Spring 2014

Vocal fundamental frequency as a function of deception and stress

Joshua David Gross
James Madison University

Vocal Fundamental Frequency as a Function of Deception and Stress

---

A Project Presented to

the Faculty of the Undergraduate

College of Health and Behavioral Studies

James Madison University

---

in Partial Fulfillment of the Requirements

for the Degree of Bachelor of Science

---

by Joshua David Gross

May 2014

Accepted by the faculty of the Department of Psychology, James Madison University, in partial fulfillment of the requirements for the Degree of Bachelor of Science.

FACULTY COMMITTEE:

____
Project Advisor: Michael D. Hall, Ph.D.
Professor, Psychology

____
Reader: Jeffrey T. Andre, Ph.D.
Professor, Psychology

____
Reader: Vivien (Kit Ying) Chan, Ph.D.
Assistant Professor, Psychology

HONORS PROGRAM APPROVAL:

____
Barry Falk, Ph.D.,
Director, Honors Program

Table of Contents

Acknowledgements

I would like express sincere appreciation to my thesis chairman and advisor, Dr. Michael Hall, for his unwavering guidance and mentorship. I would like to extend additional appreciation to my committee members, Dr. Jeffrey Andre and Dr. Vivian (Kit Yang) Chan, for their valuable comments and recommendations. Lastly, I would like to thank the James Madison University Honors Program, Department of Psychology, and College of Health and Behavioral Studies for supporting the project.

Abstract

The current investigation sought to address several important issues concerning frequency-based acoustic measures of vocal deception, including 1) to provide a direct comparison of aggregate and local measures, 2) to dissociate effects of deception from stress, and 3) to establish an approach that reduces potential adverse effects of stress. Participants were randomly assigned to conditions of truthful or deceptive intent in which they were recorded while answering a series of questions about their academic records. This was done both with and without the presence of a mild stressor. Acoustic measurements then were obtained from the recordings, including aggregate fundamental frequency (F_0), F_0 variability (% jitter), and local changes in F_0 following the beginning of potentially deceptive content. There was an observed effect of stress on aggregate F_0, as well as a corresponding marginal effect for local F_0 changes. Additionally, a marginal effect of deception on local F_0 changes was obtained. Across all observed effects, measured values decreased under conditions of stress/deception, suggesting attempted control of these parameters by participants. The fact that only local measures indicated deception (with much greater observed effect size relative to corresponding aggregate measures) further suggests that local measures are likely to provide a more reliable means by which to detect deception.

Introduction

The field of deception detection has traditionally relied on the presumption that psychogenic stress invariably induces an autonomic physiological response in the deceiver that can be directly measured in terms of blood pressure, skin conductance, heart rate variability, and muscular tension (Polygraph and Lie Detection, 2003). Polygraph testing, the most prominent field technology in past decades, has utilized these physiological by-products of stress as the primary means of detecting deception. However, the data to support the polygraph as a sensitive and reliable measure of deception have generally been negative. In a committee review by the United States National Research Council, it was determined that the scientific evidence underlying polygraph technology was "scanty and weak," and that the "physiological responses measured by the polygraph are not uniquely related to deception" (Polygraph and Lie Detection, 2003).

Polygraph evaluation is contingent solely on the circumstances established by the examiner, in that there is no direct method to verify that the physiological responses elicited by the deceiver reflect the situation the examiner has attempted to generate (Saxe, 1994). One example of this dependency upon the examiner is provided by the most commonly used system for deception detection assessment, the Guilty Knowledge Test (GKT). GKT is a multiple choice test constructed by the examiner consisting of a series of neutral (control) questions and one relevant question within one multiple choice item. The relevant question is intended by design of the examiner to be indistinguishable from control questions and foster a differential response from control questions, indicating knowledge of the incident in question (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003). However, these differential responses may be a function of stress generated by the examiner and not exclusively deception. Likewise, additional paradigms for assessing deception in academic fields are generally accompanied by stress induction imposed on participants. For example, the mock-crime scenario involves researchers instructing participants to smuggle what are suggested to be classified documents out of a federal building and attempt to deceive guards at checkpoint interviews where polygraph and voice analysis evaluations are made. Unfortunately, such paradigms tend to generate increased levels of adverse stress for participants (Hall & Watts, 2006).

In the hope of establishing a more reliable means of detecting deception, recent attempts have been made to develop technologies that focus on analyzing vocal and acoustic parameters. The Multifactor Model (Zuckerman et al., 1981) proposes that such parameters may be modulated in a deceptive individual by any combination of three theoretical frameworks, namely Arousal Theory, Cognitive Theory, and Attempted Control Theory. Arousal Theory presumes that deceptive individuals generate increased stress responses and arousal relative to truthful individuals and experience congruent emotional states such as fear, guilt, and "duping delight" (Eckman, 2001). Cognitive Theory instead posits that the act of being deceptive exhibits larger cognitive demands on the individual via the necessity of deceivers to suppress the truth, which is autonomously activated. Lastly, Attempted Control Theory suggests that deceptive individuals actively suppress or control behavioral signatures that may be stereotypically associated with deception in order to appear sincere, natural, and truthful. Although various types of stress, cognitive load, and attempted behavioral control have all empirically been shown to modulate speech (Beckford Wassink et al., 2006; Köster, 2001), discrete and definable patterns of association have not been established. However, self-report research assessing the relative contributions of these theoretical frameworks have found that each process is experienced in the individual, with the Attempted Control process being the most dominant (Caso et al., 2005).

Despite these findings, prominent speech analysis technologies have traditionally assumed the Arousal Theory to be the underlying contributor to deception.

Voice-Stress Analysis (VSA) and Layered Voice-Stress Analysis (LVA) have been among the primary voice-related techniques utilized in both research and applied settings to evaluate deception. The proposed rationale for detecting deception in these techniques is based upon the same fundamental assumption held by the polygraph test, that the cognitive process of generating deceit will invariably induce stress in the deceiver. VSA claims to detect deception via micro-muscle tremors (MMT) in the auditory tract that manipulate speech signals following the induction of stress. According to the theory, acoustic signals are modulated after the physiological stress response constricts the vocal folds and narrows the vocal tract (tract filter). VSA claims to detect deception by assessing these MMTs directly, and therefore stress. In instances of deception, it was determined that VSA was not detecting MMTs but instead, processing energy changes between 20 and 40 Hz within the spectral envelope (Hopkins, Benincasa, Ratley, & Grieco, 2005). However, these energy increases generally fall beneath the F_0 range for human speech. Furthermore, these energy increases do not necessarily indicate a change in the vibration of the vocal folds and could be due to extraneous acoustic factors, such as noise. Overall, data produced by VSA devices have traditionally yielded large error rates in identifying deceptive speech and intention (Everding, 2004). Similarly, studies that have been conducted to assess the ability of LVA technology to detect deception have consistently concluded that LVA functioned at around chance levels in detecting truth and deception in high and low vocal stress states (e.g., see Damphouse, Pointon, Upchurch, & Moore, 2007).

Given the problems with these various technologies attempting to link stress and deception, one may suggest that perhaps such a direct link cannot be presupposed. Although it will be the case that some deceptive individuals will elicit stereotypical signs of nervousness and stress, it is plausible that truthful individuals may elicit similar tension and anxiety when subjected to an interrogative paradigm or attached to a lie detection device. Therefore, it is essential that stress be dissociated when assessing the vocal components of deception (Kirchhübel & Howard, 2013).

Numerous studies have focused on assessing potential temporal features of deception such as speaking rate, pauses, hesitations and speech errors. Although the assessment of these parameters have been somewhat successful in detecting systematic differences between deception and truthfulness, results across studies remain mixed (Benus et al., 2006; Feeley and deTurck, 1998; Stroemwall et al., 2006). However, frequency-based measures have most consistently produced predicative associations with deception, and thus will provide the focus of the current investigation. Particularly, the literature has generally supported the claim that fundamental frequency (F_0) increases as a function of deception and the inherent stress underlying its auditory signal (Krauss, Geller, Olsen, & Apple, 1977; Villar, Arciuli, & Paterson, 2013; Protopapas & Lieberman, 1997; Anolli & Ciceri, 1997). F_0 can be defined as the lowest frequency (Hz) value represented in a complex periodic waveform. All subsequent frequencies within the waveform are integer multiples of F_0 and are therefore, harmonically related.

Prior research has focused almost exclusively on aggregate frequency-based measures such as mean, range, and median across an entire utterance or response (aggregate) and has implicated F_0 as the most salient acoustic parameter of deceptive speech analysis. However, aggregate F_0 measure analysis may not always accurately assess deception due to its integration of primarily irrelevant propositional content (noise). More recent research has instead suggested that the analysis of localized portions of a deceptive utterance may reflect the most salient propositional content within deceptive utterances. These localized regions have been termed critical segments, and they represent the portions of a deceptive utterance that are directly related in meaning to the topic of interest posed by the inquirer (e.g. "My current GPA at JMU is *a 4.0*). Presumably, critical segments within deceptive responses will possess more emotional salience and potentially present stronger prosodic and acoustic cues relative to truthful utterances (e.g., see Enos, Shriberg, Graciarena, Hirschberg, & Stolcke, 2007). There are some initial indications that F_0 measures within critical segments can provide reliable indications of deception. For example, Hall & Watts (2006) found predictive localized F_0 changes in responses from participants asked to be deceptive about their academic records.

Current Investigation

The current investigation was motivated by an interest to address several primary issues of importance to assessing the reliability of F_0 measures in predicting vocal deception. First, we sought to compare aggregate and critical segment measures on the same task. Evaluation under similar conditions was necessary in order to assess if critical segment analysis offers a more reliable method than traditional aggregate measures in detecting F_0 differences between deceptive and truthful individuals. A measure of aggregate F_0 variability (jitter) also was included.

We also evaluated whether there are contributions of deception to F_0 over and above the observed effects of stress. This was accomplished by recording truthful or deceptive utterances from college students about their academic records both with and without the presence of an additional stressor. Such an orthogonal design was necessary to determine whether there are direct contributions from stress to the acoustic signal.

Lastly, the current study seeks to extend on the previous work conducted by Hall & Watts (2006) by using a similar paradigm to analyze F_0 at critical segments in deceptive utterances. Stress induction was implemented in order to potentially present a novel means by which to assess stress independently of deception. In contrast to some of aforementioned paradigms, a relatively mild stressor was used in an attempt to reduce any aversive psychological impact on the participants. Such considerations are important in continuing research on detecting deception in an ethical and responsible fashion.

Methods

Participants

Twenty-six James Madison University students participated in the studies' initial collection of vocal samples regarding their academic records. Nine participants were male and 17 participants were females. All participants were enrolled in an introductory psychology course at the university and received credit for their participation. Additionally, it was required that all participants were native speakers of the English language and self-reported that they were free of hearing deficiencies. Participants were randomly assigned to either respond with truthful statements or deliberately intend to be deceptive regarding questions about their academic record. Random assignment was based upon the order in which they arrived at the lab to participate, (e.g. odd numbers were assigned to the truthful condition and even numbers were assigned to the deceptive condition).

Stimuli and Materials

Students were asked a series of questions (or otherwise were directed to provide requested information) regarding their academic records and transcript information, as these documents provided known veracity (refer to Appendix A for all items). All question stimuli were recorded in Ableton Live 8 and were downsampled to a sampling rate of 44.1 kHz (16-bit) and were presented at a peak intensity of 80 dB. E-Prime was used to present the stimuli to participants, and to control advancing to the next trial. Stimuli were presented over circumaural Sennheiser HD 25-SP II earphones.

The device emWave was utilized within the stress condition and was preset to elicit a persistent red light for the full duration of the stress condition. emWave is a biofeedback device intended for the general public that illuminates color as a function of physiological measures of stress such as heart rate variability. Color illumination is generated by detecting physiological variation conducted through the users index finger by a sensor pad located on the device. Participants placed their finger over a sensor pad for the full duration of the stress condition.

Procedure

After providing informed consent, participants entered a sound-attenuating chamber where they received instructions about the task. Each participant was told they would hear a series of sentences regarding their academic records and were to respond to these with the best of their ability in full and complete sentences (i.e. never with a single word). Participants assigned to the truthful condition were advised to answer the questions as accurately as they could. Participants assigned to the deceptive condition were instead advised to be deliberately deceptive in their responses to the questions but to intend on sounding as truthful as possible. Participants heard each question stimulus through the provided headphones and spoke their responses into a Shure PG58 microphone in fixed position in front of them and those responses were digitized as a continuous file (.wav) within Adobe Audition at a sampling rate of 44.1 kHz, 16-bit. A window appeared on the computer screen for each question, notifying the participant to respond and after completing the response, provided a button to move on to the subsequent stimuli. Participants proceeded with this process until a window appeared notifying the participant that the experiment was complete.

The procedure was completed over two separate blocks of trials that differed with respect to whether or not a stressor was induced on the participant. Because the impact of stressors are generally irreversible once imposed, the no stress condition always preceded the stress condition. For both truthful and deceptive participants, the experiment began with the no stress condition consisting of the task described above. This was followed by the stress condition, where participants first were notified by the researcher that a computer-based algorithm indicated they were not responding at the full capability of their assigned condition (i.e. not sounding sufficiently truthful in participant's responses). It was made clear to the participant that the series of questions would be presented again and if they did not meet the requirements as indicated by the algorithm, they would have to schedule a second session to complete subsequent trials. This information was, in fact, no more than an act of deception that was intended to act as a stressor for the participants due to the fact that it was deemed collectively by our laboratory personnel to be something that the participants would regard as an undesirable outcome.

To further reinforce the aforementioned stress induction, the emWave device provided feedback to the participant for the duration of the stress condition. The emWave device was preset so to elicit persistent red light which normally would indicate a high stress response. However, participants were told that the persistent red light was indicative of insufficient truthfulness.

Within each block of trials, the questions presented to the participants were randomized. Participants were afforded a rest break between the two blocks of trials if they desired. The entire experiment lasted between fifteen and twenty minutes. Upon completion of the task, the researcher debriefed the participant on the true nature of the study.

Acoustic Measurements

Initially, each sentence response from each participant within each condition was isolated and saved using Ableton Live 8, resulting in 34 separate sentence responses per participant (17 responses under stress and no stress conditions). Acoustic analyses on these individual sentence responses then were conducted within Praat analysis-synthesis software. F_0 analyses in Praat relied on the standard pitch range, between 75Hz (pitch floor) and 500Hz (pitch ceiling); F_0

values that existed outside of these bounds were not computed. An F_0 value was produced every 10ms based upon Praat's standard time-step settings (.75/pitch floor). We selected these settings so that each measured value would be based on a sufficiently long analysis window (40ms) to provide a reliable estimate of F_0. Specifically, this window length is based upon several period lengths of the waveform (# of periods/pitch floor = 3/75 = .040).

Aggregate analyses included: average F_0 across the utterance and average F_0 variability across the utterance [(relative) local jitter]. Median values were reported for average F_0 measures for each sentence item in order to minimize the influence of outlier values that normally are observed when tracking F_0 throughout a sentence-length utterance. Jitter was computed as percentage from the mean absolute (non-negative) difference of consecutive intervals in seconds divided by the mean period in seconds.

In order to assess critical segments, a pitch listing consisting of F_0 values calculated every 10ms across an entire utterance were exported to Excel in order to compute averages and define critical segments. Specifically, critical segments were calculated by finding the difference between the average of 50ms (five F_0 values) of F_0 values after the time point entrance into the critical segment and the average of 50ms of F_0 values before the time point entrance of the critical segment. 50ms averages were obtained by selecting F_0 values at least 30ms before and after the time point entrance into the critical segment in order to prevent overlap between each F_0 value's 40ms analysis window. F_0 values that existed outside the bounds of the pitch range were not calculated into the 50ms averages and therefore did not contribute to critical segment values.

Results

It was necessary to exclude the data from statistical analyses of aggregate measures for a few participants due to a failure to follow task instructions (by failing to respond to all question stimuli using complete sentences; N = 3). Data from a final randomly selected participant was excluded from these statistical analyses *a priori* in order to maintain equivalent sample sizes across truthful and deceptive conditions. This left a sample size of 22 participants for all aggregate analyses, with 11 each in truthful and deceptive conditions. Similarly, for analyses of critical segment measures it was necessary to exclude the data from one participant due to the fact that this participant's responses did not provide the required (50 ms) amount of voiced data prior to the entrance of the critical segment. This necessitated *a priori* exclusion of data from another randomly selected participant in order to maintain equivalent sample sizes across truthful and deceptive conditions. This left a sample size of 20 participants for the critical segment measure, with 10 each in truthful and deceptive conditions.

For analyses of critical segments it also became clear that responses to a particular question item (#11; see Appendix A), consistently failed to produce any voiced samples above threshold settings within the specified analysis window of 50 ms either before or after the time that defined entrance into the critical segment. As a result, it also became necessary to exclude all responses to that question item from critical segment analyses. This was done for both truthful and deceptive responses, leaving analyses based upon the remaining 16 question items.

For any given acoustic measure (e.g., aggregate F_0), values were collapsed across items (value/question) to produce one value per participant. For each acoustic measure, Shapiro-Wilk p-values of less than .05 were found when averaging across values for each question response, indicating that the response data may not be normally distributed. Therefore, for a given stress condition median values across sentence items were used in subsequent analyses to control for the influence of outliers, and thus provide a more reasonable measure of central tendency. For example, aggregate F_0 was computed by finding the grand median across the set of median F_0 values that were obtained for each sentence response within a given stress condition. Similarly, for each participant, jitter was computed by finding the grand median of the percent jitter values that were obtained across sentence items within a given stress condition. Finally, critical segment F_0 was computed for each participant by finding the grand median of the F_0 difference values that were obtained across sentence items within a given stress condition. Means and standard errors were computed for each measure across all participants to produce one value for each combination of deception and stress conditions.

Three mixed-subjects ANOVAs were conducted, one for each dependent measure (i.e. aggregate average F_0, aggregate % jitter, and critical segment F_0 difference). Within each of these ANOVAs, the between-subjects variable was deception with two levels, deception and truthfulness and the within-subjects variable was stress with two levels, stress and no stress. Levene's test for assumed homogeneity of variance was not violated for any measure within either no stress or stress conditions (for aggregate F_0, $F(1, 20) = .286$, $p = .599$ and $F(1, 20) = .290$, $p = .596$, respectively; for aggregate jitter, $F(1, 20) = .074$, $p = .789$ and $F(1, 20) = .019$, $p = .892$, respectively; for critical segment F_0 difference, $F(1, 18) = .149$, $p = .704$ and $F(1, 18) = .113$, $p = .741$, respectively).

Aggregate (F_0 and Jitter)

Figure 1 displays summary data for the aggregate F_0 measure for each combination of levels across stress and deception variables. Participants in the no stress condition (orange bars in

Figure 1) produced a higher F_0 relative to participants in the stress condition (blue bars). This trend contributed to a main effect of stress, $F(1, 20) = 4.858$, $p = .039$, $\eta_p^2 = .195$.

Figure 1 also reveals that the obtained mean F_0 values were similar across truthful and deceptive conditions. This was further indicated by the absence of a main effect of deception, $F(1, 20) = .063$, $p = .805$, $\eta_p^2 = .003$. Likewise, there was no significant interaction between deception and stress, $F(1, 20) = .964$, $p = .338$, $\eta_p^2 = .046$.

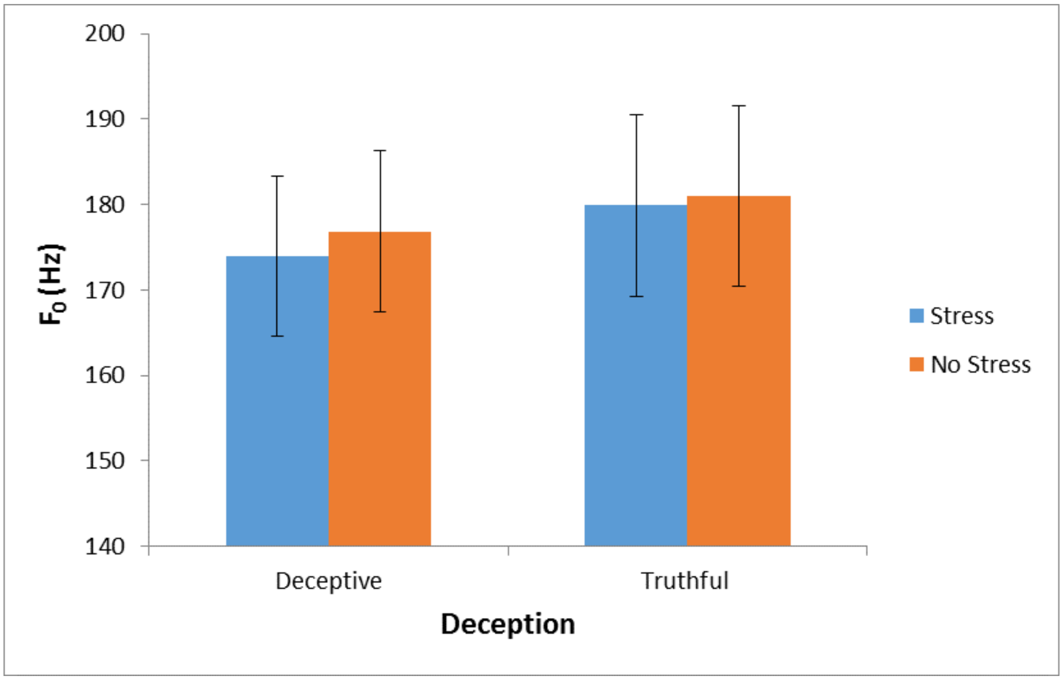

Figure 1. Mean F_0 (in Hz) and corresponding standard error bars for deceptive and truthful participants under conditions of stress and no stress.

Figure 2 displays summary data for the aggregate measure percent F_0 variability (jitter) for each combination of levels across stress and deception variables. As can be seen in the figure, jitter values were similar across levels of stress. This is indicated by the fact that there was no main effect of stress, $F(1, 20) = .352$, $p = .560$, $\eta_p^2 = .017$. Furthermore, measured jitter values were similar across levels of deception. This is indicated by the fact that there was no main effect of deception, $F(1, 20) = .119$, $p = .734$, $\eta_p^2 = .006$. An interaction between deception and stress was not observed, $F(1, 20) = 1.365$, $p = .256$, $\eta_p^2 = .064$.

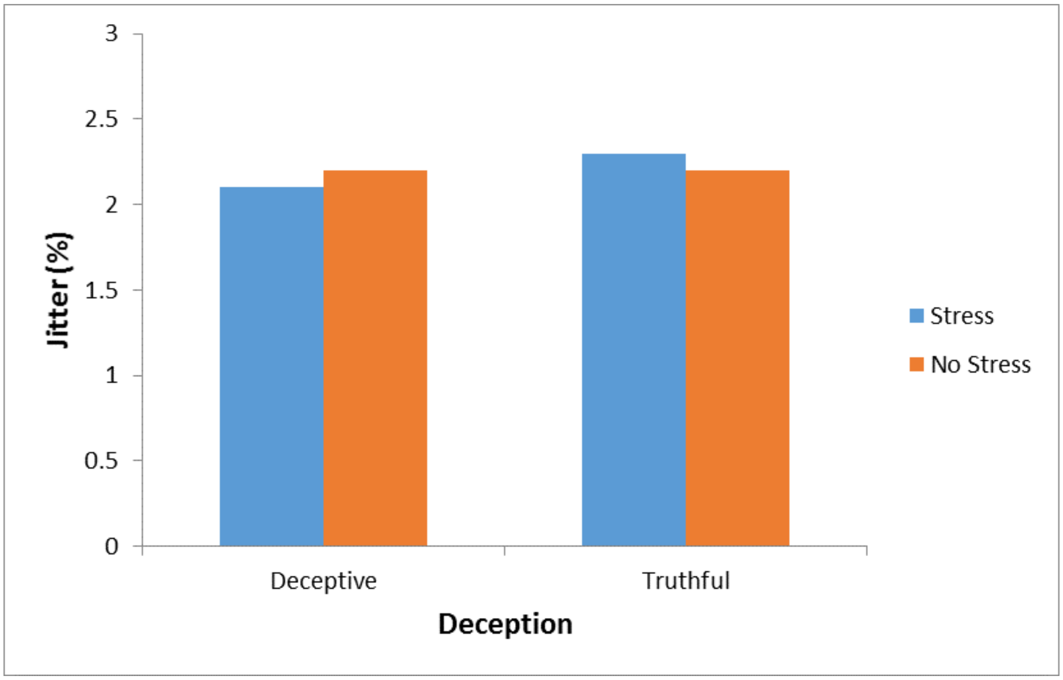

Figure 2. Mean jitter (%) for deceptive and truthful participants under conditions of stress and no stress. Standard error of measurement (not displayed) was extremely small within each condition (.01).

Critical Segments

Figure 3 displays corresponding summary data for each combination of levels across stress and deception variables for critical segments (i.e., observed F_0 differences entering critical segments). Results indicated a slight tendency for participants to increase F_0 in the no stress condition relative to the stress condition. This was confirmed by a marginally significant main effect of stress, $F(1, 18) = 3.503$, $p = .078$, $\eta_p^2 = .168$.

Furthermore, results indicated a slightly increased F_0 for truthful participants relative to deceptive participants. This tendency was indicated by a marginally significant main effect of deception, $F(1, 18) = 3.822$, $p = .066$, $\eta_p^2 = .175$. No interaction between deception and stress was observed, $F(1, 18) = .216$, $p = .647$, $\eta_p^2 = .012$.

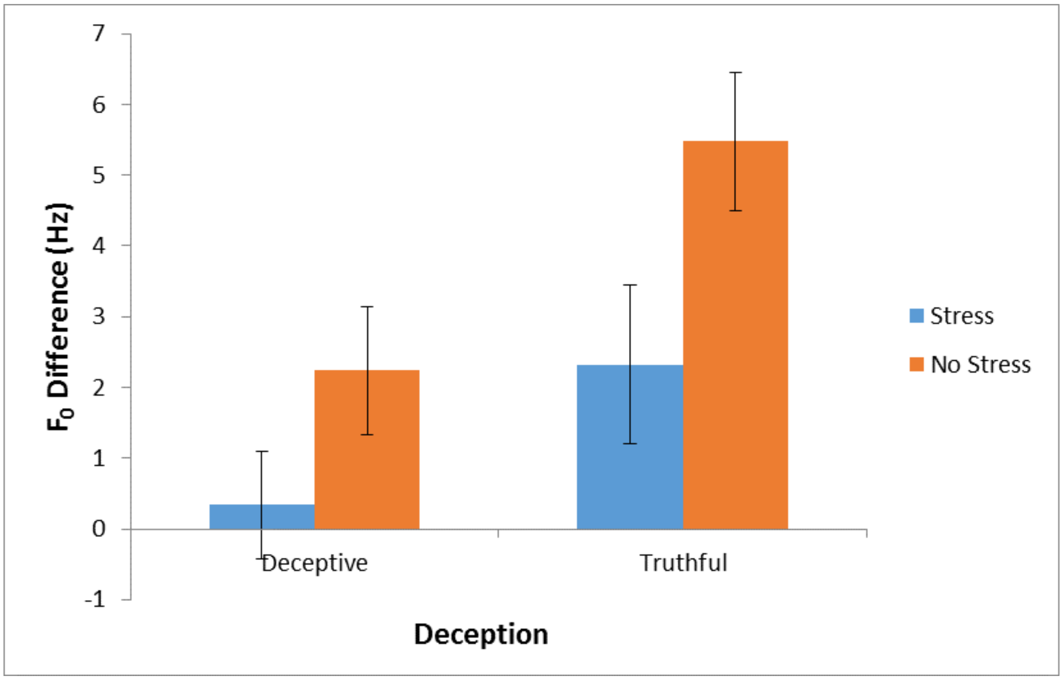

Figure 3. Observed mean F_0 difference (in Hz) entering the critical segment and corresponding standard error bars for deceptive and truthful participants under conditions of stress and no stress.

Discussion

There are suggestions from the results of the current investigation that some frequency-based measures are less likely than others to reliably indicate deception. For example, jitter was not found to significantly differ between truthful and deceptive responses, nor did it significantly vary as a function of stress. This finding is consistent with a corresponding null finding for this measure in the earlier study by Hall and Watts (2006) that used a very similar task/paradigm. Additionally, a recent study assessing F_0 variability in terms of F_0 standard deviation found no significant differences between aggregate analyses of truthful and deceptive responses (Kirchhübel & Howard, 2013). When taken together, these null findings suggest that cycle-to-cycle F_0 variability (jitter) does not effectively reveal deception. Alternatively, one could potentially argue that in the case of the current investigation and the Hall and Watts (2006) study that the failure to find significant effects for this measure was due to the paradigm. However, this seems quite unlikely given that other frequency-based measures showed much greater promise.

Similarly, no significant differences in average aggregate F_0 measures were obtained between truthful and deceptive utterances when controlling for stress. Although these results are contrary to some effects found in prior literature (Krauss, Geller, Olsen, & Apple, 1977; Villar, Arciuli, & Paterson, 2013; Protopapas & Lieberman, 1997; Anolli & Ciceri, 1997), they should not be considered particularly surprising. This is because there are considerable individual differences in the ranges of F_0 that talkers produce, and deception was manipulated as a between-subjects variable in the current investigation. For example, males typically produce lower average F_0 than females (e.g., see Traunmüller & Eriksson, n.d.). Ultimately, though, the range of produced values will change as a function of the length of, and tension on, the talker's vocal folds, which will vary across individuals. This means that the reliance on a between-subjects design to assess differences in F_0 under truthful and deceptive conditions may have increased the difficulty to detect cues to deception due to the resulting increase in the variability of aggregate F_0 that was introduced by the design.

However, it is worth noting that a recent study that instead relied upon a within-subjects design to assess mean F_0 between the truthful and deceptive utterances also produced a null result (Kirchhübel & Howard, 2013). Thus, it is possible that when measures of F_0 are aggregated across an entire utterance that the variability across those measures will often be too great to consistently permit effective evidence of deception, regardless of the choice of within- versus between-subjects design.

Aggregate F_0 data did reveal a statistically significant effect for the within-subjects variable of stress, indicating that the imposed stressors in this study did systematically impact F_0 in the voice. Interestingly, participants in the no stress condition produced a higher F_0 than those in the stress condition, in direct contrast to expectations based not only upon the primary assumptions from prominent technologies (e.g. polygraph and VSA) that there should be corresponding effects of stress and deception, but also with existing evidence that F_0 increases as a function of stress (e.g., see Demenko, n.d.). A corresponding effect of stress on F_0 also was observed within the data from critical segments, such that there was a marginally larger average increase in F_0 in the no stress condition relative to the stress condition.

One could potentially claim that the effects of stress in the current investigation could be explained if our stressor induced a differential response than the psychogenic stress assumed to underlie deception detection in polygraph and VSA (Hopkins, Benincasa, Ratley, & Grieco, 2005). However, this possibility seems unlikely given existing support for the notion that various types of stress should elicit a similar physiological response (Sapolsky, 2004), which also should apply to the physical response of the vocal folds. Given the fact that in the stress condition several deceptive participants consistently produced a lower F_0 than truthful participants, an alternative possibility is that stress simply does not affect F_0 in the same manner/direction across all tasks. In the current task, talkers appeared to actively limit/reduce F_0 under stressful conditions. This seems reasonable given the instruction set actually requested that they work to correct their productions to avoid having to return for another session. Regardless of what the appropriate explanation of the obtained effects of stress is, it should be clear from these findings that any reliable deception detection device must dissociate the effects of deception from those (potentially quite variable effects) of stress.

It is noteworthy that the only finding that was consistent with a potential effect of deception was restricted to analyses of data at critical segments, where truthful participants produced a marginally higher difference in F_0 compared to deceptive participants. Different potential explanations of this marginal effect could be posited. For instance, it could be argued that our truthful participants simply have voices that produce higher baseline F_0, and consequently provided higher observed grand means, than participants in the deceptive condition. However, several observations reduce the likelihood of this possibility. Specifically, there were more male talkers in the truthful condition (4 vs. 3 in the deceptive condition), and on average, men produce an average F_0 of 120 Hz and whereas the average F_0 for a female is 210 Hz (Traunmüller & Eriksson, n.d). Any such tendency would have lowered the baseline F_0 in the truthful condition, yet the reverse pattern was observed. Furthermore, the direction of this marginal effect was consistent with the statistically significant increase in F_0 observed by Hall & Watts (2006) at the onset of critical segments when using a within-subjects version of the task used in the current investigation. It thus appears more likely that the task and conditions produced the observed effects rather than individual differences in speech production.

This interpretation of the critical segment data leads to the suggestion that, at least under the current task conditions, that there may have been a slight tendency for liars in the current investigation to overly control F_0 in an attempt to appear truthful (e.g., for a similar explanation, see Anolli & Ciceri, 1997) and offers empirical support for the Attempted Control Theory underlying deception.

Of course, care should be taken not to make too much of a statistically marginal effect. Having acknowledged that, there are reasons to be optimistic that this trend in the critical segment data is probably a meaningful one. Specifically, effect size analyses indicated that the effect of deception on F_0 difference was moderate to large. Specifically, our results showed that approximately 18% of the variance in deception can be explained by F_0 differences. Attention should be placed on such statistical measures, as they may more accurately inform a judgment of “practical significance” (e.g., see Kirk, 1996). Consequently, it can be argued that a replication of this study with a larger sample size would likely reveal statistically significant differences. An increased sample size also would help to potentially overcome the increased F_0 variability that should be expected given the aforementioned reliance on a between-subjects design.

In contrast, our analysis of aggregate F_0 explained less than 1% of the variance in deception. The large discrepancy between critical segment and aggregate analysis of F_0 in explaining the variance of deceptive responses suggests that reliable measures of deception are more likely to be obtained from analyses of critical segments than from aggregate measures, at least for the type of task that was utilized in the current investigation. Furthermore, our data indicate that stress and deception can at least some times be dissociated phenomena that produce independent F_0 signatures. Consequently, the future development of acoustic and voice-analysis deception detection technologies must seek to disintegrate the reliance on stress induction as an indicator of deception. Additionally, the large variability of F_0 between individuals necessitates that novel technologies design protocols to reduce or control for the variability. We suggest that the GKT paradigm (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003) could potentially be repurposed to measure critical segment F_0 differences in relevant questions and compared against critical segment F_0 differences in neutral questions within one individual. By assessing frequency-based acoustic differences rather than stress responses produced by the polygraph, the novel paradigm could present a methodology that both reduces variability and dissociates stress.

Notably, all data presented in this study present deception detection implications only for a subset of frequency-based measures. The utilization of critical segment assessment could feasibly be extended to various acoustic parameters, such as formant frequency analysis, amplitude, syllable rate, and pause length. In fact, our laboratory is currently evaluating potential influences of the current paradigm on both aggregate amplitude variability (shimmer) and syllable rate at critical segments. If data from those measures coincides with the patterns that were obtained from local F_0 measures in the current investigation, then further support would be provided for the notion that Attempted Control Theory is being utilized by participants in order to appear truthful. Presumably, reproduction of data supporting the theory can provide a framework, based upon strong acoustic signatures, by which reliable technologies for detecting deception could be developed.

Appendix A

1. What is your current major at JMU?
2. What is your current cumulative GPA at JMU?
3. What is the lowest grade you've ever received at JMU?
4. What are the last four digits of your student ID?
5. Name two classes you took last semester?
6. How many semesters have you completed at JMU?
7. What state are you a resident of according to your transcript?
8. How many A's have you received at JMU?
9. Have you ever made dean's list at JMU?
10. How many psychology classes have you taken at JMU?
11. What is your middle initial?
12. What is your minor at JMU?
13. Which institutions have you attended other than JMU?
14. What year was your first semester at JMU?
15. What was your term GPA last semester?
16. Have you completed your tech level 1 and ISST tests?
17. What kind of degree are you working towards (B.A., B.S., other)?

References

Anolli, L., & Ciceri, R. (1997). The voice of deception: Vocal strategies of naive and able liars. *Journal of Nonverbal Behavior , 21(4),* pp. 259-284.

Beckford Wassink, A., Wright, R.A., & Franklin, A.D. (2006). Intra-speaker variability in vowel production: an investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *Journal of Phonetics, 35,* pp. 259-284.

Benus, S., Enos, F., Hirschberg, J., & Shriberg, E. (2006). Pauses in deceptive speech. *Proceedings ISCA: 3rd International Conference on Speech Prosody*, Dresden, Germany.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shankhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: the external validity of mock crime studies. *Journal of Experimental Psychology: Applied, 9(4)*, pp. 251-269.

Caso, L., Gnisci, A., Vrij, A., & Mann, S. (2005). Processes underlying deception: An empirical analysis of truth and lies when manipulating the stakes. *Journal of Investigative Psychology and Offender Profiling, 2,* pp. 195-202.

Damphouse, K. R., Pointon, L., Upchurch, D., & Moore, R. U.S. Department of Justice, (2007). *Assessing the validity of voice stress analysis tools in a jail setting* (219031).

Demenko, G., Oleśkowicz-Popiel, M., Izdebski, K., & Yan, Y. (2013). Variability in voice fundamental frequency in speech under stress. *TRASP'2013, P3-2,* Aix-en-Provence, France.

Ekman, P. (2001). Telling lies: Clues to deceit in the marketplace, politics, and marriage. W.W. Norton, New York.

Enos, F., Shriberg, E., Graciarena, M., Hirschberg, J., & Stolcke, A. (n.d.). *Detecting deception using critical segments*. Informally published manuscript, Columbia University , New York, NY, , Available from Google Scholar.

Everding, G. (2004). *Research casts doubt on voice-stress lie detection technology*. Unpublished manuscript, Washington University of St. Louis, St. Louis, MI.

Feeley, T.H. & deTurck, M.A. (1998). The behavioural correlates of sanctioned and unsanctioned deceptive communication. *Journal of Nonverbal Behavior, 22(3)*, pp. 189-204.

Hall, M., & Watts, C. (2006). *A multi-dimensional evaluation of vocal deception*. Unpublished manuscript, Department of Psychology & Department of Communication Sciences and Disorders, James Madison University , Harrisonburg, VA.

Hopkins, C., Benincasa, D., Ratley, R., & Grieco, J. (2005). *Evaluation of voice stress analysis technology*. *Proceedings: 38th hawaii international conference on system sciences*.

Kirchhübel, C. & Howard, D.M. (2013). Detecting suspicious behavior using speech: Acoustic correlates of deceptive speech – An exploratory investigation. *Applied Ergonomics, 44(5)*, pp. 694-702.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* pp. 746-759.

Köster, S. (2001). Acoustic-phonetic characteristics of hyper-articulated speech for different speaking styles. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 873-876.

Krauss, R. M., Geller, V., Olsen, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, *35(3)*, pp. 345-350.

National Research Council of the Nation Academies. (2003). *The Polygraph and Lie Detection* (The National Academies Press, 11-28). Washington, DC: U.S. Government Printing Office.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America*, *101(4)*, pp. 2267-2277.

Sapolsky, R.M. (2004). *Why zebras don't get ulcers*. New York, NY: St. Martin's Press.

Saxe , L. (1994). Detection of deception: Polygraph and integrity tests. *Current Directions in Psychological Science, 3*(3).

Stroemwall, L.A., Hartwig, M., & Granhag, P.A. (2006). To act truthfully: nonverbal behavior and strategies during a police interrogation. *Psychology, Crime and Law, 12(2)*, pp. 69-89.

Traunmüller, H. & Eriksson, A. (n.d.). The frequency range of the voice fundamental in the speech of male and female adults. *Institutionen för lingvistik, Stockhoms universitet, S-106* 91, Stockholm, Sweden.

Villar, G., Arciuli, J., & Paterson, H. (2013). Vocal pitch production during lying: Beliefs about deception matter. *Psychiatry, Psychology, and Law*, *20(1)*, pp. 123-132.

Zuckerman, M., Larrance, D.T., Spiegel, N.H., & Klorman, R. (1981). Controlling nonverbal cues: Facial expressions and tone of voice, *Journal of Experimental Social Psychology*, in press.