this concern the utmost care must be given to meticulous documentation of processes and methods to a level equivalent or beyond those of experimental design. Another prejudice critical of case studies is an apprehension that they provide little basis for
scientific generalization. Case studies, like experiments, are generalizable to theoretical propositions and not to populations. The goal of doing a case study and an experiment are the same, to expand and generalize theories (analytic generalization) and not to enumerate frequencies (statistical generalization). A third concern is that case studies take too long, result in massive, unreadable documents. This does not have to be the case and stems from confusion with data collection methods like ethnography and participant-observation [21], [27]. Case studies are a form of inquiry and do not rely on these methods alone [48].

One final justification for using case studies, they can offer important evidence to complement experiments -- an adjunct to experiments rather than an alternative to them. This is how we view this study.

5. Model

Examining how someone speaks leads to paralinguistic information which includes vocal characteristics of time, intensity, frequency, and fluency as well as nonverbal or linguistic based characteristics including for example syntax choice and unintentional word choices. Other aspects of what is spoken and their potential to carry deceptive cues has been studied by many researchers [7], [50], [49], [51], [52]. Zhou and Burgoon found 8 categories of linguistic cues that were above chance at detecting deception. Fuller [23] revised these categories for her study using data from a real world high stakes environment, namely quantity, specificity, uncertainty, clarity, immediacy, affect, and cognitive processing.

We believe that cues extracted from audio recordings of RWHS interviews will lead to an increase in number of significant cues as well as the number of different types of cues than from written statements.

**Table 1 Deception Detection Lexical Constructs**

<table>
<thead>
<tr>
<th>Construct</th>
<th># of Factors</th>
<th>Construct</th>
<th># of Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity</td>
<td>3</td>
<td>Immediacy</td>
<td>3</td>
</tr>
<tr>
<td>Specificity</td>
<td>5</td>
<td>Affect</td>
<td>6</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>5</td>
<td>Cognitive Processing</td>
<td>3</td>
</tr>
<tr>
<td>Clarity</td>
<td>5</td>
<td>Total:</td>
<td>30</td>
</tr>
</tbody>
</table>

Our study will contain linguistic-based constructs from Fuller [23] and the fluency construct from Meservy [34]. Fuller’s study looked at 370 written suspect statements given during law enforcement interviews following real world high stakes criminal cases. The seven cue domains, or constructs, Fuller examined are in Table 1.

**Table 2 Linguistic-based constructs and their measurements**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Brief Description</th>
<th>Construct Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity</td>
<td>Length of message</td>
<td># of Words, Verbs, &amp; Sentences</td>
</tr>
<tr>
<td>Specificity</td>
<td>Amount and type of details in the message</td>
<td>Sensory ratio, Spatial ratio, Temporal ratio, Content Word Diversity, Bilogarithmic Type-Token-Ratio</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Relevance, directness, and certainty of message</td>
<td>Certainty Terms, Tentative Terms, Modal Verbs, Passive Voice, Generalizing Terms</td>
</tr>
<tr>
<td>Clarity</td>
<td>Message clarity and comprehensibility</td>
<td>Redundancy, Sentence Length, Complexity Ratio, Average Word Length, Causation Terms.</td>
</tr>
<tr>
<td>Immediacy</td>
<td>Attempts to disassociate oneself from the events described</td>
<td>1st person pronouns, 2nd person pronouns, 3rd person pronouns</td>
</tr>
<tr>
<td>Affect</td>
<td>Emotions present in the message</td>
<td>Positive Activation, Negative Activation, Positive Imagery, Negative Imagery, Positive Pleasantness</td>
</tr>
<tr>
<td>Cognitive Processing</td>
<td>Increased or decreased cognitive processing and cognitive information present in the message related to veracity</td>
<td>Exclusive Verbs, Motion Words, Cognitive Processing Terms.</td>
</tr>
</tbody>
</table>

Fuller’s constructs were chosen because they generated almost 74% accuracy in deception
detection, the data was field data taken in high stakes LE environments with solid ground truth validation, and the unit of measure were entire written statements.

This matches our model with the exceptions that our data is a transcript of a LE interview and our unit of measure is at the question/response level. An eighth construct of severity was also considered important in fuller’s study however it is not a part of the current study because its measure would be a constant across the data set. As will be discussed later our data comes from a serial rapist, the punishment for which was life in prison. The lead detective in this case, as well as most rational people, would assign the max severity score of 5 on the scale used by Fuller. Therefore the severity construct will be left for further research when more data becomes available. The constructs along with their measurements are described in Table 2.

### Table 3 Vocal-based constructs and their measurements

<table>
<thead>
<tr>
<th>Construct</th>
<th>Brief Description</th>
<th>Construct Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Speech disturbances other than “um”s, “ur”s, “ah”s, and other such words. General speech errors. Overlaps of subject and interviewer that results in a change of turns</td>
<td>1. Non-ah disturbances 2. Speech errors 3. Interruptions</td>
</tr>
</tbody>
</table>

The vocal constructs examined by Meservy [34] were selected for this study because they were a very thorough coverage of the audio channel, they were successfully used on field data, and tools exists to measure each which consists of: duration, tempo, intensity, frequency, fluency, and voice quality. However, only the construct of fluency contains measurements that can be captured from a transcript. The fluency construct is described along with its measurements in Table 3. Study of the remaining vocal constructs in the context of a RWHS setting will be left to further research. Our measurement consists of 33 total measures across the eight deception detection constructs as depicted in Figure 2.

For our study we want to consider any non-linguistic sources of information that could lead to improved deception detection but could still be captured from a transcript. As described above, the vocal-based channel is a potentially rich source of deceptive cues. These vocal-based cue domains are time, intensity, frequency, and fluency [39]. The measurements used by Meservy [34] for each domain were time, intensity, frequency, and fluency. However, the only cues that can be captured from a transcript are from the fluency domain, namely non-ah pauses, speech errors, and interruptions. The examination of all vocal-based cue domains under RWHS environments will be left for further research.

Because our model is conceptual and construct validation has not been performed, the direction of each measurement with respect to the DV is notional and taken from previous literature. The principal reason cue measurement and construct validation have not been performed at this time is a matter of resources. Cue measurement of this scale in the area of audio analysis is extremely labor intensive, requiring technical skills and equipment still being acquired and the process for automation is under development.

### Figure 2 Measurement Model

![Figure 2 Measurement Model]

6. Methodology & data preparation

Since our dependent variable is dichotomous the preferred exploratory analysis methods to determine which cues are related are discriminant analysis and linear regression [10]. We will use a stepwise linear probability model (stepwise linear regression for binary DV) to discover the combinations of cues that when taken together predict the maximum amount of variance in deceptive outcomes. One goal of this study is to identify the cues that distinguish truthful and deceptive messages in the high-stakes, real-world context. To ensure that our cues are measuring what we believe they should, factor analysis will be conducted.
For hypotheses testing logistic regression is preferred [10]. Logistic regression can help answer the questions, “How well does the model predict or explain group membership in the binary dependent variable?”; “To what extent does each predictor variable contribute to the probability of a case being in one group or the other of the dependent variable?”; and “What is the probability that a particular case is in one group or the other of the dependent variable?”.

Our data preparation process will follow the steps shown in Figure 3. First the raw video stored on DVD will be processed with Adobe Soundbooth to isolate the audio from the video portion. The digital audio files will then be passed through DC Live Forensic 7.5 to improve audibility in preparation for transcription. Once audibility quality is improved speakers will be segmented into separate channels while maintaining time coding. The suspects audio channel will then undergo transcription using Docsoft: AV. The transcript will be manually verified for 100% accuracy. Linguistic cues will be measured from transcripts using Generalized Architecture for Text Engineering (GATE) [12], [12], [13] for text processing as was done by Fuller [23]. Waikato Environment for Knowledge Analysis (WEKA) [47] is used for classification based on the initial text processing steps. This transcript will be compared to the LE transcript and ground truth and deceptive statements will be identified in the full transcript.

7. The case study subject

In Nov 2004 the subject of this case study was sentenced in federal court to 470 years in prison for creating child pornography, rape, sexual exploitation of children, child sexual assault and kidnapping. It is one of the longest sentence for sex crimes in history and there is no parole option in the federal system, i.e., RWHS situation. LE videotaped three consecutive days of interviews totaling 14 hours and 27 min of video. Interviews were conducted by the same lead detective and their partner in the same room and under the same conditions with the subject and his attorney. Interaction was primarily between lead detective and the subject, only minor contributions (less than 5 min total) were made by the second detective and the subject’s attorney; their voice will be removed before analysis. A 200 page transcript was generated by the lead detective immediately after the interviews. Both the videotaped interviews and LE transcripts were used in federal court. Ground truth was established by the lead detective based on credible evidence admissible in a federal court. The lead detective identified two types of statements: known lies and suspected lies. Know lies were those statements proven to be false by evidence admissible in court. Suspected lies were those statements LE personnel believed in their expert opinion to be false but for which they had little or no hard evidence. For this study only known lies were considered valid deceptive statements.

8. Future research

Should this conceptual model prove useful additional research is warranted. Initially addition of other vocal-based constructs should be studied from the RWHS environment. Automated classification methods in deception detection utilizing decision
support systems such as artificial neural networks and decision trees have shown promise and should be examined within the framework of RWHS audio analysis. A potential research question might be, “can decision support systems be used to automate audio-based cues for DD in real world high stakes environments?”

There also appears to be a research gap in the area of deception detection research namely the lack of studies looking at what amount of information is necessary to detect deception, in other words deception episode granularity. A possible research question might be, “can we increase accuracy without decreasing reliability by decreasing granularity of episodes?” Also is deception detection enhanced when episode granularity is at the question/response level vs the statement/interview level? Also the work of Buller and Burgoon [5] on IDT begs the question, “can audio-based indicators of deceit can be measured with repeated measures analysis without loss of significance?” Finally the development of a “stakes” construct to examine the impact real world and high stress have on deception detection would be very interesting. If such an construct could be developed a meta-analytic study could be performed examining previous deception detection research.

9. Implications & conclusions

Law enforcement personnel have a difficult task when it comes to sorting truths from lies. And their payoff ranges from the very small -- catching a kid stealing a toy, to the extreme -- stopping an act of terrorism. Technologies such as polygraphs, metal detection, and x-ray baggage screening have proven effective, but are static and not as adaptable as human find new ways to circumvent them. The suspect interview is a common tool used by law enforcement that could be enhanced with non-invasive, covert, real-time analysis methods. Before these tools can be developed we need a better understanding of deceptive cues in real life settings.

Field studies in deception detection have several substantial hurdles to overcome, namely a lack of randomization, difficulty acquiring large data samples, and the wide range of quality and format of data samples. But perhaps the most difficult hurdle to overcome in deception detection field studies is establishing ground truth. LE personnel often “know” someone is lying but that is not sufficient for academic rigor.

Although understanding deception and its detection are worthy goals, the problem addressed here is how best to analysis available cues. Our goal is to further the understanding of vocal-based cues to deception (text and speech) and their potential for automation. It is our hope that our research will reveal whether vocal and linguistic cues to deception can be effectively combined to improve deception detection accuracy. If proven to be successful at detecting deception, automating cue extraction and analysis in a covert, non-invasive manner and could have significant impact to operations involving RWHS interviews.

10. References


