Evaluating the Credibility Assessment Capability of Vocal Analysis Software

Aaron C. Elkins
University of Arizona
aelkins@cmi.arizona.edu

ABSTRACT

Vocal analysis software used for credibility assessment was investigated using a repeated measures deception experiment with 96 subjects. The vocal analysis software's built-in deception classifier performed at the chance level. When the vocal measurements were analyzed independent of the software's interface, the variables FMain (Stress), AVJ (Cognitive Effort), and SOS (Fear) significantly differentiated between truth and deception. The results of the present study suggest the claim that vocal analysis software measures stress, cognitive effort, or emotion cannot be completely dismissed.

1 INTRODUCTION

Remember the last time you called your insurance or credit card company? After navigating the maze of automated operators, you were greeted by a human voice and the words "This conversation may be recorded for quality assurance purposes". Most of us do not think twice about this seemingly innocuous statement; however, perhaps we would if we knew these recorded conversations are increasingly being subject to Vocal Risk Analysis (VRA). VRA is the process of evaluating the credibility of a person by analyzing their voice with specialized vocal analysis software.

The UK government recently invested an additional £1.5 million to expand their usage of VRA to assess and investigate claims made over the phone for housing and social security benefits [1, 2]. Based on a 20 minute conversation with an agent, a decision is made based on the results of the VRA to approve, deny, or investigate the claim further. Not strictly confined to phone calls, analysis of voice to detect deception is gaining wider adoption worldwide for rapid screening in airports and investigations by law enforcement. The Los Angeles county Sheriff's department is now using vocal analysis software to aid in criminal interrogations [3].

In all the rush to employ newer and better technology to combat fraud, terrorism, and crime, very few empirical attempts have been made to assess the validity of the vocal analysis software. The vocal analysis software claims to detect deception as well as levels of emotion, cognitive effort, and stress. These claims have been investigated in experimental and field settings and found the system was unable to detect deception above chance levels [4-6]. Still, the software vendors refute these findings by arguing the built-in algorithms only work in the real world where tension, stress, and consequences are high. To address this claim, this research intends to explore the vocal measurements independent of the software's interface and built-in algorithms to determine their validity and potential to predict emotion, cognitive effort, stress, and deception.

1.1 Vocal Deception Detection

Differences in acoustic vocal behavior exist between liars and truth tellers [7-11]. Vocal cues fall into three general categories, which include time (e.g., speech length, latency), frequency (e.g., pitch), and intensity (e.g., amplitude) [12]. Previous research demonstrated that relative to truth tellers, deceivers speak in shorter durations, with slower tempos, less fluency, and exhibit greater response latencies [8, 10, 13]. It has been postulated that deceivers, particularly during extemporaneous speech, are more reticent to provide extra details and require more cognitive effort to fabricate their responses [10, 14]. An increase in pitch or frequency has also been associated with arousal during deceptive responses which presumably results from the anxiety of being caught and facing negative consequences [7, 8, 11, 15].

1.1.1 Previous Generation of Vocal Stress Analysis Software

The generation of software for analyzing voice to detect deception preceding the vocal analysis

software under investigation is called Vocal Stress Analysis (VSA) and has consistently failed to reliably detect deception in experimental or field settings [4, 6]. Despite the richness of features present in the voice, previous VSA systems focused on a very small frequency band of 8-12Hz [6]. This is because the human body exhibits periodic contractions of the muscles known as microtremors on this narrow and low frequency range [16, 17]. VSA systems attempted, unsuccessfully, to measure this frequency produced by the larynx muscles.

VSA systems assumed that a reduction in the power of the microtremor frequency implies deception because it is caused by a stress induced drop in blood pressure. The microtremors do occur at the low frequency range; however, existing recording technologies may not have the sensitivity required to accurately measure and subsequently calculate this low frequency. Additionally, even if microtremors can be measured via the voice, the relationship between lower blood pressure and deception is tenuous.

1.1.2 Full Spectrum Vocal Analysis

The focus of this study is on modern vocal analysis software that use the full spectrum of the vocal information contained in the voice. In addition to measuring frequency and intensity, modern vocal analysis software measure indicators of cognitive effort through speech disfluencies or plateaus. The vocal analysis software looks for variation, length, and total micro-momentary drops in amplitude during speech. When examining the vocal waveform these appear as plateaus and reflect speech interrupted by additional thoughts or cognitive load.

Not only does the modern vocal analysis software differ from VSA by using the full vocal spectrum and including measurements of cognitive effort, but it also measures frequency using thorns, which represent peaks or valleys of amplitude in the vocal waveform. The measurements provided by the vocal analysis software will explained in more detail in the subsequent vocal measurements section.

1.2 Deception Experiment

This study is investigating the validity and deception detection ability of vocal analysis software using audio recordings from a deception experiment. The experiment consisted of an interview that required participants to alternate between deceptive and truthful responses.

The focus of the experiment was to identify systematic patterns of vocal behavior that vary as a function of truth or deception. Using the measurements provided by the vocal analysis software the following hypotheses were specified.

H1: There is a difference in vocal measures between liars and truth tellers.

H2: Liars will exhibit higher vocal measurements of cognitive effort than truth tellers.

H3: Liars will exhibit shorter message length.

H1 is an exploratory hypothesis because many of the system provided vocal measurements are dissimilar from those used in previous nonverbal research and any unexpected significant findings will be corrected to reflect the experiment wise error of testing 13 simultaneous vocal measurements. At the α=.05 level this corresponds to 48.7% chance of Type-I error [18].

H2 is based on previous deception research findings that lying is more cognitively demanding than telling the truth [7]. It is very difficult to recreate a sufficiently perilous situation or conditions to induce negative stress or arousal. However, the extra cognitive effort required to fabricate lies should exist in both experimental and real world settings [19].

H3 is also based on previous research that found deceivers exhibited shorter response lengths, talking time, and lengths of interactions [7, 13]. The reduced response time is explained as a deceptive individual's reticence to provide more information than necessary [10].

2 METHOD

2.1 Participants

220 international participants were recruited from a southwestern university for a study on cross-culture interviewing behavior. Because of differences in recording and malfunctioning equipment or poor audio quality (vocal responses below the noise floor), only 96 of the original 220 participant were included in this study of which 53 were male and 43 female. The mean age was 26.1 (S.D. = 11.2) with a range of 18 to 77 years. The ethnicity breakdown of participants was: 53% White, 28% Asian, 8% African American, 7% Hispanic, and 3% Other.

Upon arrival, participants completed a pre-survey that measured their pre-interaction goals and demographics. Participants were instructed to complete a 13-question interview during which they should, for designated truthful questions, give answers that are "the truth, the whole truth, and nothing but the truth" and on remaining questions

give answers that depart substantially from this standard. A teleprompter that was not visible to their professional interviewer notified them for each question whether they were to "tell the truth" or "not tell the truth."

Participants were awarded $15 for their participation and offered an extra $10 if deemed credible by their professional interviewer. The professional interviewers were blind to the experimental manipulation.

Participants were randomly assigned to one of two lie and truth sequences. There were 47 participants in the first sequence and 50 in the second. The sequences varied the questions to which they were instructed to lie or tell the truth. (D is Deception and T is Truth)

SEQUENCE ONE: DT DDTT TD TTDD T
SEQUENCE TWO: DT TTDD TD DDTT T

Both sequences had participants lie and tell the truth concurrently on questions 1, 2, 7, 8, and 13. The primary focus of this study was on the eight questions that contained variation both between and within question that was attributable to lying or telling the truth. The eight questions are listed in Table 1 and were intended to be answered quickly to facilitate comparability with the vocal analysis software. The questions were designed to be either neutral or charged. Charged questions were expected to evoke a greater emotional or stressful reaction than neutral questions.

Following the interview, participants completed a post-survey that measured their arousal, cognitive difficulty, attempted behavioral control, self-construals, culture, and social skills.

Table 1
Short answer questions from the experiment

Question
1. Where were you born? (N)
2. Did you ever take anything from a place where you worked? (C)
3. Did you bring any keys with you today? (C)
4. If I asked you to empty your wallet purse or backpack would anything in it embarrass you? (C)
5. What city did you live in when you were 12 years old? (N)
6. Did you ever do anything you didn't want your parents to know about? (C)
7. Name the country stamped most often in your passport? (N)
8. Did you ever tell a lie to make yourself look good? (C)

Note. C is a charged question and N neutral questions.

2.2 Instruments

2.2.1 Vocal Analysis Software

A commercial and full spectrum vocal analysis software package currently being used by law enforcement, government, and private industry was used to analyze the short answer questions [20]. Of the 13 short answer questions there were 1,181 valid vocal responses. The mean response length of each vocal measurement was .47 seconds (S.D. = .4) and consisted of primarily one word responses (e.g., "Yes", "No").

2.2.2 Vocal Segmentation Procedure

96 participant audio files were analyzed with the vocal analysis software. Prior to analysis, the audio files were required to be converted to 11.025 KHz sampling rate, 8 bits, and mono channel. The converted audio files were segmented using the Offline mode of the software, which means post-processing of audio files instead of live real-time analysis with a microphone. The segmenting process involves listening to each audio file and marking the portions that are noise, the participant, and relevant. For each of the segmented audio files, the 13 question responses were marked as relevant. The vocal analysis software generated vocal measurements for each segment marked relevant.

2.3 Vocal Measurements

The vocal analysis software provides measurements intended to reflect deception, emotion, cognitive effort, and stress. The variables generated for each of the 13 questions are listed and described in Table 2 based on the software documentation. It is important to note that there is no theoretical support for the descriptions provided by the vendor

Table 2
Vocal measurement descriptions

Variable	Description
SPT	Emotional level – Number of thorns
SPJ	Cognitive Level – Average number of Plateaus
JQ	Stress Level – Standard error of Plateau length
AVJ	Thinking Level – Average Plateau length
SOS	Indication of fear or unwillingness
FJQ	Imagination – Uniformity of low frequency
FMAIN	Stress Level – Most significant frequency
FX	Level of Concentration – Frequencies above FMAIN
FQ	Deception – Uniformity of frequency
FFLIC	Embarrassment or conflicting thoughts - Harmonics
ANTIC	Anticipation
SUBCOG	Subconscious cognition
SUBEMO	Subconscious emotion

While most of the variables involve measurements of frequency calculated using traditional Fourier Transforms, SPT, SPJ, JQ, AVJ are not. The SPT measurement is the average number of thorns per sample. Thorns are defined as three successive amplitude measurements following the pattern either high, low, high, or low, high, low. Figure 1 below illustrates three thorns graphically in a .002 second portion of audio, which corresponds to 24 samples at an 11.025 KHz sampling rate.

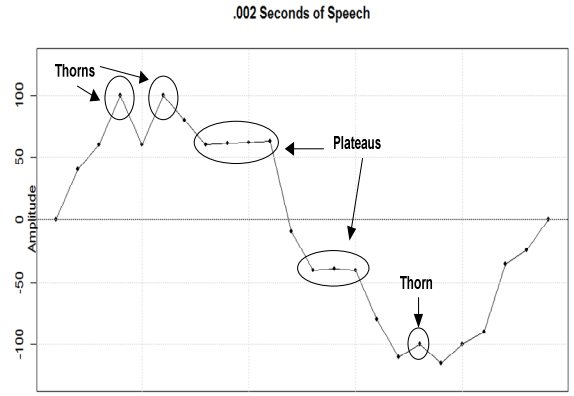

Figure 1: Thorns and plateaus in voice segment

The SPJ, AVJ, and JQ measurements are based on plateaus. Plateaus are defined as a local flatness of amplitude containing consecutive samples less than a threshold. Two plateaus can be seen graphically in Figure 1. AVJ measures the average length of the plateaus, which is intended to reflect speech interrupted by cognitive effort. SPJ measures the average number of plateaus and JQ the standard error or variation of plateau length.

Eriksson and Lacerda contend that the thorns and plateaus identified by the vocal analysis software may be artifacts that occur when the audio is converted from analog to digital [21].

2.3.1 Standardization

All of the reported and analyzed vocal measurements were converted to their corresponding z-scores for ease of interpretability and comparison.

3 DECEPTION EXPERIMENT

3.1 Results of Vocal Analysis Software Built-in Classifier

3.1.1 Lie and Truth Detection

For each processed audio segment, the software provides a probability of deception. Using these predicted probabilities, the system had an overall accuracy of 52.8% for detecting either truth or deception and an area under the curve (AUC) of .50. Based on Signal Detection Theory, AUC reflects the tradeoff of the true positive rate (TPR) and false positive rate (FPR) [22]. An AUC score of .50 can be interpreted as a 50% probability that the system will find a liar more deceptive than a truthful person. The software's detection accuracy was at the chance level. There was no significant difference in the software predicted lie probability between liars or truth tellers, $F(1,735) = .59$, $p = .44$.

The Receiver Operator Characteristic (ROC) curve in Figure 2 provides more detail on the software's deception detection performance. This curve displays the continuous relationship between TPR and FPR as the classifier decreases the cutoff for a deceptive classification. An optimal classifier would have a line reaching the top left corner, which corresponds to 100% TPR and 0% FPR. The grey diagonal line represents prediction at the chance level. The software's deception detection performs best with a conservative probability cutoff, which results in a 26% TPR vs. 19% FPR. Depending on the scenario, a higher TPR at the expense of FPR may be acceptable; however, the software performed close or worse than chance on the remainder of the curve.

Per question, the vocal analysis software had an overall accuracy ranging from 48.86%-57.89% and an AUC ranging from .46-.59. The software performed best on the question "Did you ever take anything from a place where you worked?" where it had a 62% TPR vs. 36% FPR at the more conservative side of the curve. This charged question may have caught the participants off guard and resulted in increased stress or negative arousal, which the system is intended to measure.

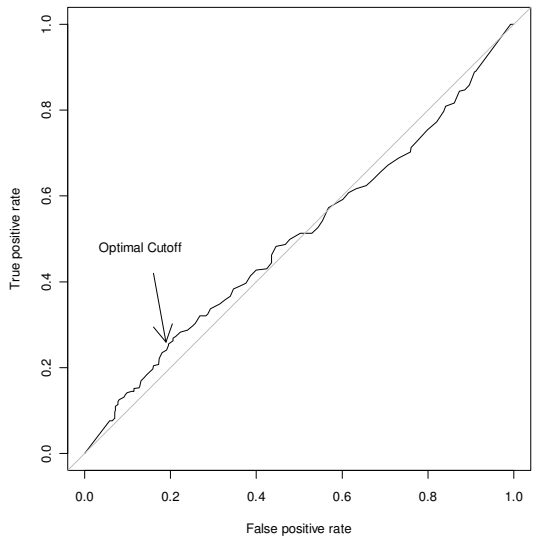

Figure 2: ROC curve of vocal analysis software deception detection

The poor lie detection results using the built-in algorithms are congruent with previous research utilizing the vocal analysis software [4]. The vendor of the vocal analysis software contends that the built-in algorithms are tuned for real world conditions which involve jeopardy or consequences and are not replicated in an experimental environment [21].

In addition to a probability of deception, the vocal analysis software provides a classification of the emotion, stress, or truthfulness for each response. [23]. The Excited classification provided the best discrimination between liars and truth tellers and had a standardized residual of -1.68; however there was no significant relationship between deceptive communication and the software's classifications, χ^2 (9,N=730)=11.51, p=.24.

3.1.2 Question Type

In order to fully explore the relationship between the classification and emotion or stress the classifications were compared against charged and neutral question types. Questions designated as charged were designed to evoke an emotional or stressful response from the participants. There was a highly significant relationship between the software's classification and charged or neutral questions, $\chi 2$ (9,N=730)=58.94, p<.001. In the Stressed category, 70% of the responses were from neutral questions. Contrastingly, 87.5% of the responses classified as Excited were from charged questions. If the classifications are valid, this may mean charged questions caused excitement, but not stress to the participants. Similarly, the system categorized neutral questions less often as Truth, likely because it found the responses stressful.

3.1 Analysis of Vocal Measurements

In order to test the experimental hypotheses on a sample containing unbalanced observations, a multilevel regression model was used in place of traditional ANOVA [24, 25]. The unbalanced observations resulted from responses that the vocal analysis software was unable to process. With longer interviews the likelihood of missing at least one time point or response is high, particularly with short physiological measurements [26]. A listwise case deletion to attain a balanced dataset would have resulted in the loss of 27 cases and 153 observations.

A multilevel model was specified for each vocal measurement (N=737) as the response variable, a dummy coded Truth variable (1 = Truth, 0 = Lie) as a fixed effect parameter, and varying intercepts for random Subject (N=96) and Question (N=8) effects. This model adjusts the standard errors to reflect the uncertainty that arises from variation within subject and question.

To test the H1 and H2 hypotheses, the specified models were compared to the unconditional models, which omit any fixed effect of lying or telling the truth. To test if the Truth condition provides a significant improvement to the fit of the data, the models were compared using deviance-based hypothesis tests. Deviance reflects the improvement of log-likelihood between a constrained model and a fully saturated model [24].

3.1.1 Results of Experimental Treatment

Table 3 reports the results of the deviance hypothesis tests for each vocal measurement. The test statistic for a significant (α=.05) difference between the unconditional and specified model is χ^2 (1,N=737) >3.84. The χ^2 statistic is calculated by subtracting the deviance of the specified model from the unconditional model.

H1: There is a difference on vocal measures between liars and truth tellers.

The H1 hypothesis was supported by the finding of a significant χ^2 for JQ, AVJ, FFlic, and FMain. FMain and FFlic were unexpected and after a Bonferroni correction (.05/13=.0038) only FMain remained significant. FMain is documented as being the numerical value of the most significant frequency in the vocal spectrum. Previous research has found increased pitch or frequency to be associated with deception [27, 28].

Table 3
Results of Deviance-Based Hypothesis Tests on Vocal Measurements (=737, 96 Subject, 8 Questions)

	d.f.	χ^2	p
SPT	1	3.23	0.07
SPJ	1	0.32	0.57
JQ	1	5.15*	0.02
AVJ	1	4.91*	0.03
SOS	1	2.65	0.10
FJQ	1	0.03	0.85
FMAIN	1	10.99*	<.001
FX	1	1.57	0.21
FQ	1	0.73	0.39
FFLIC	1	4.18*	0.04
ANTIC	1	0.03	0.87
SUBCOG	1	0.80	0.07
SUBEMO	1	0.23	0.63

The FMain results can be qualified by examining Table 4 where fixed effect coefficients are listed for each significant vocal measurement. FMain is negatively related to telling the truth in our sample data. This means that on average, across all questions in the interaction, participants telling lies had FMain values greater than participants telling the truth.

H2: Liars will exhibit higher vocal measurements of cognitive effort than truth tellers.

The H2 hypothesis was supported by finding a significant χ^2 for JQ and AVJ shown in Table 3 in addition to significant negative coefficients for the Truth condition found in Table 4. This suggests that participants in our sample had higher average AVJ and JQ scores when lying than when telling the truth.

AVJ and JQ appear to be capturing speech interruptions or disfluencies (hesitations, pauses, responses latency) that prior research has found to be associated with high cognitive load [19, 29, 30].

The random effects in Table 4 display a high degree of variability within subjects across all of the vocal measures, particularly AVJ. This likely explains why the standard error of the intercepts was so high. Until this within subject variability is accounted for, predicting deception through vocal behavior will be imprecise.

H3: Liars will exhibit shorter message length.

Table 4
Results of fitting multilevel models for predicting FMain, AVJ, and JQ (N=737, 96 Subject, 8 Questions)

	AVJ	JQ	FMain
Fixed Effects			
Intercept	0.05	0.04	0.11
	(0.09)	(0.15)	(0.08)
Truth	-0.13*	-0.13*	-0.22*
	(0.06)	(0.06)	(0.07)
Random Effects - Variance Components			
Within-Subject	0.35	0.29	0.18
Within-Question	0.02	0.15	0.01
Residual	0.63	0.57	0.79

Note. Significant coefficients (b < 2 SE) are denoted by *; models were fit by maximum likelihood estimate.

The H3 hypothesis was discredited. There was no significant difference in response length between liars and truth tellers, $F(1,734)=2.47$, $p>.05$. This could be attributed to the short response interview format that did not facilitate enough variation to find a significant effect (Response Length M= .55 sec, SD = .45). However, there was a significant difference between the responses to charged or neutral questions, $F(1,734)=189.48$, $p<.001$. Responses to charged questions were an average of -.32 seconds or 57% shorter than responses to neutral questions.

While the act of lying did not result in any reluctance to give longer responses, charged questions such as, “Did you ever do anything you didn't want your parents to know about?” did.

Figure 3 illustrates the negative relationship to charged questions. This implies that lying alone is not enough; one needs to be fearful of the repercussions of a wrong answer, which accompanies deception in more interactive contexts.

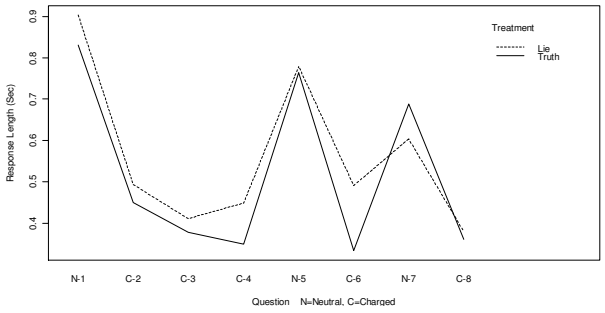

Figure 3: Interaction of question and truth on response length

3.1.1 Moderators of Lying on Vocal Measurements

3.1.2 Question Effect

JQ demonstrated relatively high levels of within question variability with a variance of .15 compared to .02 and .01 for AVJ and FMain respectively. Examining the estimated random effect intercepts for each question reveals the pattern illustrated in the bottom of Figure 4. Charged questions such as "Did you ever tell a lie to make yourself look good?" were on average 33% lower than JQ scores for neutral questions such as, "Where were you born?". More charged questions result in less vocal interruption or disfluency variation than neutral questions. The JQ pattern mirrors the response length relationship for each question and in fact JQ and response length are highly correlated, r(735)=.82, p<.001).

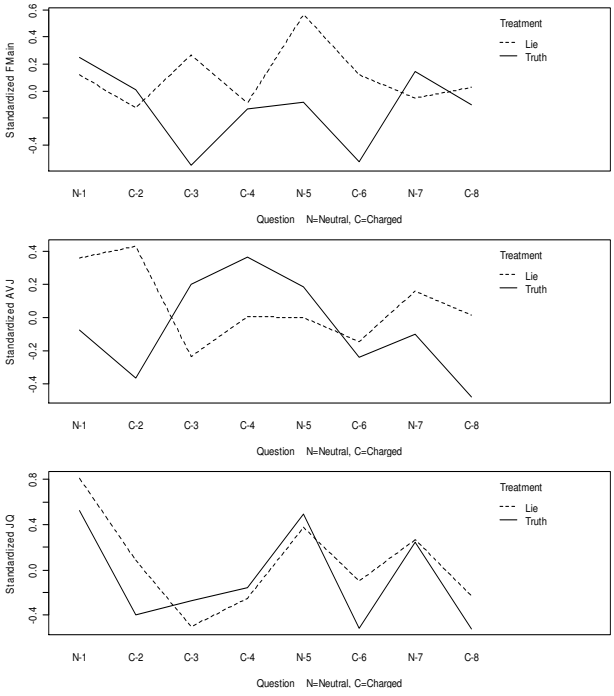

Figure 4: Interaction of question and truth treatment on FMain, AVJ, and JQ

Figure 4 illustrates the interaction between the question, question type (charged or neutral), and experimental treatment on FMain, AVJ, and JQ. While JQ does not appear to provide a clear separation between liars and truth tellers, it does move predictably negative for charged questions and positive for neutral questions. A multilevel model regressing JQ on Truth, Question Type, and the interaction between Truth and Charged Question was specified with subject as a random effect. There difference in JQ levels between question types was significant, F(1,734)=171.8, p<.001.

The vendor of the vocal analysis software refers to higher levels of JQ as corresponding to increasing levels of stress. This coincides with the disproportionate amount of neutral questions categorized as Stress by the systems' built-in classifier. However, the finding of neutral questions as more stressful is curious; perhaps, in the case of shorter responses to charged questions, less variation in vocal disfluencies is actually indicative of stress.

There was a significant interaction, F(1,734)=4.64, p<.05, between lying and charged question on SOS. The variable SOS, or "Say or stop" is defined as an indication of fear or unwillingness to discuss. Figure 5 illustrates the interaction. Only during charged questions does SOS provide separation between liars and truth tellers. Both liars and truth telling participants had similar SOS scores for neutral questions; however, charged questions resulted in higher SOS values for liars. The main effect, F(1,734)=33.89, p<.05, of lower SOS values for charged questions seems to contradict that SOS measures fear, unless the only real fear as registered by SOS, occurred when participants lied to charged questions.

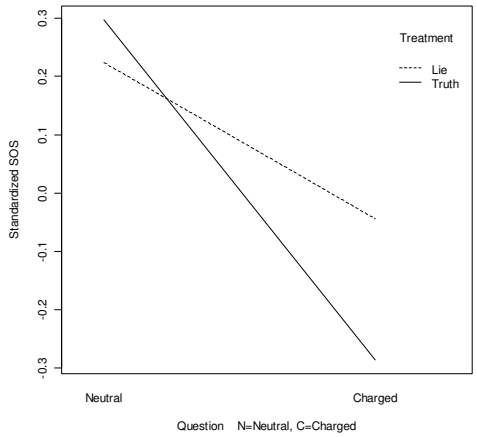

Figure 5: Interaction of charged question and truth on SOS

4 DISCUSSION

Mirroring the results of previous studies, the vocal analysis software's built-in deception classifier performed at the chance level [6]. However, when the vocal measurements were analyzed independent of the software's interface, the variables FMain, AVJ, and SOS significantly differentiated between truth and deception. This suggests that liars exhibit higher pitch, require more cognitive effort, and during

charged questions exhibit more fear or unwillingness to respond than truth tellers.

Previous research has found measurements similar to FMain or the fundamental frequency to be predictive of deception or stress [10]. However, the measurement of AVJ which is based on average plateau length is novel and may reflect cognitive effort through micro-momentary speech interruptions. Future research should further investigate this measurement and its diagnostic potential to detect cognitive effort or thinking.

The measurement of JQ, which is described as reflecting stress level, was highly predictive of charged questions designed to evoke stressful or emotional responses from participants. FMain was a highly significant discriminator of deception and was partially explained. Consistent with prior research, it appears stress may cause higher frequency or elevated pitch [10].

The results of the present study also suggest the claim that vocal analysis software measures stress, cognitive effort, or emotion cannot be completely dismissed.

4.1 Future Vocal Deception Research

In order to improve the predictive power of future models incorporating vocal measurements, covariates should be investigated and included to account for the within subject variance. Additionally, the vocal analysis software should be evaluated and compared across different modalities and environments to determine measurement invariance and robustness (e.g., telephone, noisy room, etc.).

Future deception research should focus more on vocal behavior over the entire interaction. While some of the deception predictions using vocal measurements performed better than chance, there is still much unaccounted variability in vocal behavior. Interpersonal deception theory (IDT) predicts that deceptive behavior is dynamic and varies as a function of sender, receiver, time, deception, suspicion, motivation, and social skills [31]. However, most deception experiments and even the polygraph exam focus on behavior difference scores over a set of questions [19]. Using this design ignores all of the important contextual information.

It may be more appropriate to think of deceptive behavior as constantly changing over time in either a negative or positive direction in response to environmental stimulus. Multilevel regression and latent growth curves using structural equation models can be used to model this behavioral change over time [24, 26, 32]. However, deception experimental designs would need to be reoriented from prediction of difference scores to rates of change. Regardless of modeling approach, unless the entire interaction is accounted for, we will have to be satisfied with deception prediction models that are in one instance remarkably accurate and in another, remarkably inaccurate depending on the person, time, place, or context.

5 REFERENCES

[1] *Whose pants on fire?* The Economist, City, 2008 May 8.

[2] Walker, J. *Phone lie detector led to 160 Birmingham benefit cheat investigations*. City, 2008 May 8.

[3] Holguin, R. *L.A. Co. gets cutting edge lie detector*. KABC, City, 2008 Dec 12.

[4] Damphousse, K., Pointon, L., Upchurch, D. and Moore, R. *Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting*. 2007.

[5] Gamer, M., Rill, H. G., Vossel, G. and Gödert, H. W. Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology*, 60, 1 2006), 76–87.

[6] Haddad, D., Walter, S., Ratley, R. and Smith, M. *Investigation and evaluation of voice stress analysis technology*. AIR FORCE RESEARCH LAB ROME NY INFORMATION DIRECTORATE, City, 2001.

[7] DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K. and Cooper, H. Cues to deception. *Psychological Bulletin*, 129, 1 2003), 74-118.

[8] DePaulo, B. M., Stone, J. I. I. and Lassiter, G. D. I. Deceiving and detecting deceit. *The self and social life*1985), 323.

[9] deTurck, M. and Miller, G. Isolating the Behavioral Correlates of Deception. *Human Communication Research*, 12, 2 1985), 181–201.

[10] Rockwell, P., Buller, D. and Burgoon, J. The voice of deceit: Refining and expanding vocal cues to deception. *Communication Research Reports*, 14, 4 1997), 451-459.

[11] Zuckerman, M., DePaulo, B. and Rosenthal, R. Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, 14, 1 1981), 59.

[12] Scherer, K. R. *Methods of research on vocal communication: paradigms and parameters*. New York: Cambridge University Press, City, 1985.

[13] deTurck, M. and Miller, G. Deception and arousal. *Human Communication Research*, 12, 2 2006), 181-201.

[14] Vrij, A. *Detecting lies and deceit: Pitfalls and opportunities*. Wiley-Interscience, 2008.

[15] Apple, W., Streeter, L. A. and Krauss, R. M. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37, 5 1979), 715–727.

[16] Lippold, O. Physiological tremor. *Scientific American*, 224, 3 1971), 65–73.

[17] Lippold, O. C. J., Redfearn, J. W. T. and Vučo, J. The rhythmical activity of groups of motor units in the voluntary contraction of muscle. *The Journal of physiology*, 137, 3 1957), 473.

[18] Rice, W. Analyzing tables of statistical tests. *Evolution*, 43, 1 1989), 223-225.

[19] Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, R. and Bull, R. Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, 32, 3 2008), 253-265.

[20] Nemesysco Ltd. *Layered Voice Analysis (LVA) 6.50*. Netania, Israel, 2009.

[21] Eriksson, A. and Lacerda, F. Charlatanry in forensic speech science: a problem to be taken seriously. *International Journal of Speech, Language and the Law*, 14, 2 2007), 169-193.

[22] Green, D. and Swets, J. Signal detection theory and psychophysics1966).

[23] Cleveland, W. *Visualizing data*. Hobart Press, 1993.

[24] Singer, J. and Willett, J. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA, 2003.

[25] Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press New York, 2007.

[26] Moskowitz, D. and Hershberger, S. *Modeling intraindividual variability with repeated measures data: Methods and applications*. Lawrence Erlbaum Associates, 2002.

[27] Hocking, J. and Leathers, D. Nonverbal indicators of deception: A new theoretical perspective. *Communication Monographs*, 47, 2 1980), 119-131.

[28] Apple, W., Streeter, L. and Krauss, R. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37, 5 1979), 715-727.

[29] Smith, V. and Clark, H. On the course of answering questions. *Journal of Memory and Language*, 321993), 25-25.

[30] Goldman-Eisler, F. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press New York, 1968.

[31] Buller, D. and Burgoon, J. Interpersonal deception theory. *Communication Theory*, 61996), 203-242.

[32] Fitzmaurice, G., Laird, N. and Ware, J. *Applied longitudinal analysis*. Wiley-IEEE, 2004.