VOCALIC MARKERS OF DECEPTION AND COGNITIVE DISSONANCE FOR AUTOMATED EMOTION DETECTION SYSTEMS

by

Aaron C. Elkins

A Dissertation Submitted to the Faculty of the

COMMITTEE ON BUSINESS ADMINISTRATION

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Aaron C. Elkins entitled Vocalic Markers of Deception and Cognitive Dissonance for Automated Emotion Detection Systems and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 8/11/2011
Jay F. Nunamaker, Jr.

_____ Date: 8/11/2011
Judee K. Burgoon

_____ Date: 8/11/2011
Elyse Golob

_____ Date: 8/11/2011
Paulo B. Goes

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 8/11/2011
Dissertation Director: Jay F. Nunamaker, Jr.

_____ Date: 8/11/2011
Dissertation Director: Judee K. Burgoon

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____
Aaron C. Elkins

ACKNOWLEDGEMENTS

The Department of Homeland Security's National Center for Border Security and Immigration (BORDERS) and the National Center for Credibility Assessment (NCCA) provided significant support for this research. Similarly, funding from the Center for Identification Technology Research (CITeR), a National Science Foundation (NSF) Industry/University Cooperative Research Center (I/UCRC), supported portions of this dissertation.

The vocal analysis software investigated in the research was provided for evaluation purposes by Nemesysco Ltd.

I would like to thank my committee, Jay F. Nunamaker, Jr., Judee K. Burgoon, Elyse Golob, and Paulo Goes for their friendship and tireless efforts to help me become a better academic and researcher.

I would also like to thank Jeff Stone for his mentorship and friendship. It was a privilege to learn and research his Self and Attitudes lab.

Finally, I would like to acknowledge and thank Doug Derrick my friend and colleague from the doctoral program for developing the Embodied Conversational Agent investigated in this research.

DEDICATION

I dedicate this dissertation to my mother Lourdes Elkins. None of my success or accomplishments, including this document, would have been possible without her unconditional love and limitless support.

I would also like to thank my sister Hashaw Elkins and father David Elkins for their support and encouragement over the past four years.

I would also like to thank both my Psychologist/Social Worker mother and Engineer father for cultivating my interest and curiosity for both the technical and behavioral sciences.

It may be said with some assurance that if no one has calculated the orbit of a fly, it is only because no one has been sufficiently interested in doing so. The tropistic movements of many insects are now fairly well understood, but the instrumentation needed to record the flight of a fly and to give an account of all the conditions affecting it would cost more than the importance of the subject justifies. Difficulty in calculating the orbit of the fly does not prove capriciousness, though it may make it impossible to prove anything else. The problems imposed by the complexity of a subject matter must be dealt with as they arise. Certainly no one is prepared to say now what a science of behavior can or cannot accomplish eventually. Advance estimates of the limits of science have generally proved inaccurate. The issue is in the long run pragmatic: we cannot tell until we have tried.

-- B.F. Skinner, *Science and Human Behavior* (1953)

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF TABLES - Continued

LIST OF FIGURES

LIST OF FIGURES - Continued

LIST OF FIGURES - Continued

ABSTRACT

This dissertation investigates vocal behavior, measured using standard acoustic and commercial vocal analysis software, as it occurs naturally while lying, experiencing cognitive dissonance, or receiving a security interview conducted by an Embodied Conversational Agent (ECA).

In study one, vocal analysis software used for credibility assessment was investigated experimentally. Using a repeated measures design, 96 participants lied and told the truth during a multiple question interview. 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. Using these measurements, a logistic regression and machine learning algorithms predicted deception with accuracy up to 62.8%. Using standard acoustic measures, vocal pitch and voice quality was predicted by deception and stress.

In study two, deceptive vocal and linguistic behaviors were investigated using a direct manipulation of arousal, affect, and cognitive difficulty by inducing cognitive dissonance. Participants (N=52) made verbal counter-attitudinal arguments out loud that were subjected to vocal and linguistic analysis. Participants experiencing cognitive dissonance spoke with higher vocal pitch, response latency, linguistic Quantity, and Certainty and lower Specificity. Linguistic Specificity mediated the dissonance and attitude change. Commercial

vocal analysis software revealed that cognitive dissonance induced participants exhibited higher initial levels of Say or Stop (SOS), a measurement of fear.

Study three investigated the use of the voice to predict trust. Participants (N=88) received a screening interview from an Embodied Conversational Agent (ECA) and reported their perceptions of the ECA. A growth model was developed that predicted trust during the interaction using the voice, time, and demographics.

In study four, border guards participants were randomly assigned into either the Bomb Maker (N = 16) or Control (N = 13) condition. Participants either did or did not assemble a realistic, but non-operational, improvised explosive device (IED) to smuggle past an ECA security interviewer. Participants in the Bomb Maker condition had 25.34% more variation in their vocal pitch than the control condition participants.

This research provides support that the voice is potentially a reliable and valid measurement of emotion and deception suitable for integration into future technologies such as automated security screenings and advanced human-computer interactions.

1 INTRODUCTION

When we remember the last time we spoke to a close friend or parent, we could easily determine if they were angry or happy from just their voice. Our parent spoke louder, faster, and in a higher pitch than usual after discovering, for instance, that their grandmother's vase was broken. Contrast this with a close friend who recently had a death in their family. They sounded depressed and spoke much slower and in a lower volume than an angry parent. With the thoughts of their loved ones on their mind they would sound distracted, speak in shorter responses, and with more frequent vocal interruptions. As social creatures, we can quickly and automatically determine emotional state or mood from the voice.

Despite how effortlessly we can interpret emotion and mood from the voice, developing computer software to replicate this feat is exceedingly difficult. Computers require very specific and predictable inputs and cannot deal well with unbounded contexts and the chaotic nature of conversation. We take for granted how complex conversations are and how quickly they branch and weave back and forth between topics and ideas. We even alternate between moods and emotions in just one conversation, from anger when recounting a mean boss and back to joy when discussing an upcoming celebration.

In addition to the complexity of conversation contexts, the science of measuring and classifying emotion and deception using the voice is in its infancy. Fear, for instance, is characterized by fast speech rate, higher mean pitch, low

pitch variability, and lower voice quality (Juslin & Laukka, 2003, Juslin & Scherer, 2005). However, the relationship between vocal measures and emotion has not been well explored beyond correlational analyses, leading to conflicting results and alternative vocal profiles for emotions such as fear (Bachorowski & Owren, 1995, Juslin & Scherer, 2005).

Using the voice as a means to determine someone's emotional state is more than just a social convenience. If you heard someone angrily yelling in close proximity you would immediately, without thinking, interpret danger or a possible threat and take the appropriate actions. However, what if the person was not so overt and did not yell to conceal their anger and malicious intent? If they were ostensibly friendly would their voice reveal their true intent?

1.1 Vocal Emotion and Deception Detection for National Security and Law Enforcement

We could conjure a myriad of personal reasons why someone may try to deceive us. However, it is in the context of national security and law enforcement that the gravity of deception is evinced. Despite numerous border controls in the United States, 18 hijackers boarded planes on September 11th 2001. These terrorists deceived their way through multiple security checks including State Department Visa applications and consular interviews, U.S. Customs and Border Protection screenings, and airport security (Kean et al., 2004).

There were multiple opportunities to identify each hijacker before they boarded the plane. Ultimately, politics, human subjectivity and fallibility, and

dilapidated information systems and sharing made observing and identifying their deceptions and violent intent extremely difficult.

Despite all the advances in technology, very little progress has been made in technologies to support law enforcement and national security efforts to identify criminals and people with hostile intentions. None of the layers of security standing between the 9/11 hijackers had technology more sophisticated than identification tools (finger print scanners, identification databases) or basic environment sensors (metal and explosives detection) to aid credibility assessment. The events of 9/11 proved current technological tools are only reliable for predictable threats.

The only certainty we have is that future attackers maintain violent and hostile intentions. These behaviors co-occur with many possible emotions regardless of their specific plot or plan of attack. This reality motivates the urgency and need to measure and classify emotions in real-time in a screening and credibility assessment context. Despite this ever-present need, science and technology designed with the expressed purpose of detecting and measuring emotion using technology and advanced sensors is very limited. The polygraph examination, developed over 60 years ago, is still the best behavior analysis and deception detection technology available to law enforcement.

The technology and science behind the polygraph was primarily developed between 1895 and 1945 (Inbau, 1948, Reid, 1947, Skolnick, 1960). Moreover, the protocol for administering the polygraph examination requires a lengthy (3-5

hours) and multiphase interview to obtain reliability. These interviews are often preceded by background investigations that provide polygraph examiners additional information used to interpret and guide interviewing.

The polygraph examination still remains an indispensable tool for law enforcement, but its reliance on a lengthy interview and physically attached physiological instruments (i.e., blood pressure cuff, respiratory rate pneumograph, and galvanic skin resistance galvanometer) make it unsuitable for rapid screening environments such as the airport or border.

Most officers at a congested border entry in the United States must make a credibility assessment within 30 seconds. This rapid credibility assessment is further confounded by their divided attention to their physical environment, monitoring of behavior, and operation of technology (e.g., querying criminal and identification databases). In sharp contrast to the background investigation that typically precedes a polygraph, an officer at a pedestrian or vehicle border crossing has no advance knowledge of who is coming across the border until they arrive.

In light of the challenges faced by law enforcement to secure borders and airports, commercial security technology companies have emerged to service this niche industry. However, unlike the polygraph, none of these technologies or software systems are supported scientifically nor validated empirically. These technologies are marketed directly to law enforcement and security organizations and are unrequired to prove their deception detection capabilities.

Carl Sagan (1980) popularized the expression "extraordinary claims require extraordinary evidence." This expression captures the ethos of the modern scientist well. It also explains why the academy virtually ignores and dismisses commercial emotion and deception detection systems. Skepticism is an admirable trait, but it should not lead to myopia. The vacuum left in scientific research for predicting emotion, deception, and behavior using technology was filled by non-scientists and the commercial sector. It is hubris to believe that all non-scientific developments in detection technologies should be dismissed.

Law enforcement customers are also dubious of commercial detection systems. They believe vendors are "selling solutions in search of a problem," that they offer "one-size-fits-all technologies" with exciting feature lists. These systems depend on specific operating characteristics (e.g., polygraph style, rapid screening) and rely on single modalities (e.g., the voice) that may not be compatible with the screening and security environment.

Implementing an unreliable and invalid detection technology could place the country's security in jeopardy by failing to detect actual threats. Just as deleterious, however, would be to dismiss technology, such as vocal analysis software, before it has been thoroughly examined. This would deprive law enforcement of a valuable tool for detecting threats and scientists new innovations and insight into the science of emotion and deception detection.

The next section will introduce the current research, technology, and measurements of vocal emotion, dissonance, and deception.

2 VOCAL MEASUREMENTS OF EMOTION AND DECEPTION

2.1 Commercial Vocal Analysis Software

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 his or her voice with specialized vocal analysis software.

The UK government has invested heavily to expand usage of VRA to assess and investigate claims made over the phone for housing and social security benefits (Walker, 2008). 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 (Holguin, 2008).

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 (Damphousse, Pointon, Upchurch, & Moore, 2007, Gamer, Rill, Vossel, & Gödert, 2006, Haddad, Walter, Ratley, & Smith, 2001). Harnsberger and Hollien (2009, 2006, 2008) have evaluated vocal stress and layered voice analysis technology extensively and found no sensitivity to deception and high false positive rates (incorrect deception judgment). 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, study one explores the vocal measurements independent of the software's interface and built-in algorithms to determine their validity, composition, and potential to predict emotion, cognitive effort, stress, and deception.

2.1.1 Vocal Deception

Differences in acoustic vocal behavior exist between liars and truth tellers (DePaulo et al., 2003, DePaulo, Stone, & Lassiter, 1985, deTurck & Miller, 1985, Rockwell, Buller, & Burgoon, 1997a, Zuckerman, DePaulo, & Rosenthal, 1981). Vocal cues fall into three general categories, which include time (e.g., speech

length, latency), frequency (e.g., pitch), and intensity (e.g., amplitude) (Scherer, 1985). Previous research demonstrated that relative to truth tellers, deceivers speak in shorter durations, with slower tempos, less fluency, and exhibit greater response latencies (DePaulo et al., 1985, deTurck & Miller, 1985, Rockwell, Buller, & Burgoon, 1997a).

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 (Rockwell, Buller, & Burgoon, 1997a, Vrij, 2008). An increase in pitch or frequency has also been associated with arousal during deceptive responses (Apple, Streeter, & Krauss, 1979, DePaulo et al., 2003, Rockwell, Buller, & Burgoon, 1997a, Zuckerman et al., 1981), which presumably results from the anxiety of being caught and facing negative consequences (Apple et al., 1979, DePaulo et al., 2003, Zuckerman et al., 1981).

2.1.2 Vocal Stress Analysis Software

The previous generation of software for analyzing voice to detect deception preceding is called Vocal Stress Analysis (VSA) and has consistently failed to reliably detect deception in experimental or field settings (Damphousse et al., 2007, Haddad et al., 2001). Despite the richness of features present in the voice, previous VSA systems focused on a very small frequency band of 8-12Hz (Haddad et al., 2001). This is because the human body exhibits periodic contractions of the muscles known as microtremors on this narrow and low frequency range

(Lippold, 1971, Lippold, Redfearn, & Vučo, 1957). VSA systems attempt, unsuccessfully, to measure this frequency produced by the larynx muscles.

VSA systems assume 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.

The two primary VSA programs in use today are the National Institute for Truth Verification Federal Services' CVSA (2011) and X13-VSA (X13-VSA Ltd., 2011).

2.1.3 Full Spectrum Vocal Analysis

Modern vocal analysis software uses the full spectrum of the vocal information contained in the voice. In addition to measuring frequency and intensity, modern vocal analysis software measures 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 be explained in more detail in the subsequent vocal measurements section. Nemesyco is currently the primary developer of full spectrum vocal analysis software, which they refer to as Layered Voice Analysis (LVA). Nemesysco develops software customized for a number of scenarios from home use (eX-Sense) to security investigation (LVA 6.50). The security investigation software LVA 6.50 is the vocal analysis software utilized in this research.

2.1.3.1 Vocal Measurements

The vocal analysis software provides measurements intended to reflect deception, emotion, cognitive effort, and stress. The variables calculated by the primary software investigated, Nemesysco's LVA 6.50, are listed and described in Table 1 based on the software documentation. It is also important to note that there is no current theoretical explanation or support for the descriptions provided by the vendor, which is part of the impetus for investigating this software.

Table 1. Vocal Measurement Descriptions

Variable	Description	Measures
SPT	Emotional level	Average 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	"Say or Stop", indication of fear or unwillingness	
FJQ	Imagination	Uniformity of low frequency
FMAIN	Stress Level	Most significant frequency in the range
FX	Level of Concentration	Frequencies above FMAIN
FQ	Deception	Uniformity of frequency spectrum
FFLIC	Embarrassment or conflicting thoughts	Frequency spectrum 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 of 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.

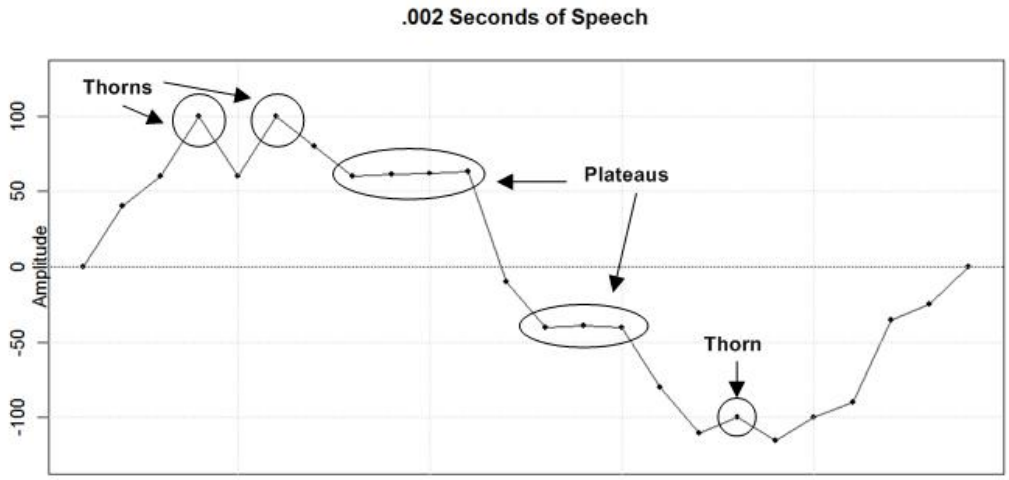


Figure 1. Thorns and Plateaus in Voice Segment

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 (2007) 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 or resampled.

2.2 Deception

Ekman and Friesen's (1969) leakage hypothesis predict that liars leak verbal and non-verbal behaviors that can discriminate them from truth tellers. These cues leak in response to underlying arousal, negative affect, cognitive

effort, and attempted control (Buller & Burgoon, 1996). Buller and Burgoon (1996) introduced Interpersonal Deception Theory (IDT), which expanded deception into a strategic interaction between a sender and receiver. This reconceptualization of deception predicts that liars must simultaneously manage information, behavior, and image during the interaction. The multitude of interaction responsibilities paired with dyadic relationship, history, motivation, modality, skill, and contextual factors further explain and predict behavioral differences between liars and truth tellers.

IDT predicts that arousal, affect, and cognitive effort contribute to leakage cues by liars. However, these conditions are difficult to manipulate using traditional deception paradigms in an experimental setting. Participants are typically asked to make sanctioned lies in exchange for a monetary bonus for success with no aversive consequences. While fabricating a consistent message during an interaction should leak cognitive effort cues, the dynamic interplay between both affect and arousal is missing in most deception experiment paradigms. Study two uses a manipulation of cognitive dissonance to influence the affect and arousal experienced during deceptive communication.

2.3 Cognitive Dissonance

Cognitive dissonance is a psychological discomfort or tension felt when there is inconsistency between cognitions (Festinger, 1957). This discomfort is described as a drive state, similar to hunger, that motivates the person to reduce the inconsistency. The magnitude of the dissonance felt is predicted to stem from

the importance of the inconsistency. Strategies available to reduce cognitive dissonance are a functions of how open the inconsistent cognition is to change (Festinger, 1957, Olson & Stone, 2005).

For example, if smokers believe that smoking is bad for their health, they would experience cognitive dissonance. Smokers aware of the dangers of smoking could remove this inconsistency by quitting smoking or convincing themselves that smoking is not unhealthy. However, these strategies can be very difficult to accomplish. Alternatively, smokers can add additional cognitions, such as "I am more likely to die in a car accident", or, "the scientific evidence is inconclusive" to reduce the importance of the inconsistency, and thereby reduce the cognitive dissonance.

2.3.1 Cognitive Dissonance and Arousal

The unpleasant feeling or arousal that accompanies important inconsistent cognitions is predicted by cognitive dissonance theory. The earliest support for the existence came from studies that demonstrated performance impairment and enhancement when dissonance was induced (Olson & Stone, 2005, Pallak & Pittman, 1972, Waterman, 1969). Later, the research focused on attitude change through misattribution. Zanna and Cooper (1974) found that when participants misattributed their cognitive dissonance induced negative feelings to a placebo pill, which they were told would cause tension, no attitude change occurred. This occurred because participants believed their negative

feelings were not a result of their inconsistent behavior or cognitions. Thus, they had no motivation to change their beliefs.

There have been few studies that attempt to measure arousal directly. The physiological measure of galvanic skin response has been used to demonstrate that an arousal occurs in greater levels for cognitive dissonance induced participants (Elkin & Leippe, 1986, Losch & Cacioppo, 1990, Olson & Stone, 2005). Moreover, these studies suggest that arousal may motivate an attitude change, but the affect of the participants actually leads to dissonance reduction, not the reported change in attitude. Reminding participants that they acted inconsistently by asking them to report their attitude actually sustains the arousal (Elkin & Leippe, 1986).

2.3.2 Cognitive Dissonance and Deception

The traditional Induced-Compliance paradigm for inducing cognitive dissonance requires all participants to lie or make counter-attitudinal arguments. The liars' degree of arousal and motivation is varied by manipulating the degree of choice participants felt when agreeing to lie. High choice liars are motivated to reduce their dissonance and predicted to be more aroused. Low choice liars attribute their behavior to being forced to lie and do not experience cognitive dissonance and its concomitant arousal.

2.4 Vocalics and Linguistics

Vocalics refer to qualities of speech distinct from the verbal or linguistic content (Juslin & Scherer, 2005). Vocalics falls in the category of non-verbal communication referring to "how" something was said instead of "what" was literally said. Linguistics encompasses the verbal message, or "what" was said.

We often take for granted how effortlessly we can infer emotional information from the voice of our speaking partner. Even if our significant other says "I am not upset", we can infer from their voice that they may actually be sad or angry. The incongruence between the vocal tone and verbal message is diagnostic of their emotion. This underscores the importance of investigating both "what" and "how" a message is communicated.

2.4.1 Arousal and Cognitive Effort

The relationship between traditional vocal measures (e.g., F_0 or Pitch, Intensity, and Tempo) and emotion is not clear. Fear, for instance, is characterized in the majority of vocal studies as fast speech rate, higher mean pitch, low pitch variability, and lower voice quality (Juslin & Laukka, 2003, Juslin & Scherer, 2005). However, the relationship between vocal measures and emotion has not been well explored beyond correlational analyses, leading to conflicting results and alternative vocal profiles for fear (Bachorowski & Owren, 1995, Juslin & Scherer, 2005).

Previous research has found that an increase in the fundamental frequency, which is heard as pitch, is related to stress or arousal (Bachorowski &

Owren, 1995, Streeter, Krauss, Geller, Olson, & Apple, 1977). Pitch is a function of the speed of vibration of the vocal chords during speech production (Titze & Martin, 1998b). Females have smaller vocal chords than men, requiring their vocal chords to vibrate faster and leading to their higher perceived pitch.

When we are aroused, our muscles tense and tighten. When the vocal muscles become tense they vibrate at a higher frequency, leading to a higher pitch. Similarly, previous research has found that when aroused or excited, our pitch also exhibits more variation and higher intensities (Juslin & Laukka, 2003).

Deceptive speech is also predicted to be more cognitively taxing, leading to non-strategic or leakage cues (Buller & Burgoon, 1996, Rockwell, Buller, & Burgoon, 1997b). These cues, specific to cognitive effort, can be measured vocally. Cognitively-taxed speakers take longer to respond (response latency) and incorporate more nonfluencies (e.g., “um” “uh”, speech errors).

2.4.2 Emotion

Our thoughts and emotions are communicated and articulated into words. For example, the words “love” or “nice” connote more positive emotion than “hurt” or “ugly” when used in speech or text (Francis & Pennebaker, 1993, Newman, Pennebaker, Berry, & Richards, 2003, Tausczik & Pennebaker, 2010). Using automated text analysis and validated emotion dictionaries, previous research has revealed 21 linguistic cues and their corresponding categories that discriminate between deceptive verbal messages (Newman et al., 2003, Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). The categories are word Quantity,

Complexity, Certainty, Immediacy, Diversity, Specificity, and Affect. Table 2 details these categories and their corresponding linguistic cues.

Table 2. Linguistic Cues

Category	Cues
Quantity	Word and Verb Count
Complexity	Word Length
Certainty	Modal Verbs, Modifiers
Immediacy	Passive Voice, Impersonal Pronouns
Diversity	Lexical Diversity
Specificity	Sensory (see, hear, feel), Temporal, Spatial Imagery
Affect	Emotion, Pleasantness, Activation

The linguistic cues used in study two of this research were extracted using the automated linguistic analysis software Structured Programming for Linguistic Cue Extraction (SPLICE), which incorporates the Dictionary of Affect in Language (DAL), and the Linguistic Inquiry and Word Count (LIWC) (Francis & Pennebaker, 1993, Moffitt, 2010, Whissell, 1989).

IDT and previous deception research would predict that liars would display less Certainty, Immediacy, Quantity, Complexity, Diversity, Specificity, and Affect words (Buller & Burgoon, 1996, Zhou et al., 2003). However, the

correspondence between previous deception research and cognitive dissonance-induced lying is unclear and is explored in study two.

3 STUDY ONE – PREDICTING AND VALIDATING VOCAL DECEPTION AND EMOTION

3.1 Introduction

Using a deception experiment, this study examines how reliable and valid commercial vocal analysis software and standard acoustic measurements of the voice are for predicting emotion and deception in security screening contexts. While research exists that evaluates current vocal analysis software's built-in classifications, there is a gap in our understanding on how it may actually perform in a real high stakes environment.

Previous research on vocal analysis software for deception detection has relied on experiments with low-stakes and sanctioned lying. This is a necessary limitation for scientists staying within the bounds of ethical treatment of their human subjects. We would, for instance, not induce participants to lie under threat of incarceration or bodily harm. This leads to an inconsistency between the intended vocal analysis software operating environment and the experimental or evaluative environment. Specifically, that poor deception detection by vocal analysis software could be because the built-in classification is insensitive to the unrealistic experimental conditions.

To address this alternative explanation this study examines the variables produced by commercial vocal analysis software for predictive potential and statistical validity in identifying emotion and deception. It is unrealistic to rely

completely on the voice to detect deception and hostile intent for all people and all situations. But, by exploring the vocal variables used by the software, we will be better able to correspond and fuse them with other detection technologies for higher prediction reliability and accuracy.

3.2 Deception Experiment

The experiment consisted of an interview that required participants to alternate between deceptive and truthful responses that were recorded and analyzed with vocal analysis software. The focus of the experiment was to identify systematic patterns of vocal behavior that vary as a function of truth or deception.

Previous deception research has found that lying is more cognitively demanding than telling the truth (DePaulo et al., 2003). 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 (Vrij et al., 2008).

Deceivers also exhibited shorter response lengths, talking time, and lengths of interactions (Burgoon, 1983, deTurck & Miller, 2006). The reduced response time is explained as a deceptive individual's reticence to provide more information than necessary (Rockwell, Buller, & Burgoon, 1997b).

Using the measurements provided by the vocal analysis software the following hypotheses were specified.

H1: Liars will exhibit higher vocal measurements of cognitive effort than truth tellers.

H2: Liars will exhibit shorter message lengths than truth tellers.

Due to the absence of theory surrounding the vocal measurements calculated by the software, a research question (R1) exploring the differences on vocal measures between liars and truth tellers was specified. All of the unexpected significant findings will be corrected to reflect the experiment wise error of testing 13 simultaneous vocal measurements. At the $\alpha=.05$ level, this corresponds to 48.7% chance of Type-I error (Rice, 1989).

R1: Is there a difference on vocal measures between liars and truth tellers?

In addition to testing the above hypotheses and research question, this study evaluates the classifications provided by the software, explores the factor structure of the vocal measurements, and models and compares custom deception classifiers using statistical and machine learning methods.

3.3 Method

3.3.1 Participants

International participants (N = 220) were recruited from a southwestern university for a study on culture and credibility in interviews. In exchange for their participation, they were offered information on effective interviewing, a $15 payment, and the opportunity to earn up to an additional $20 if successful in convincing the interviewer of their credibility. Because of differences in recording equipment or poor audio quality, only 96 of the original 220 participants were included in this study. Low signal-to-noise ratio of recordings was the primary contributor to the reduction in usable audio. The recording environment noise levels were high because of experimental equipment (thermal camera refrigeration unit). This noise level overwhelmed the audio of participants speaking in a low volume.

Of the 96 participants, 53 were male and 43, female, with a mean age of 26.1 (SD = 11.2) and ranged from 18 to 77 years. By ethnicity and nationality, 53% reported themselves as Caucasian, 28% reported an Asian ancestry, 8% self-identified as African American, 7% were Hispanic (either U.S. or from a Spanish-speaking country), and 3% fit other categories.

3.3.2 Procedures

Upon arrival at the research site, participants completed a consent form and a questionnaire that measured pre-interaction goals and demographics.

These measures were used in the validation and interpretation of the factor structures. They were informed that in an upcoming interview, they would be instructed to answer some questions truthfully and some, deceptively, and that their goal was to convince the interviewer of their credibility and truthfulness. Success in doing so would earn them the bonus payment.

They then joined the interviewer, a professional examiner, in a separate room equipped with audiovisual recording equipment and a teleprompter which was hidden from the interviewer's view. The teleprompter instructed the interviewee to tell the truth or lie on each of the questions. Of interest to the current experiment are the initial 13 questions that elicited brief, one-word answers and were meant to provide some opportunity to acclimate to the environment plus supply the interviewer with baseline exposure to the interviewee's response patterns.

To counterbalance truth and deceit, participants were randomly assigned to one of the following two deception (D) and truth (T) sequences:

SEQUENCE ONE: DT DDTT TD TTDD T

SEQUENCE TWO: DT TTDD TD DDTT T

The questions listed in Table 3, required short, one to two word answers and were designed to be either charged or neutral. Neutral questions such as, "Where were you born?" and "What city did you live in when you were 12 years old?" were meant to be straightforward questions devoid of any emotion or stress. In contrast, charged questions such as, "Did you ever take anything from a

place where you worked?" and "Did you ever do anything you didn't want your parents to know about?" were intended to evoke enhanced emotional responses because of the implications of the answer. For instance, saying that you stole something from a place where you work is normatively inappropriate and should induce more stress or cognitive effort in the response as compared to neutral questions. Of the 13 questions, only 8 of the questions varied the lie and truth condition between participants.

Table 3. Short Answer Questions

Question
1. Is today Sunday? (N)
2. Are there any lights on in this room? (N)
3. Where were you born? (N)
4. Did you ever take anything from a place where you worked? (C)
5. Did you bring any keys with you today? (C)
6. If I asked you to empty your wallet purse or backpack would anything in it embarrass you? (C)
7. Is the door closed? (N)
8. Are you now sitting down? (N)
9. What city did you live in when you were 12 years old? (N)
10. Did you ever do anything you didn't want your parents to know about? (C)
11. Name the country stamped most often in your passport? (N)
12. Did you ever tell a lie to make yourself look good? (C)
13. If I told you I didn't believe you, would that bother you (C)

Note. C refers to charged questions and N to neutral questions.

Following the interview, participants completed post-measures and were debriefed while interviewers recorded their assessments of interviewee truthfulness and credibility.

3.4 Instrumentation

3.4.1 Self-Report Measures

Prior to the interview, participants completed an 18-item measure of interaction and relationship goals (α=.74), self presentation goals (α=.84), and motivation to appear credible (α=.82) developed by Burgoon, White, Ebesu, Koch, Alvaro, and Kikuchi (1998). If participants are unmotivated during an interaction, their deceptive and truthful performances will not be representative of what occurs outside the laboratory.

Following the interview, participants reported the degree of stress (α=.88) and cognitive effort (α=.85) they experienced during the interview. To assess the communication skill of participants, which should reflect the ability and skill to communicate both truth and deception, an abbreviated version of the 120 item Riggio (1986) Social Skills Inventory was used to capture Social Expressivity (α=.81), Social and Emotional Control (α=.72), Social and Emotional Sensitivity (α=.69), and Emotional Control (α=.67). Social Skills was measured to serve as potential covariates during analysis because participants with greater social skills may leak few vocal cues to deception.

Participants completed three measures of cultural orientation: the Singelis, Triandis, Bhawuk, and Gelfand (1995) Horizontal and Vertical Dimensions of Individualism and Collectivism, the Gudykunst and Lee (2003) Interdependent and Independent Self Construal, and the Park and Guan Positive and Negative Face scale (2006). Descriptions and reliabilities for the culture

measurements are listed in Table 4. These measurements were collected to control for cultural variation from our international participants. For instance, participants high on the Independent Self Construal may be more likely to take offense and become stressed when asked direct or confrontational questions. This would lead to a more stressed voice regardless of deception.

Table 4. Culture Measurements

Cultural Measurement	Reliability (α)	Description
Horizontal individualism	.65	Orientation toward individual uniqueness, responsibility and action
Horizontal collectivism	.76	Extent to which group harmony takes priority over personal preferences and goals
Vertical individualism	.77	Extent to which individual is competitive and puts self advancement over that of others
Vertical collectivism	.66	Degree to which individual subordinates self-interest to those of the family and superiors
Self positive face	.86	Degree to which individual is concerned with own self presentation and favorable image
Self negative face	.61	Degree to which individual values own freedom and independence
Other face	.63	Extent to which individual is concerned with protecting other person's face and not imposing on the other

The relationship between cultural dimensions and the vocal measurements is previously unexplored. However, it was expected that whether or not a participant speaks English as their first language should affect

measurements focused on cognitive effort. This is because participants might be translating the question and response in their native language in their mind.

3.4.2 Vocal Analysis Software and Segmentation

Although the past validation efforts have been discouraging about the role of the voice for discriminating truth from deception, the quest for better instruments and for features that are reliable indicators of stress has continued unabated, motivated by the continued belief that the voice remains a rich source of information about cognitive and emotional states and by the desired to find an automated solution to detecting them. One such system, a new-generation, commercial vocal analysis software, called Layered Voice Analysis (LVA), which is in use today by international law enforcement, was utilized for this study. LVA analyzes the full spectrum of the voice instead of merely a narrow or micro frequency band. The full spectrum software not only claims to predict stress, but also emotion, cognitive effort, thought, and deception (Nemesysco, 2009a).

The LVA 6.50 full spectrum software package was used to analyze the 96 participant audio files. Each of the recorded interviews were listened to in real-time to mark segments as noise, interviewer speech, or participant response. The vocal analysis software generated vocal measurements for each segment marked as “participant.” Of the 13 short answer questions there were 1,181 valid vocal responses. The mean response length of each vocal measurement was .47 seconds (SD = .40) and consisted of primarily one word responses (e.g., "Yes", "No").

3.4.3 Standardization

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

3.5 Deception Experiment

3.5.1 Results of Vocal Analysis Software Built-in Classifier

3.5.1.1 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) (Green & Swets, 1966). 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.

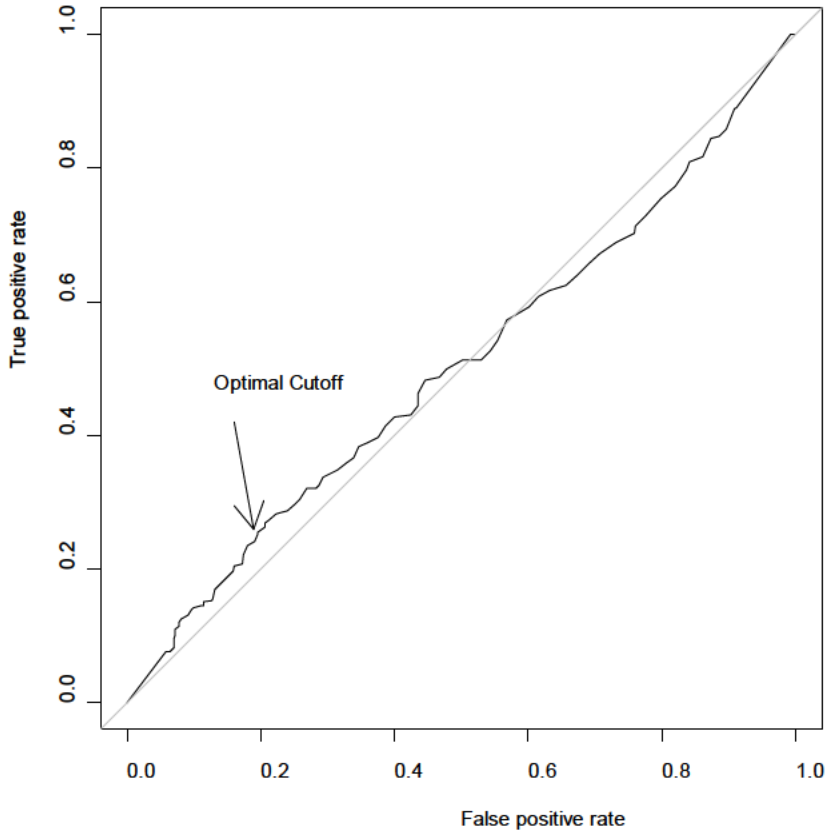


Figure 2. Roc Curve of vocal analysis software Deception Detection

Per question, the vocal analysis software had an overall accuracy ranging from 48.86%-57.89% and an AUC ranging from .46-.59. The full set of ROC curves for each question is included in Appendix A. 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.

Table 5. AUC And Total Accuracy of Vocal Analysis Software Deception Classification

Question	Accuracy	AUC
1. Where were you born?	51.04%	0.56
2. Did you ever take anything from a place where you worked?	57.30%	0.56
3. Did you bring any keys with you today?	48.86%	0.39
4. If I asked you to empty your wallet purse or backpack would anything in it embarrass you?	49.47%	0.46
5. What city did you live in when you were 12 years old?	52.63%	0.49
6. Did you ever do anything you didn't want your parents to know about?	49.45%	0.46
7. Name the country stamped most often in your passport?	57.89%	0.59
8. Did you ever tell a lie to make yourself look good?	55.68%	0.53

The poor lie detection results using the built-in algorithms are congruent with previous research utilizing the vocal analysis software (Damphousse et al., 2007, Harnsberger et al., 2009, H. Hollien et al., 2008). 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 (Eriksson & Lacerda, 2007).

In addition to a probability of deception, the vocal analysis software provides a classification of the emotion, stress, or truthfulness for each response. The visualization in Figure 3 of the classification category by the truth or lie condition provides extra clarity on the internal classifications of the vocal

analysis software (Cleveland, 1993). The frequencies or counts of each LVA classification by Lie (o) and Truth (+) are made comparable. It indicates greater discrimination by the classification when the symbols for Lie (o) and Truth (+) are farther apart. 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.

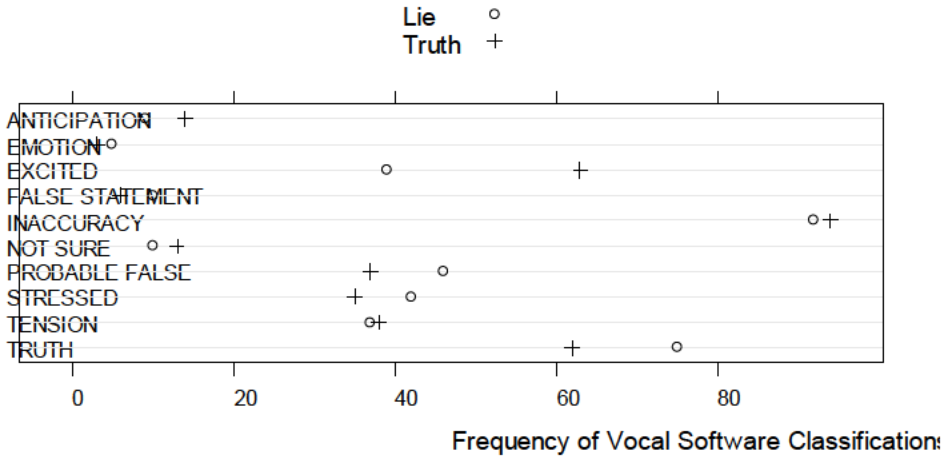


Figure 3. Dot Plot of Vocal Analysis Software Classification Frequency (Count) by Truth or Lie

3.5.2 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. 70% of all the Stressed classifications occurred during responses to neutral questions. Contrastingly, 87.5% of the responses classified as Excited occurred during charged questions. If the classifications are valid, this may mean charged questions caused excitement, but not stress to the participants.

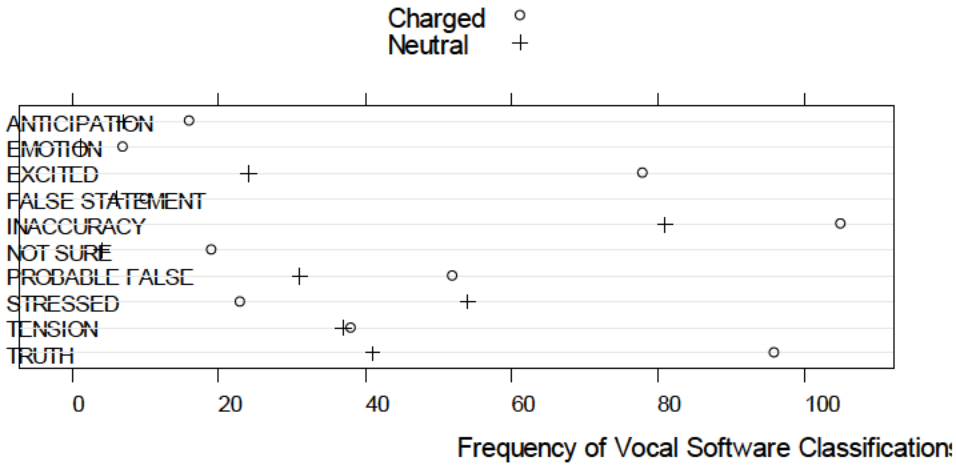


Figure 4. Dot Plot of Vocal Software Analysis Classification Frequency (Count) by Question Type

3.6 Analysis of Vocal Measurements

In order to test the experimental hypotheses and research question on a sample containing unbalanced observations, a multilevel regression model was used in place of traditional ANOVA (Gelman & Hill, 2007, Singer & Willett, 2003). 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, especially with physiological measurements (Moskowitz & Hershberger, 2002). 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 R1 and H1 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 (Singer & Willett, 2003).

3.6.1 Results of Experimental Treatment

Table 6 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.

R1: Is there a difference on vocal measures between liars and truth tellers?

The R1 research question was affirmed 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 (Apple, Streeter, & Krauss, 1979; Hocking & Leathers, 1980).

The FMain results can be qualified by examining Table 7 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.

H1: Liars will exhibit higher vocal measurements of cognitive effort than truth tellers.

The H1 hypothesis was supported by finding a significant χ^2 for JQ and AVJ shown in Table 6 in addition to significant negative coefficients for the Truth condition found in Table 7. This suggests that participants in our sample had higher average AVJ and JQ scores when lying than when telling the truth.

Table 6. Results of Deviance-Based Hypothesis Tests on Vocal Measurements (N=737, 96 Subjects x 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

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 (Goldman-Eisler, 1968, Smith & Clark, 1993, Vrij et al., 2008).

Table 7. 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 random effects in Table 7 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 high. Until this within subject variability is accounted for, predicting deception through vocal behavior will be imprecise.

H2: Liars will exhibit shorter message lengths than truth tellers.

The H2 hypothesis was not supported. 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 2 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.

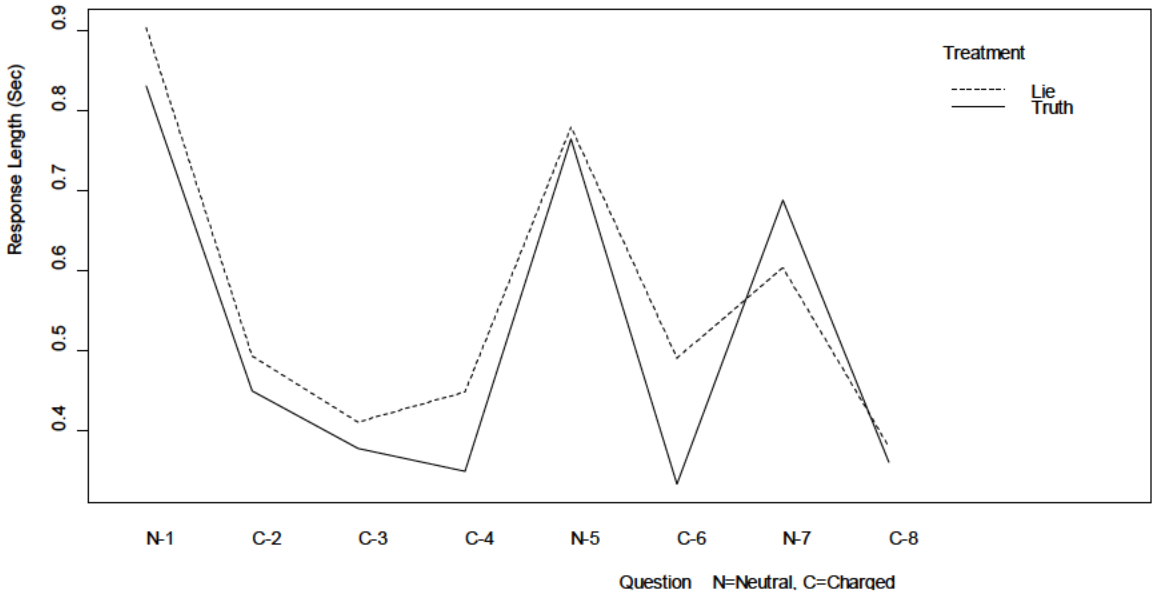


Figure 5. Interaction of Question and Truth on Response Length

3.6.2 Moderators of Lying on Vocal Measurements

3.6.2.1 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 in Table 8 reveals the pattern illustrated in the bottom of Figure 6. 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, $r(735)=.82$, $p<.001$).

Table 8. Random Intercepts of JQ for Each Question

Question	Random Intercept
1. Where were you born?	0.70
2. Did you ever take anything from a place where you worked?	-0.10
3. Did you bring any keys with you today?	-0.33
4. If I asked you to empty your wallet purse or backpack would anything in it embarrass you?	-0.14
5. What city did you live in when you were 12 years old?	0.48
6. Did you ever do anything you didn't want your parents to know about?	-0.24
7. Name the country stamped most often in your passport?	0.31
8. Did you ever tell a lie to make yourself look good?	-0.33

Note. All vocal measurements were standardized to z-scores

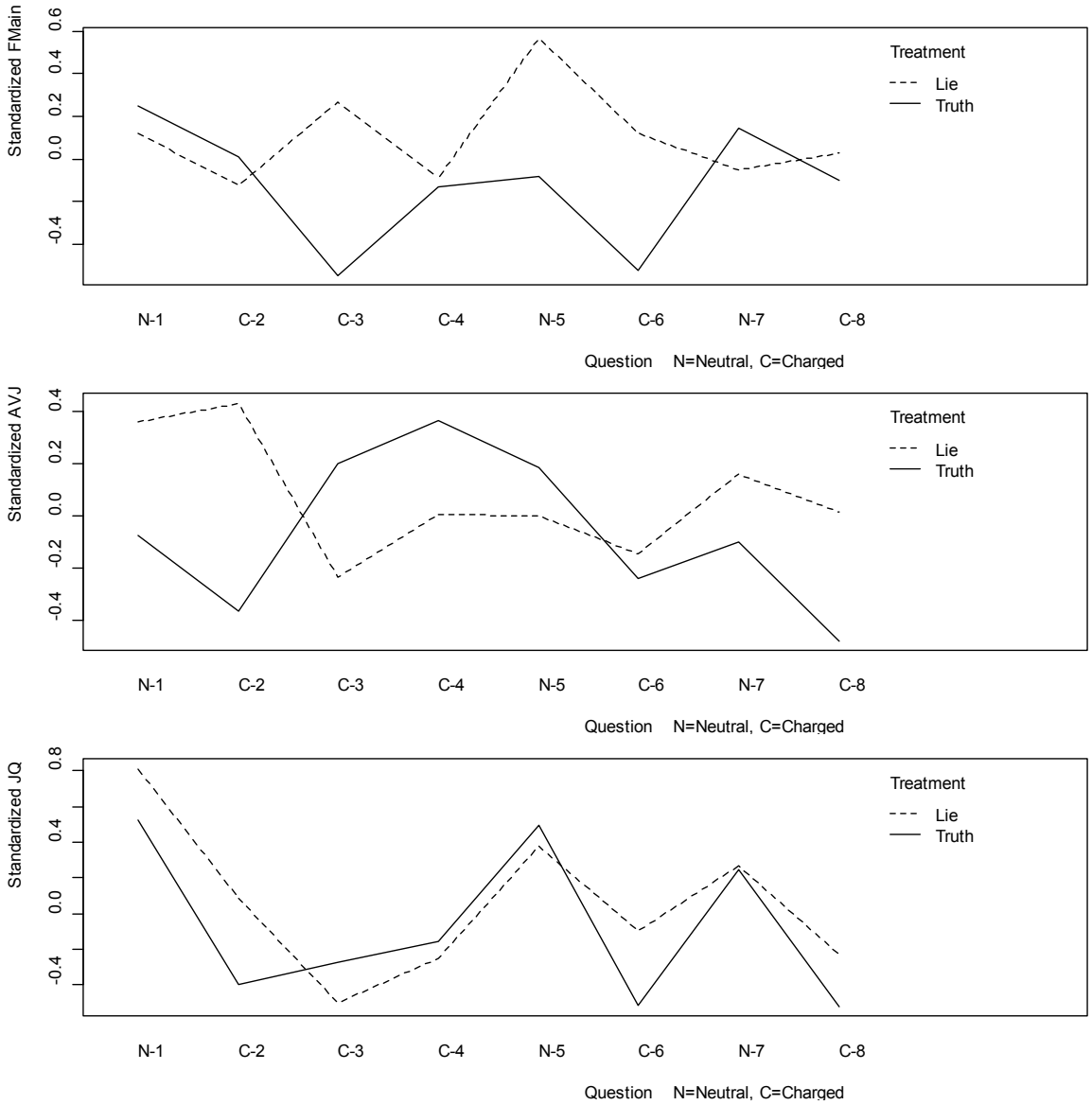


Figure 6. Interaction of Question and Truth Treatment nn FMain, AVJ, and JQ

Figure 6 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. The 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 7 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.

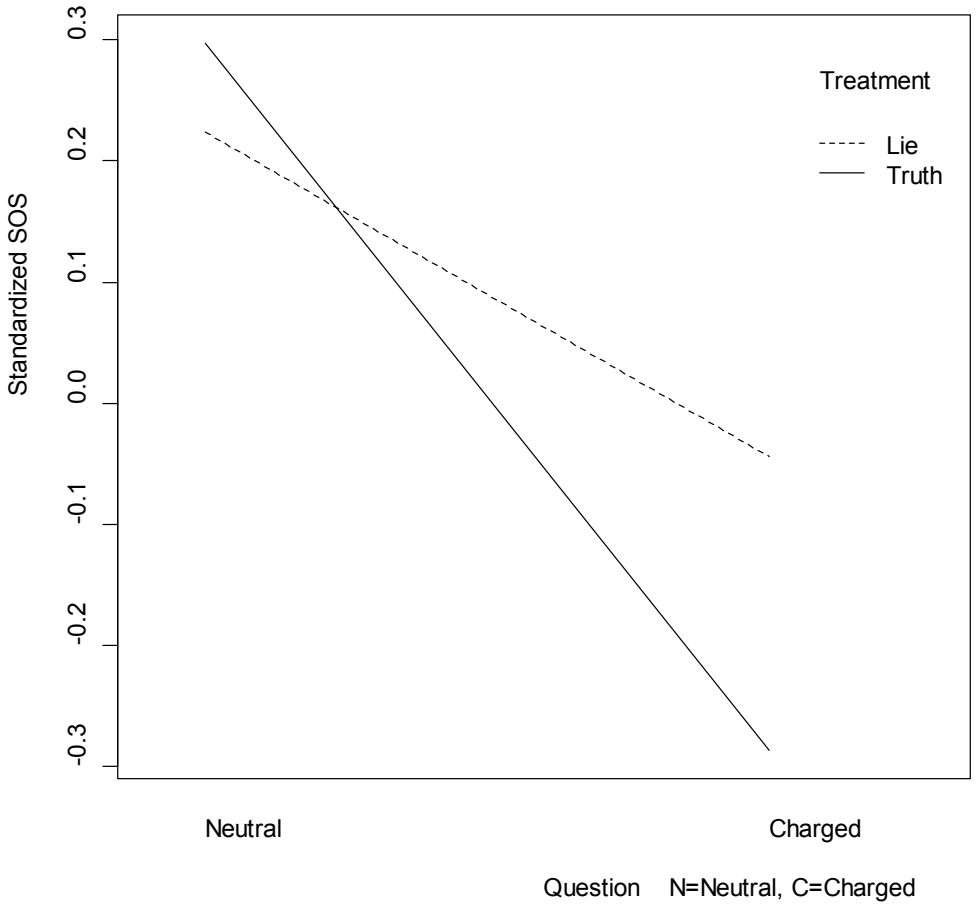


Figure 7. Interaction of Charged Question and Truth on SOS

In order to improve the predictive power of future models incorporating vocal measurements, covariates should be included to account for the within subject variance. However, there is very little known on what the vocal measurements are actually measuring. The next section will explore the vocal measurements more fully using multilevel factor analysis (MFA) and lasso regression.

3.7 Vocal Measurements

3.7.1 Multilevel Factor Analysis

The first step in validating the vocal measurements is to identify the relationships among the different measured variables, the underlying dimensionality, and to interpret any latent factors that emerge.

Each of the participants responded to 13 short answer questions during the interview; each response was processed with the vocal analysis software to generate 13 sets of vocal measurements (96 subjects x 13 questions). While the experimental analysis only focused on the 8 questions that varied by treatment, this analysis is incorporating every vocal measurement from the study. Some of the responses could not be processed by the software because of high signal-to-noise ratios (i.e., very quiet responses). Despite missing responses, there were 1,181 total sets of vocal measurements in the sample.

Traditional factor analysis assumes independence of observations (Rummel, 1970). In order to use all of the information provided by the 1,181 observations in an Exploratory Factor Analysis (EFA), the dependency of observations must be accounted for using modern methods such as Multilevel Factor Analysis (MFA). Muthén (1991) and Reise et al. (2005) have developed a set of procedures to follow when conducting an MFA. The four step procedure of conducting an MFA on the repeated vocal measurements is detailed next.

3.7.1.1 Step 1: Factor Analysis of the Total Correlation Matrix

The total correlation matrix containing the 1,181 observations were treated as independent and submitted to an exploratory factor analysis using the Maximum Likelihood method and Geomin oblique factor rotation to allow correlated factors. A four factor solution was extracted from the sample correlation matrix with eigenvalues of 2.67, 2.50, 1.32, and 1.21 ($\chi^2(32)=174.54$, p=<.001, CFI=.965, RMSEA = .061). Despite the significant χ^2 statistic, the CFI and RMSEA fit statistics suggest a moderately good fit. The χ^2 test of model fit is sensitive to large sample sizes that increase power and over emphasize even minor deviations between the estimated population correlation matrix and the sample correlation matrix (Bollen, 1989). The resulting factor loadings are displayed in Table 9 and are illustrated in Figure 8.

The interpretation of factors extracted from the total correlation matrix is misleading because it assumes no reliable between-individual differences (Reise et al., 2005). However, based on the description of the variables provided by the vocal analysis software vendor, we could interpret factor 1 as Emotional Stress, factor 2 as Thinking, factor 3 as Conflicting Thoughts, and factor 4 as Cognitive Fear.

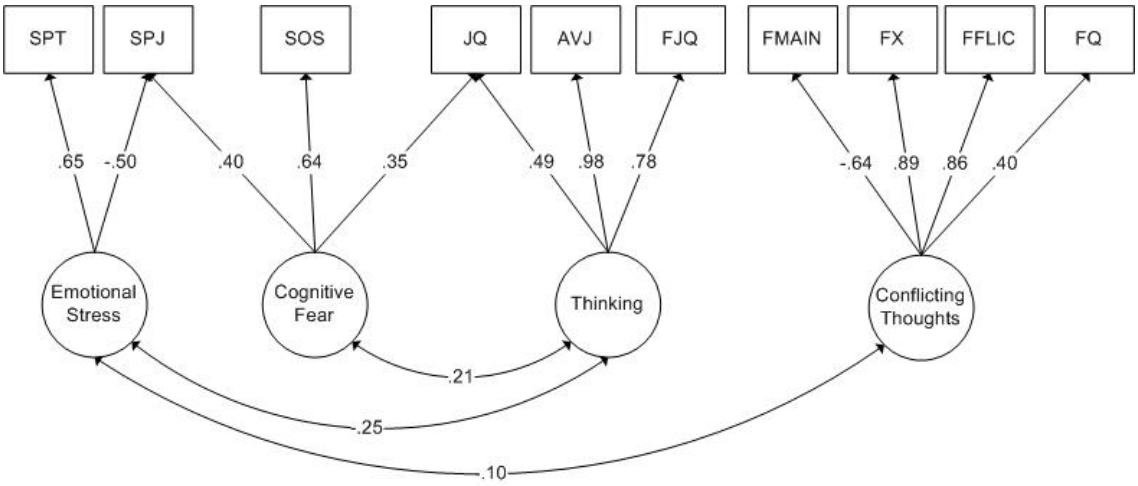


Figure 8. Exploratory Factor Analysis of Vocal Measurements Based on the Total Correlation Matrix

3.7.1.2 Step 2: Establishing Between-Individual Variation

Intraclass correlations (ICC) measure how much variance in a variable is attributable to between subject variance (Muthén, 1991). ICC values for each of the vocal measurements are listed at the bottom of Table 24 available in Appendix A. ICC for the vocal measurements ranged from .09 to .61, suggesting a high degree of between subject variance that could seriously impact the extraction of factors if clustering is ignored. Muthén and Satorra recommend the following measurement of Design Effect, which is the magnitude of distortion to methods when clustering is ignored (1995).

Where var_C is the variance under cluster sampling, var_{SRS} is the variance assuming simple random sampling, c is the cluster size, and ρ is the ICC value.

$$Design\ Effect = \frac{var_C(\widehat{K})}{var_{SRS}(\widehat{K})} = 1 + (c-1)\rho \qquad (1)$$

With 96 subjects, the Design Effect of ignoring clustering for the vocal measurements range from 9.55 to 58.95. This can be interpreted, roughly, as resulting in an underestimation of standard errors by about 10 to 60 percent. This calculated Design Effect also relates to other methodological decisions, such as using OLS with complete-pooling in place of a multilevel or hierarchal regression. The high Design Effect supports the decision to use MFA to examine the vocal measure's factor structure.

3.7.1.3 Step 3: Factor Analysis of Within Matrix

The total correlation matrix is partitioned into separate within and between matrices. The total, within, and between correlation matrices can be seen in Table 24 available in Appendix A. The within matrix, was submitted to an exploratory factor analysis using the Maximum Likelihood method and Geomin oblique factor rotation. A four factor solution was extracted from the within-sample correlation matrix with eigenvalues of 2.52, 2.18, 1.32, and 1.19 ($\chi^2(64)=230.11$, p=<.001, CFI=.954, RMSEA = .047). The factor loadings for the total extracted factors are displayed in Table 9 and Table 10.

Table 9. Factor Loadings for Total Analyses

Item	1	2	3	4
SPT	0.65	0.06		0.04
SPJ	-0.50	0.06		0.40
JQ	-0.11	0.49	-0.08	0.35
AVJ	0.09	0.98		
SOS	0.09	-0.03	0.06	0.64
FJQ		0.78		
FMAIN			-0.64	0.34
FX			0.89	
FQ		0.08	0.40	-0.14
FFLIC	-0.05		0.86	
ANTIC	0.15		0.23	0.10
SUBCOG	-0.17	0.32	0.12	
SUBEMO		-0.07	0.12	0.15

Note. Factor loadings not significant at the .05 level are omitted; underline represents loadings greater than .50.

Table 10. Factor Loadings for Within and Between Analyses

	Within				Between			
Item	1	2	3	4	1	2	3	4
SPT	-0.59		0.52		1.09			
SPJ	0.53				-0.44	0.55		
JQ	0.43	0.43				0.62	0.77	
AVJ	-0.04	1.03		0.02		0.94		
SOS	0.15		0.48	0.05			0.91	
FJQ		0.63				1.03		
FMAIN		0.05	0.38	-0.60			0.21	
FX				0.89				0.99
FQ			-0.15	0.38				0.54
FFLIC	0.07			0.83				0.98
ANTIC				0.17	0.33			0.42
SUBCOG	0.11	0.06		0.11		0.67		
SUBEMO			0.29	0.12				0.15

Note. Factor loadings not significant at the .05 level are omitted; underline represent loadings greater than .50.

As illustrated in Figure 9, interpretation of the factor loadings could be factor 1 as Cognitive Effort, factor 2 as Thinking or Imagination, factor 3 as Emotional Fear, and factor 4 as Conflicting Thoughts.

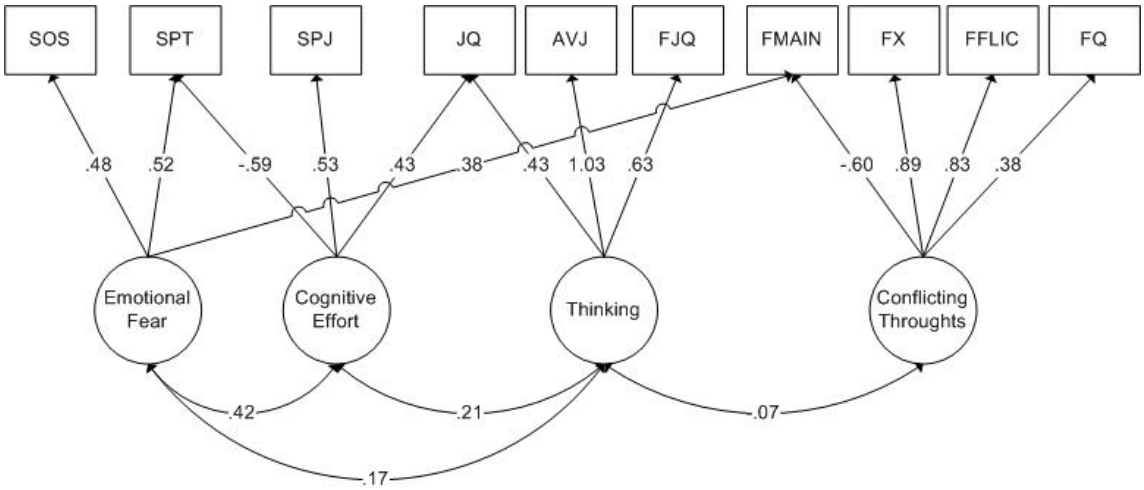


Figure 9. Exploratory Factor Analysis of Vocal Measurements Based On the Within Correlation Matrix

3.7.1.4 Step 4: Factor Analysis of Between Matrix

The estimated between-sample correlation matrix was entered into an exploratory factor analysis using the Maximum Likelihood method and Geomin oblique factor rotation. A four-factor solution was extracted from the estimated between-sample correlation matrix with eigenvalues of 4.31, 2.6, 1.23, and 1.23. The model fitness statistics are identical to the within factor analysis because the within-sample and between-sample factors are extracted at the same time. Interpretation of the between-sample factor loadings is very different from the within factor loadings. The factors can be interpreted as the between-subject

mean differences on the vocal measurements. For instance, on the Cognitive Effort factor, there is a very high loading of 1.09 on SPT. This suggests that there are individual differences on how SPT (average number of thorns) is measured by subject, in effect mirroring the ICC scores discussed earlier.

The between factor has a different pattern of loadings than the within factor loadings. This may be indicative of a measurement that varies by individual. This hypothesis should be tested by a confirmatory factor analysis in a future study. It should be noted, that when the software is used in Online mode there is a calibration that occurs for each subject. This study used the Offline mode to segment the recordings and generate the vocal measurements because vocal analysis occurred post-experiment.

3.8 Interpretation of Vocal Measurement Factors

3.8.1 Lasso Regression

In order to better interpret the results of the factor analysis, the vocal measurements were regressed on the participant's self-reported cognitive effort, stress, motivation, cultural orientation, interaction goals, and motivation. A lasso regression was conducted to explore covariates that might aid interpretation of the extracted factors and measurements. The lasso solves an ordinary least squares (OLS) regression constrained to a total sum of absolute standardized estimate coefficients less than a tuning value (Tibshirani, 1996). This promotes shrinkage (i.e., coefficients = 0) and parsimonious model selection (Wright &

London, 2009). As the tuning value grows, the lasso approaches an OLS regression. The lasso uses least angle regression (LARS) to solve the following equation simultaneously for increasing values of the tuning value t (Efron, Hastie, Johnstone, & Tibshirani, 2004).

Where (x^i, y_i) are the predictor and response variables, $i = 1,2, \dots, N$

and $x^i = (x_{i1}, \dots, x_{ip})$ for p predictor variables:

$$(\hat{\alpha}, \hat{\beta}) = argmin \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \quad subject\ to \sum |\beta_j| \leq t \qquad (2)$$

Using the equation above, each of the vocal measurements calculated by the vocal analysis software was regressed on the entire set of self-reported covariates (x^i) measured during the experiment for each participant.

3.8.2 Results of Lasso Regression and Factor Interpretation

It is important to note that these results are for exploratory purposes and interpretation of the vocal measurements. The Lasso regression does not account for clustering within subject. Table 11 summarizes the lasso regression results for the Conflicting Thoughts factor. The six predictors that explain the most variance in the vocal measurement are listed from left to right in order of magnitude (i.e., contribution to the R^2). The results of each Lasso regression are organized by the factors that emerged from the extracted within matrix. Table 11 contains the vocal variables that loaded highest on factor 1, the Conflicting Thoughts factor. The reported R^2 reflects the total explained variance when including all six predictors in the regression.

The first result, relevant to all vocal measurements, is that all the variables in the Conflicting Thoughts factor vary by gender. This can be explained by the lower fundamental frequency of the male (85 to 155 Hz) versus female voice (165 to 255 Hz) (Titze & Martin, 1998c). Additionally, the highest loading variables FX and FFlic both have stress accounting for the next highest amount of variance. This suggests that this factor may pick up on tension or negative arousal. The variance explained by cognitive difficulty and the truth condition of the subject may coincide with conflicting thoughts.

Table 11. Conflicting Thoughts Factor Lasso Regression Results

Loading		R^2	1	2	3	4	5	6
0.89	FX	.04	Female	Stress	Horiz Ind	Horiz Coll	Response Length	Social Control
0.83	FFLIC	.06	Female	Stress	Cog Diff	Motivation	Response Length	Horiz Ind
-0.60	FMAIN	.11	Female	Response Length	Stress	Truth	Motivation	Social Control
0.38	FQ	.04	Female	Response Length	Self Neg Face	EFL	Horiz Coll	Stress

A multilevel regression of FMain with random subject effects revealed significant fixed effects on Female, Response Length, Stress, Truth, and Motivation. The full table comparing the fit and significance of FMain models accounting for within-subject variance is available in Table 25 available in Appendix A. It should be noted that with these variables included in the model there is a within-subject variation of SE=.05, down from SE=.18 by including the Truth condition of the subject.

The Thinking factor summarized in Table 12 has the largest amount of variance explained (i.e., factor items have highest R^2s) by the self-report predictors. The inclusion of age and cognitive effort as predictors for all of the Thinking factor variables supports that the factor reflects thinking or thoughts. Aged adults have been shown to require more cognitive effort in free recall tasks than younger adults (Macht & Buschke, 1983). The response length might reflect longer responses due to extra thinking or age.

The distinction between the Conflicting Thoughts and Thinking factor is a negative valence. The Conflicting Thoughts factor reflects negative arousal stemming from maintaining inconsistent cognitions, which might be explained by the theory of cognitive dissonance (Festinger & Carlsmith, 1959). This theory predicts that there is a negative drive state when an individual maintains inconsistent or dissonant cognitions (Aronson, 1969). In the case of the sample data, the inconsistent cognition might be the lie told to the interviewer.

Table 12. Thinking Factor Lasso Regression Results

Loading		R^2	1	2	3	4	5	6
1.03	AVJ	.15	Female	Response Length	Stress	Age	Cog Diff	Other Face
0.63	FJQ	.11	Female	Cog Diff	Age	Response	OtherFace	Horiz Coll
0.43	JQ	.72	Response Length	Female	Age	Cog Diff	Stress	Self Pres Goals

The Cognitive Effort factor in Table 13 is more difficult to interpret. The variables of SPJ and JQ are related to the average number of plateaus and standard error of plateaus respectively. The major departure is the negative

loading of SPT, which is the average number of thorns per sample. The vendor lists this variable as related to high frequency and emotional level. Interestingly, English as First Language and Social/Emotional Sensitivity were predictive of SPT. It is possible that a participant high on Social/Emotional sensitivity could be affected by the perceived valence of the lie interaction and partner. It may be more appropriate to refer to this factor as Emotional Cognitive Effort to reflect the interplay between decreased cognitive fluency due to increased emotions.

Table 13. Cognitive Effort Factor Lasso Regression Results

Loading		R^2	1	2	3	4	5	6
-0.59	SPT	.07	EFL	Soc/Emo Sen	Soc Exp	Female	Age	Horiz Ind
0.53	SPJ	.152	Female	Self Pres Goals	Age	Soc Sen	Soc Exp	EFL
0.43	JQ	.72	Response Length	Female	Age	Cog Diff	Stress	Self Pres Goals

The Emotional Fear factor summarized in Table 14 is positively related to SPT, its highest loading factor. The combination of SOS described as an indication of fear and higher FMain levels indicating stress, suggest this factor reflects increased emotional levels accompanied by negative arousal because of trepidation to respond. Interestingly, Horizontal Individualism and Self Negative Face partially accounted for the variance in SOS, which includes responses to items such as, “I want my privacy respected” and “It is important for me to not make my conversation partner look bad.” These cultural orientations towards interpersonal communication are reflected by SOS as hesitations or fear during conversation.

Table 14. Emotional Fear Factor Lasso Regression Results

Loading		R^2	1	2	3	4	5	6
0.52	SPT	.07	EFL	Soc Sen	Soc Exp	Female	Age	Horiz Ind
0.48	SOS	.08	Response Length	Horiz Ind	Truth	Female	Self Neg Face	Stress
0.38	FMAIN	.11	Female	Response Length	Stress	Truth	Motivation	Social Control

A confirmatory factor analysis on a new dataset would need to be performed to assess the fit of the extracted factor model (Brown, 2006). The next step in assessing the validity of the vocal measurements would be to test the factor structure's invariance to time or groups. If the vocal measures had differing factor structures across time or groups, the vocal analysis software's validity would be called into question. For example, the vocal analysis software might measure voice over the phone differently than in an airport. In this situation, the system would not provide reliable predictions if the same underlying prediction algorithms were employed in both cases.

3.9 Predicting Deception

3.9.1 Methodology

In order to test the predictive ability of the models and prevent over fitting to the sample data, the data were randomly partitioned into a training set (N=368) and testing set (N=369) (Han & Kamber, 2006). A logistic regression, decision tree using recursive partitioning, and Support Vector Machine (SVM) were fit and tuned to the same training data. The final models were then used to

predict the lie or truth classification on the testing dataset and their results compared. The built-in software lie prediction was also compared against the test set and had an overall accuracy of 49% and an AUC of .52.

3.9.2 Logistic Regression

Based on the experimental findings, a multilevel logistic regression was specified to predict the probability of deception using the fixed effects of FMain, AVJ, question type, SOS, and the interaction between SOS and question type. Question and subject were included as random effects. Consistent with the experimental findings, Table 15 details the fixed effects for the logistic model with FMain providing the strongest contribution to predicting the probability of deception.

Table 15. Results of Fitting Multilevel Logistic Model for Predicting Deception on Training Data (N=368)

Fixed Effects	b (SE)
Intercept	1.630** (0.514)
FMain	-0.033*** (0.009)
AVJ	-0.011~ (0.007)
CQ	1.425~ (0.812)
CQ * SOS	-0.180~ (0.094)

~p<.10; *p<.05; **p<.01; ***p<.001

Note: Models were fit by Laplace approximation.; CQ is Charged Question

To predict deception, Equation 3 was specified, which takes the inverse logit of the sum of coefficients multiplied by the vocal measurements. The equation was then calculated against the test data set to produce a set of probability estimates of deception.

$$\frac{e^{1.63 - 0.033FMain - .011AVJ + 1.425CQ - .18CQ*SOS}}{(1 + e^{1.63 - 0.033FMain - .011AVJ + 1.425CQ - .18CQ*SOS})} \quad (3)$$

The final logistic regression prediction resulted in an overall accuracy of 55% and an AUC of .51. Examining the ROC curve in Figure 10 for the logistic regression reveals that it performed best at 67% TPR vs. a 59% FPR at the more liberal end of the prediction cutoff. Looking at the accuracies by question in Table 16 suggests this model is more appropriate for charged questions.

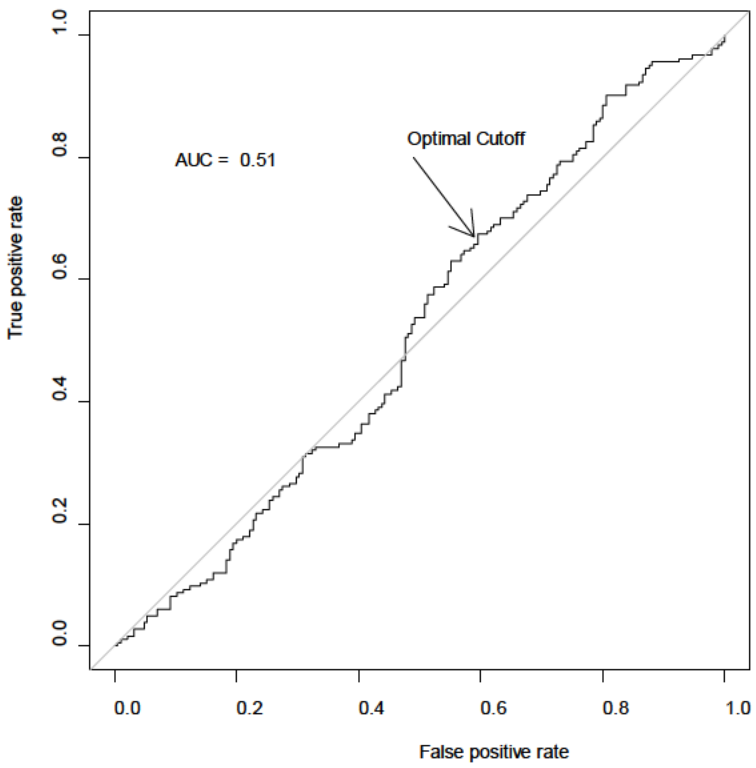


Figure 10. ROC Curve of logistic regression deception classification

Per question, the logistic regression had a prediction accuracy ranging from 46% to 62%. It performed best on charged questions, suggesting that its reliance on FMain as the primary predictor requires enhanced levels of stress or negative arousal to differentiate between the frequency or pitch of liars and truth tellers.

Table 16. Detection Accuracy Comparison by Question

	Built-in	Logistic	Tree	SVM
Where were you born? (N)	51.04%	**53.19%**	42.55%	**53.19%**
Did you ever take anything from a place where you worked? (C)	57.30%	**59.52%**	54.76%	54.76%
Did you bring any keys with you today? (C)	48.86%	**62.50%**	52.50%	52.50%
If I asked you to empty your wallet...would anything in it embarrass you? (C)	49.47%	47.17%	**50.94%**	**50.94%**
What city did you live in when you were 12 years old? (C)	52.63%	**59.18%**	55.10%	55.10%
Did you ever do anything you didn't want your parents to know about? (C)	49.45%	**62.79%**	51.16%	**62.79%**
Name the country stamped most often in your passport? (N)	**57.89%**	51.06%	38.30%	46.81%
Did you ever tell a lie to make yourself look good? (C)	**55.68%**	45.83%	54.17%	50.00%

Note: Highest accuracy in bold.

3.9.3 Decision Tree

In contrast to the logistic regression, a more exploratory tool was implemented to examine the structure of the vocal measurements and their relationship to lying or telling the truth. A decision tree that performs greedy recursive partitioning of the data was fit to the training data set (L. A. Clark & Pregibon, 1992, Koziol et al., 2003).The advantage of this method is the ease of interpreting the decision tree results, which reflect the subsets and condition of variables that best differentiate liars and truth tellers (Han & Kamber, 2006).

The entire set of vocal measurements was included as candidate predictors in the initial model to classify truth or deception. A cross validation was run on the initial model to prune variables and levels of the decision that resulted in the least prediction error. This was done to reduce over fitting to the training data set.

The final decision tree pictured in Figure 12 suggests that FMain and FQ provide the best separation between liars and truth tellers. While FMain represents the main frequency of the voice, FQ is meant to capture the uniformity of the frequency spectrum. The vendor of the vocal analysis software indicates that deception is indicated as the FQ value approaches high or low levels.

The decision tree model would find voices over a certain main frequency (FMain) or pitch threshold as deceptive. However, if they had low main frequencies, then the model would look for abnormally low uniformity of the frequency value to classify deception. However parsimonious, these rules proved too naïve to predict deception; the decision tree achieved an overall accuracy of 52% and an AUC of .51. The ROC curve for the decision tree classifier reveals no redeeming prediction ability beyond the chance level.

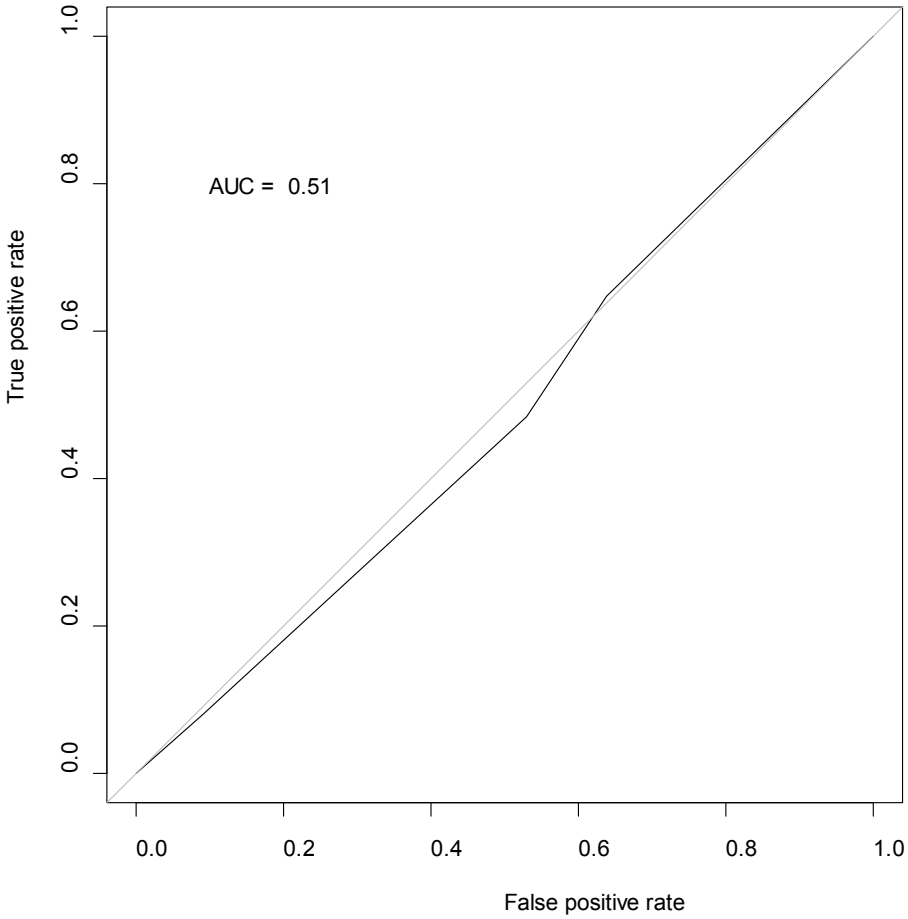


Figure 11. ROC curve of decision tree classification

The decision tree performed poorly on individual questions with an overall accuracy ranging from 38% to 55%. The decision tree classifier was out-performed by the software's built-in deception detection classification.

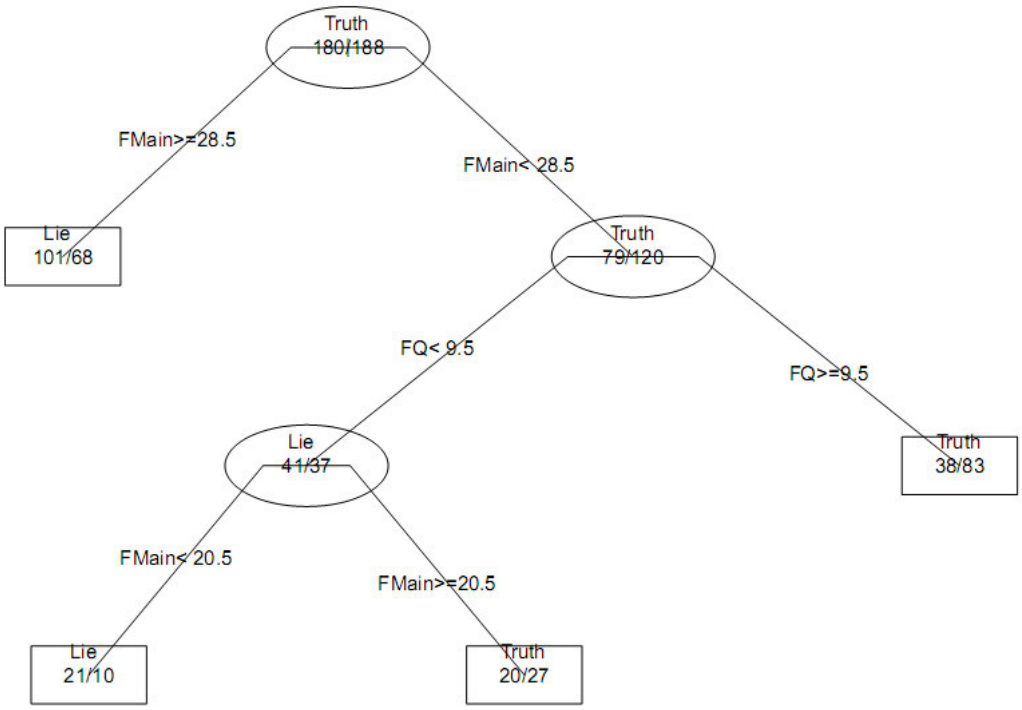


Figure 12. Decision Tree for Classifying Truth or Deception Using Vocal Measurements

3.9.4 Support Vector Machine

A SVM with a Radial Basis kernel function was used to develop a classifier of truth or deception (Chang & Lin, 2001, R Development Core Team, 2011) on the training data. The SVM approach maps the solution to high dimensional space to find the greatest separation between liars and truth tellers among the predictor variables in the model. Unlike the previous classifiers, this method assumes no linearity and can be difficult to interpret outside of its accuracy values (Chen & Lin, 2006, Efron et al., 2004).

The optimal Cost and γ parameters for the SVM model were selected by performing a grid search with 10-fold cross validation on the training set (Optimal Cost=1 and γ=.125). The cost parameter increases the penalty in the search for misclassifying, while γ can be considered a smoothing parameter. A 10-fold cross validation was conducted to avoid over fitting the model to the training data set (Han & Kamber, 2006).

The SVM underperformed the logistic regression using the predictors implied by the experimental results. To represent the entire dimensionality of the vocal measurements and promote interpretability, surrogates from the within-subjects factors extracted from the MFA and found in Table 9 on page 66 were included in the model. The variables with the highest loadings on each factor were included to reflect the Conflicting Thoughts, Thinking, Emotional Cognitive Effort, and Emotional Fear factors. The final SVM model included SPT, AVJ, SOS, and FX.

The SVM prediction model had an overall deception detection accuracy of 53% and an AUC of .56. Unlike the logistic regression classifier, the SVM performed consistently above chance throughout most of the ROC curve illustrated in Figure 13. The SVM achieved the best overall TPR-FPR ratio, with optimal prediction at the 46% TPR vs. 38% FPR cutoff.

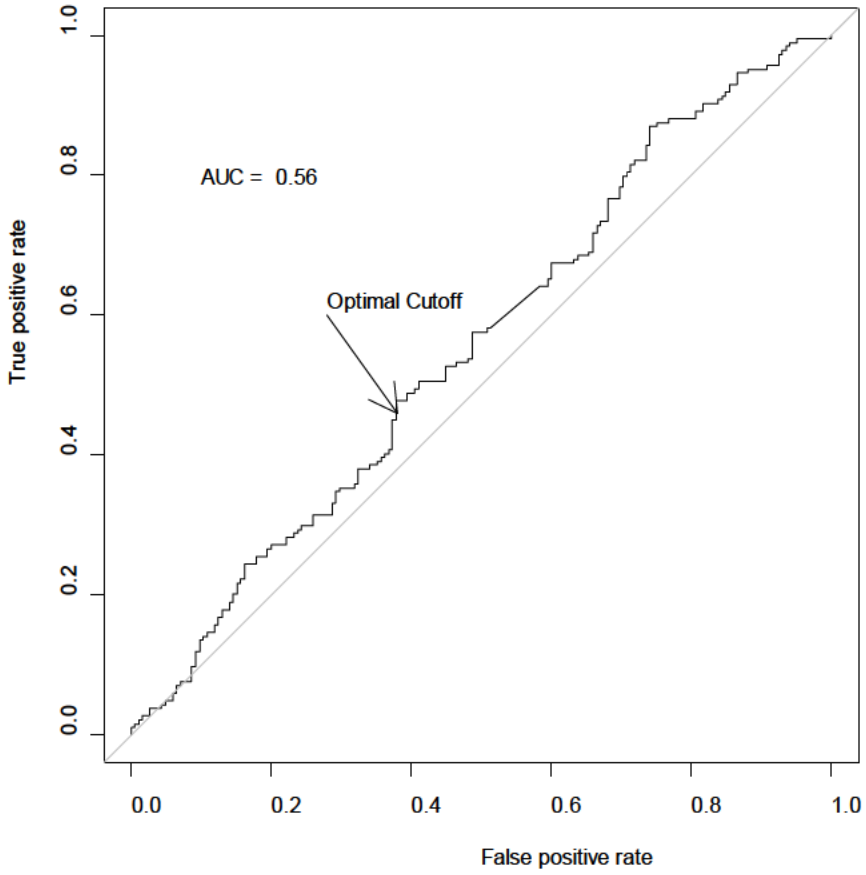


Figure 13. ROC curve of SVM classification

Per question, the SVM had a prediction accuracy ranging from 47% to 63%. The SVM performed poorly on all questions except the charged questions, “Did you ever do anything you didn't want your parents to know about?” which is similar to the results attained by the logistic regression.

3.10 Standard Acoustic Vocalics

For researchers one obstacle towards understanding modern vocal analysis software is its black box nature. The software and measurements are based on proprietary technology closely guarded to protect a commercial

investment. This makes validating the reliability and validity of such technology difficult, as demonstrated by this study, due to the absence of available theory. Even more distressing, is how little progress can be made in understanding and learning from the vocal analysis software research as a result. This may be a major contributor to the current skepticism surrounding vocal analysis software.

To relate the results of this experiment with current and vocalic research, standard phonetic measurements were calculated on the same experimental vocal recordings using the Phonetics software Praat (Boersma, 2002). A script containing the Praat code used to calculate all of the measurements is available in Appendix C.

3.10.1 Vocal Measurements

Previous research has found that people speak with a higher more varied pitch or fundamental frequency when under increased stress or arousal (Bachorowski & Owren, 1995, Streeter et al., 1977). However, there are many other factors that can contribute to variation in pitch. For instance, the words spoken can strongly influence the average pitch of an utterance because different phonemes or vowels emphasize higher or lower pitches. There is also variation between people, who have different resonance characteristics, accents, language proficiency, gender, and intonation (Ladefoged, 2001, Titze & Martin, 1998a).

One aim of this analysis using standard vocal measurements is to identify statistical controls for individual variations and reliably predict vocal pitch as a function of stress. Absent linguistic content, all of the included vocal measures

are meant to control for the variation contributed by different speakers and their choice of words.

Pitch, the primary dependent measure of this study, was calculated using the autocorrelation method (Boersma, 1993). The Harmonics-to-Noise ratio was calculated to serve as an indicator of voice quality (Boersma, 1993). Originally intended to measure speech pathology (Yumoto, Gould, & Baer, 1982), the Harmonic-to-Noise ratio is included to account for the unique speaking characteristics of different participants (measured in dB and larger values reflect higher quality).

Vocal Intensity was calculated to partially control for the influence of different words and vowels. Vowels that require more open mouth and used in words such as "saw" and "dog" result in 10 dB more than the words "we" and "see" (Ladefoged, 2001). Humans perceive a 10dB increase in intensity as a volume four times as loud. The third and fourth formants were calculated and reflect the average energy in the upper frequency range reflecting specific vowel usage in speech. The fourth formant is regarded as an indicator of head size (Ladefoged, 2001). In order to correct the third formant for the unique resonance characteristics of different speakers, it was divided by the fourth formant. This ratio of third to fourth formant was included to account for the effect high frequency vowels have on overall pitch.

In addition to the vocal measures, the participant's gender, if they were born in the US, spoke English as their first language, answered a stressful

question, or lied were included. All of the selected measures used in this study were meant to be representative of limited individual differences variables that can be readily acquired without extensive biographical or demographic information, consistent with an automated screening scenario.

3.10.2 Results

A multilevel regression model was specified (N=760) using mean pitch as the response variable, the vocal and individual measurements previously described as fixed effects and Subject (N=81) and Question (N=13) as random effects. To reflect the repeated measures experimental design, all measurements in the model were nested within Subject. The full model is reported in Table 8.

To test if the specified model provides a significant improvement to the fit of the data, it was compared to an unconditional model using a deviance-based hypothesis test. Deviance reflects the improvement of log-likelihood between a constrained model and a fully saturated model (Singer & Willett, 2003). The difference in deviance statistics (12,256 – 7,865) = 4,391, greatly exceed the test statistic of χ^2 (14, N=760) = 36.12 at the $p < .001$ level. This allows us to reject the null hypothesis that the specified model does not fit the data.

The primary interest of this model is to explore the relationship between emotional states and vocal pitch. The factors manipulated to evoke emotional responses were the instructions to lie and the asking of stressful questions. To test the hypothesis that pitch is affected by whether participants were answering stressful questions or lying, a deviance-based hypotheses test was conducted and

compared the full model against the full model with the fixed effects of lying and stressful questions removed. The inclusion of lying indicators and stress questions significantly improves the fit of the model to the data, χ^2 (4, N=760) = 177, p <.001.

The average pitch for males was 128.68Hz and for females was 200.91 Hz. By examining the full model coefficients in Table 17, we see a pattern of vocal behavior consistent with previous research. Responding to stressful questions resulted in the predicted increase of pitch, b = 23.58, t(760) = 2.80. In contrast, deceptive vocal responses had a lower pitch than truthful responses, b = -18.14, t(760) = -2.10. This may be because responding honestly was more stressful for participants in this study, particularly when the lies were sanctioned and inconsequential. Additionally, being a native English speaker or born in the United States results in lower pitch. This might be explained by a lower anxiety when being interviewed in one's native language.

The significant interactions between voice quality (Harmonics-to-Noise ratio) and the measures in the model qualify the simple effects for predicting pitch. Specifically, when answering stressful questions, pitch decreases as Voice Quality increases the b = -1.61, t(760) = -2.42 and lying results in higher pitch as Voice Quality increases, b = 1.37, t(760) = 2.06.

Table 17. Model of Pitch as Response Variable

	β	SE β
Fixed Effects		
(Intercept)	-236.59*	86.40
Voice Quality	20.92*	6.41
Female	42.35*	12.25
Stress Question	23.58*	8.39
Born in US	-66.65*	18.15
English First Lang	38.51*	19.43
Lie	-18.14*	8.35
High Freq Vowels	491.70*	107.77
Intensity	0.88	0.53
Voice Quality * Female	3.34*	0.83
Voice Quality * Stress Question	-1.61*	0.67
Voice Quality * Born in US	5.11*	1.34
Voice Quality * English First Lang	-3.67*	1.41
Voice Quality * Lie	1.37*	0.66
Voice Quality * High Freq Vowels	-34.25*	8.86

Note. $p < .05$ *; models were fit by maximum likelihood estimate. All continuous variables mean centered.

To more fully understand Voice Quality, a multilevel regression was specified with Voice Quality as the response variable, stress as a fixed effect and Subject (N=81) and Question (N=13), both modeled as random effects. Stress was measured after the interview when participants reported how nervous, flustered, relaxed, uneasy, and stressed they felt during the interview. These items measured on a 7-point scale were then averaged into a composite ($\alpha = .89$) measuring stress. Reported levels of stress predicted increases in Voice Quality, $b = .66$, $t(722) = 2.92$. A deviance-based hypothesis test comparing the model against the unconditional model reveals that stress provides a significant

improvement to the fit of the data, χ^2 (1, N=722) = 739.31, $p < .001$. In light of these results it appears that both Voice Quality and pitch reflect how stressed a person feels while speaking.

3.11 Discussion

3.11.1 Experimental Results

Mirroring the results of previous studies, the vocal analysis software's built-in deception classifier performed at the chance level (Haddad, et al., 2001). 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 (Rockwell, et al., 1997). However, the measurement of AVJ which is based on average plateau length is novel. Future research should further investigate this measurement and its diagnostic potential to detect cognitive effort or thinking.

3.11.2 Factor Structure and Robustness of Vocal Measurements

The current investigation offers evidence that automated analysis of the voice is no longer out of reach. A newer-generation commercial software program successfully extracted several key features of the voice that could be combined

through multilevel factor analysis into four key dimensions related to thought processes, emotions and cognitive effort. These dimensions in turn successfully discriminated between truthful and deceptive responding. Thus, despite past failures with commercial software and even the failure of the current instrument when its own predictions of veracity were made, the raw measures produced by the system were successful in predicting veracity.

A multilevel factor analysis produced both between-subject and within-subject factor loadings. The factor structure extracted from the estimated within-sample correlation matrix suggests the existence of latent variables measuring Conflicting Thoughts, Thinking, Emotional Cognitive Effort, and Emotional Fear. The Conflicting Thoughts factor consists of four features related to fundamental frequency (which humans hear as pitch): FMain, Fx, FFlic and FQ. These features are thought to tap into deception, embarrassment or conflicting thoughts, stress, and concentration. One might expect this factor to be featured most prominently when detecting deceit (as opposed to other kinds of stressors).

The Thinking factor, comprised of AVJ, JQ, and FJQ, is more related to stress associated with thinking and imagination. It might be more implicated in responding that requires thoughtful deliberation or perhaps fabricating imaginary versus real accounts of events. The Cognitive Effort factor, comprised as it is of elements of JQ, SPJ, and SPT, is thought to tap into a mix of thinking, cognitive, and emotional levels that jointly might be expected when one is expending considerable cognitive effort. Finally, the Fear factor, comprised of

SPJ, FMain and SOS, overlaps somewhat with other factors but should be especially sensitive to the negative emotional state of fear. It might be expected to be more evident when stakes are high and consequences of being detected are dire, as when undergoing an interrogation or facing the possibility of being tortured.

The use of an MFA on repeated measures data provides preliminary support that the factor structure is invariant over time. However, a confirmatory factor analysis should be performed to test the hypothesis that the identified factor structure fits a new dataset. Confirming the factor structure extracted from this controlled experiment on data collected over the phone or in a screening environment would support validity and refute claims that the system is only measuring artifacts of the digitization process.

3.11.3 Validity of Measurements

The results of the present study suggest that the claim that vocal analysis software measures stress, cognitive effort, or emotion cannot be completely dismissed. 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. Additionally, the thinking measurement of AVJ, defined as the average plateau length, was partially explained by response length, stress, age, and cognitive difficulty. Controlling for these variables should reduce within subject variability and improve the ability of AVJ to discriminate truth from deception. This combination of variables supports the validity of AVJ

measuring cognitive effort through micro-momentary speech interruptions. FMain was a highly significant discriminator of deception and was partially explained by the stress measure. Consistent with prior research, it appears stress may have caused higher frequency or elevated pitch.

The emotional measurement SPT, based on the average number of thorns, was partially accounted for by emotional and social sensitivity. Participants high on emotional or social sensitivity are more likely to be emotionally affected by the interaction and SPT could be reflecting this. The SOS measure, which reflects fear or hesitation, was partially accounted for by stress, horizontal individualism, and self negative face. Taken together, this pattern of variables is consistent with the expectation that individuals who report high individualistic tendencies would perceive the interaction as more invasive and thus, be more hesitant to respond or comply.

3.11.4 Predicting Deception

The logistic regression provided the best classification of deception using the vocal measures. However, only when properly tuned did perform optimally, meaning that a simple cutoff of 50% to determine deception results in a lower TPR to FPR ratio.

When evaluating the results by question, the logistic regression and SVM classifiers do not contradict the claim that the system works optimally during excited or stressful conditions found in the real world. The best performing classifier, the logistic regression, had the highest prediction accuracy of 62.8% on

the question “Did you ever do anything you didn't want your parents to know about?” This question appeared to successfully evoke emotional reactions and thoughts in participants, which are conditions that provided the best prediction accuracy.

4 STUDY TWO - THE EFFECT OF COGNITIVE DISSONANCE ON VOCAL ARGUMENTS

4.1 Introduction

This study investigates vocal and linguistic behavior using a more direct manipulation of arousal, affect, and cognitive difficulty for deceivers by inducing cognitive dissonance. A novel variation of the Induced-Compliance paradigm was implemented, requiring participants to make verbal arguments out loud to facilitate vocalic and linguistic analysis.

4.2 Method

4.2.1 Participants

Fifty-two female undergraduate students participated in a study advertised as "Campus Policy Issues." Only female participants were recruited so as to reduce the variance in vocal pitch across subjects. Men have a much lower pitch range (100-130Hz) than females (200-230Hz), requiring more power for the vocal analysis. There is no evidence that cognitive dissonance affects males differently than females.

Prior to recruitment, all participants completed a survey containing a series of 11-point scales (1 = strongly disagree; 11 = strongly agree). Participants indicated their opinion on several campus funding issues, including the critical

item “The university should decrease funding toward facilities and services for people with physical disabilities on campus.” The 52 students who successfully participated were strongly opposed (M = 1.73, SD = 1.25) to a cut in funding for campus services for people with physical disabilities. All participants received course credit for their participation.

4.2.2 Procedure

Upon arrival, participants were told that they were going to complete measures intended to measure how they think and discuss campus policy issues, specifically in conversation. Their first task was to make arguments in support of a cut in funding for people with physical disabilities on campus.

Participants were randomly assigned to one of two conditions, either High or Low choice. In the High choice condition, participants were asked if they would help out the study because the researchers were in need of more arguments supporting cuts. Participants in the Low choice condition, in contrast, were told they had been randomly assigned to make arguments in support of a cut in funding.

Participants were then given one minute to prepare and deliver two arguments in support of a cut in funding into a microphone. These arguments, they were told, would be listened to by a university committee to help them make their final decision on cutting funding to balance the budget. This was emphasized to increase the consequences for their arguments.

Figure 14 depicts the script used by participants when delivering their arguments. This was done to enforce structure in argumentation and facilitate segmentation for vocalic analysis. Also, by stating their names, participants further attached themselves to their message and created a public commitment to it.

> My name is _____ and I represent the University of _____ class of 20__
>
> The first reason the university should decrease funding toward facilities and services for people with physical disabilities on campus is....
>
> The second reason the university should decrease funding toward facilities and services for people with physical disabilities on campus is....

Figure 14. Participant Argument Script

After making their arguments, participants completed a brief survey measuring their attitude towards cutting funding, how much choice they felt they had to decline making the arguments, and were debriefed.

4.2.3 Vocal and Linguistic Processing

All of the participants' arguments were recorded digitally to 48kHz mono WAV files. The recordings were listened to in real-time to manually segment and identify the time points for each of the two arguments. The mean length of each vocal argument was 19.07 seconds (SD = 16.47). All of the vocal recordings where resampled to 11.025kHz and normalized to each recording's peak amplitude. The standard vocal measurements used in this study were then calculated using the

Phonetics software Praat (Boersma, 2002) and LVA 6.50 vocal analysis software (Nemesysco, 2009a).

All of the verbal arguments were transcribed and submitted to automated linguistic analysis using SPLICE and LIWC (Francis & Pennebaker, 1993, Moffitt, 2010, Tausczik & Pennebaker, 2010).

4.3 Results

4.3.1 Manipulation Check

Following the argument recordings, participants responded to a survey item meant to check the efficacy of the choice manipulation. The 11-point scale item (1 = strongly agree; 11 = strongly disagree) asked participants' agreement with the statement "I felt free to decline to state the recorded arguments." As illustrated in Figure 15 participants in the High choice (M = 2.00, SD=1.69) condition reported that they felt more free to decline to make the arguments than participants in the Low choice (M = 4.79, SD=3.76) condition, $F(1,50) = 11.22$, $p < .01$.

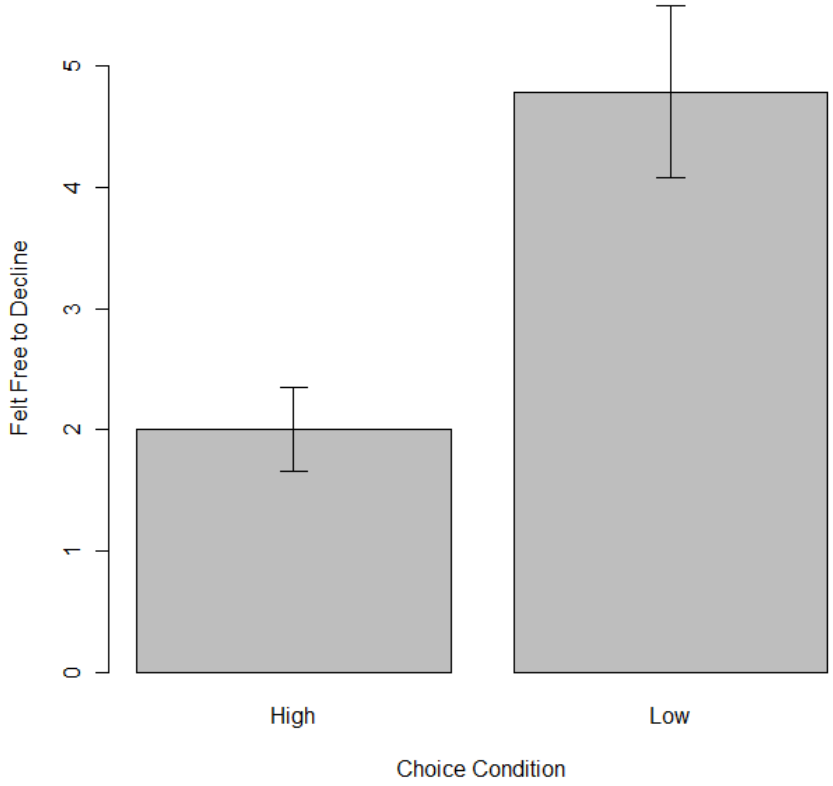


Figure 15. Manipulation Check for High and Low Choice Conditions

4.3.2 Attitude Change

The primary indicator that participants experienced cognitive dissonance, using the classic Induced-Compliance paradigm, is attitude change for High choice participants following their vocal arguments. Participants responded to an 11-point scale item (1 = strongly agree; 11 = strongly disagree) "The university should decrease funding toward facilities and services for people with physical disabilities on campus" immediately after making their arguments.

Participants in the High choice condition were predicted to change their attitudes in order to reduce their cognitive dissonance after freely making counter-attitudinal arguments. Difference scores were calculated between the participants' initial reported attitude towards cutting funding for the physically disabled and their attitude after making arguments in support of a cut in funding (higher numbers reflects greater support for the cut).

Participants in the High choice (M = 3.23, SD=2.53) condition had significantly greater attitude change, $F(1, 50) = 5.91$, $p = .02$, than participants in the Low choice (M = 1.61, SD=2.28) condition.

These choice and attitude change effects successfully replicate the classic Induced-Compliance effect and support the interpretation that participants in the High choice condition experienced more cognitive dissonance and concomitant arousal than participants in the Low choice condition.

4.3.3 Arousal

A repeated measures ANOVA was conducted to compare the effect of the choice manipulation on each of the arousal DVs, mean vocal Pitch (F_0 variation and mean), Intensity (sound pressure level), and Tempo (words per minute) across both arguments (time points one and two) (Ihaka & Gentleman, 1996). There was a significant effect between High and Low choice participants on mean pitch, $F(1,50) = 4.43$, $p = .04$.

Participants in the High choice condition had an average pitch 10Hz greater (High M = 195Hz, Low M = 185Hz). This difference is approximately one

semitone, which in music corresponds to one note higher; 185Hz would be perceived as the note G and 195Hz G#.

There was no significant difference between High and Low choice participants on Pitch Variation, Intensity, or Tempo. The lack of difference between High and Low choice participants could suggest a ceiling where increased arousal does not correspond to ever-increasing tempo and intensity.

All participants made arguments at an average of 158 words per minute with an average intensity of 66.03 dB. The intensity for the typical conversation voice at 1 meter is from 40-60dB and the average speaking tempo is 120-150 words per minute. Participants in this study spoke above the normal threshold for speaking intensity and tempo. This may have been a result of the speaking task that induced a baseline level of arousal or excitement.

Figure 17 depicts the mean vocal arousal measures by choice condition and argument. Apparent from the figure is a decrease in pitch and intensity over time from argument one to two. This is supported by a significant within argument effect for both pitch, $F(1,50) = 4.90$, $p = .03$, and intensity, $F(1,50) = 7.40$, $p < .01$. This suggests that all participants experienced a reduction in arousal over time. This occurred at the same rate for both Low and High choice participants as there were no significant interactions between choice condition and argument.

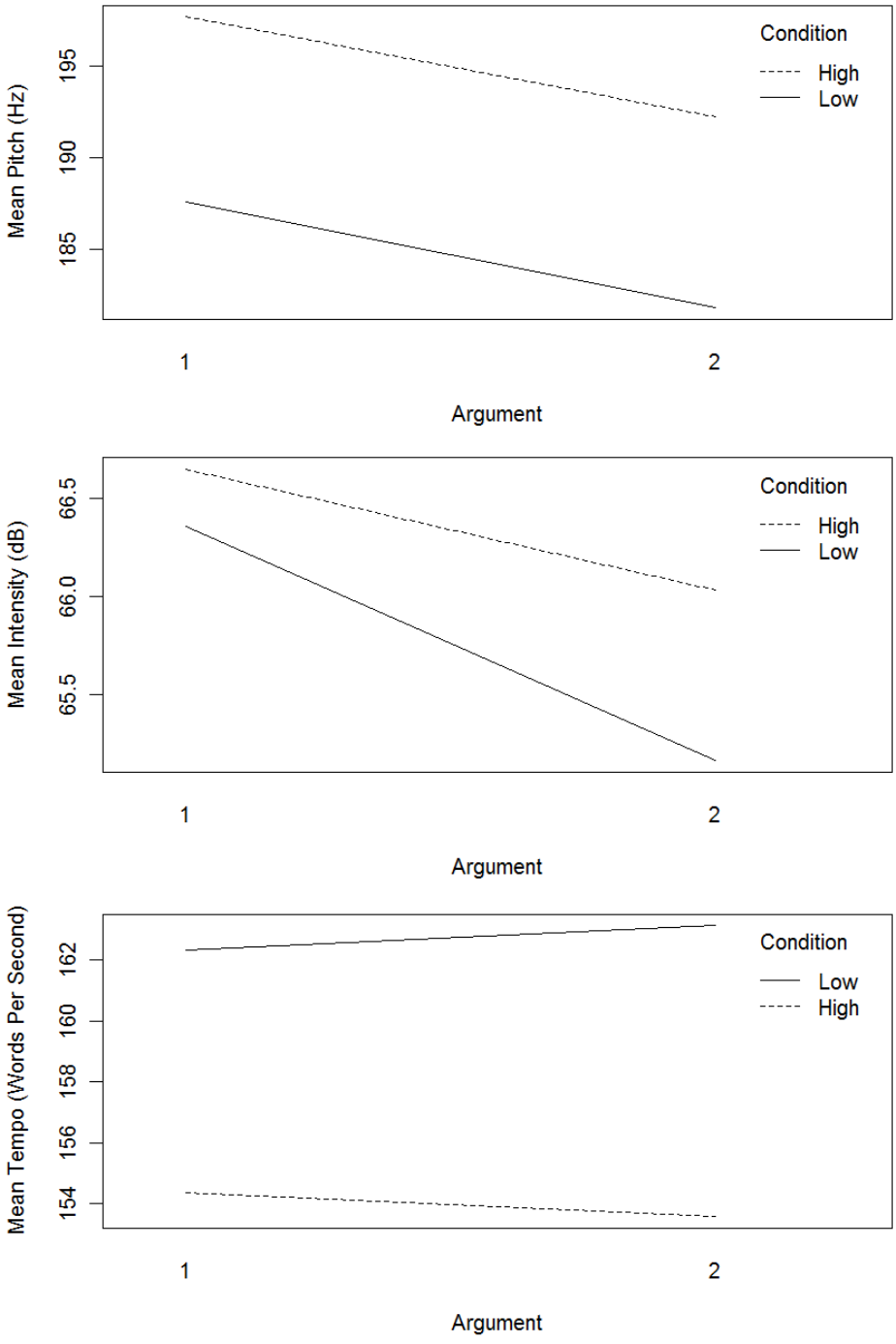


Figure 16. Mean Pitch, Intensity, and Tempo By Choice and Argument

4.3.4 Cognitive Difficulty

The vocal measures of cognitive difficulty, response latency (time in seconds from start of argument after stating stem) and nonfluency-to-word ratio were submitted to a repeated measures ANOVA. There was a significant effect between High and Low choice participants on response latency, $F(1,50) = 4.13$, $p < .05$ and no significant effect on nonfluencies. Participants in the High choice condition ($M = 1.26$s) took nearly twice as long as those in the Low choice condition ($M = 0.65$s). This suggests that in addition to arousal, cognitive dissonance induced participants also experience additional cognitive load and performance reductions when making their arguments.

Both High and Low choice participants experienced increased cognitive load when stating their second argument. This is supported by a significant within argument effect for both response latency, $F(1,50) = 4.53$, $p = .04$, and nonfluencies, $F(1,50) = 4.03$, $p = .05$. Participants had more difficulty making their second argument, likely because their limited preparation time (1 min) was used up on their first argument.

Examining the interaction plot for response latency in Figure 17 we find the difference between High and Low choice participants occurs entirely in the second argument. This suggests that argument main effect on response latency should not be interpreted. Including only the interaction term (Condition x Argument) a repeated measures ANOVA reveals a significant interaction, $F(2, 50) = 4.17$, $p = .02$.

Both High and Low choice participants started with the same first argument response latency (High M = 0.62, Low M = 0.60). Additionally, High and Low choice increased their response latencies on their second arguments, but High choice participants increased at a higher rate (High M = 1.90, Low M = 0.70). This suggests that cognitive dissonance moderated and increased cognitive difficulty when delivering arguments.

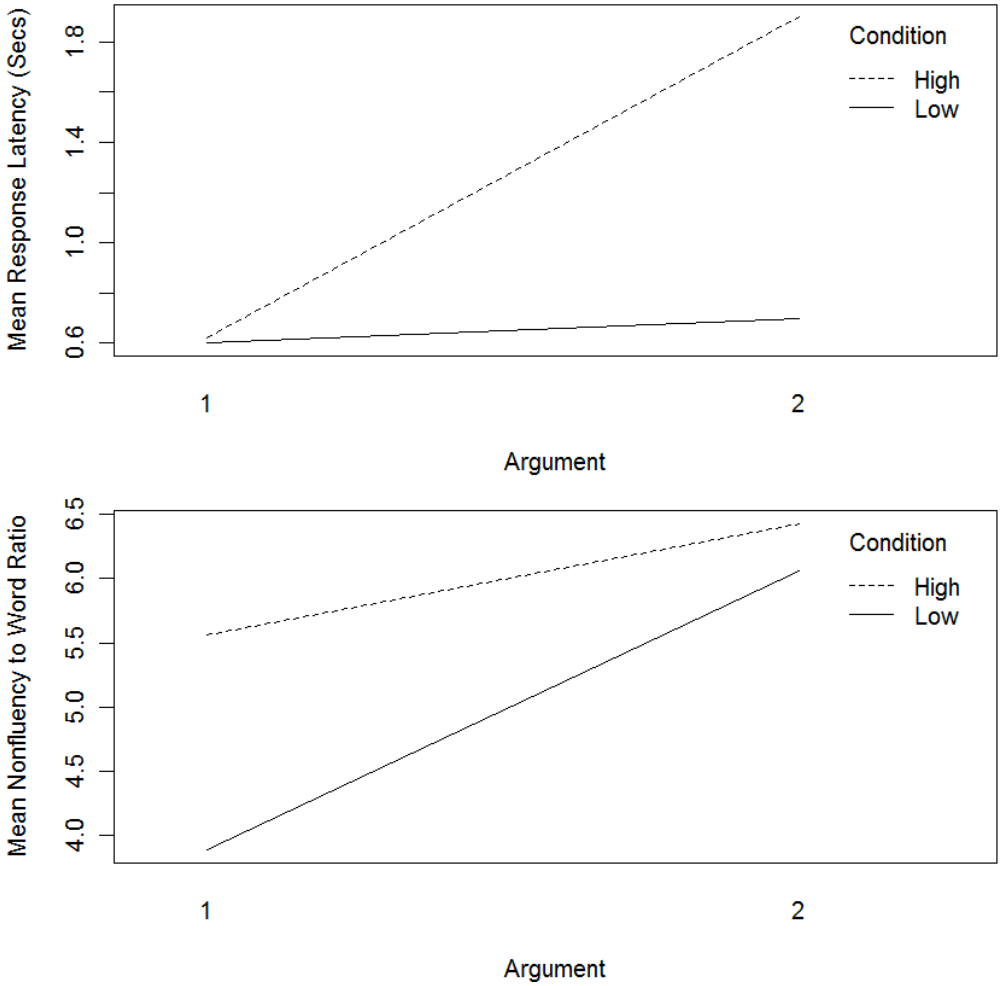


Figure 17. Mean Response Latency and Nonfluencies by Choice and Argument

4.3.5 Emotion

A repeated measures ANOVA (Choice x Argument) was conducted to compare the effect of the choice manipulation on each of the linguistic DVs. The results, summarized in Table 18, reveal that High choice participants spoke in greater Quantity and Certainty and with lower Specificity than Low choice participants. There was no difference in the Immediacy and Affect language between High and Low choice participants.

Participants experiencing cognitive dissonance spoke with 3.2% more verbs, $F(1,49) = 4.36$, $p = .04$, and .6% more modal verbs, $F(1,49) = 8.37$, $p < .01$, when making their arguments. This higher quantity and more certain speech likely reflect the cognitive dissonance reduction process. High choice participants believed their arguments more in order to remove the inconsistency between their behavior counter-attitudinal argument) and beliefs (disagreement with funding cuts).

In congruence with previous linguistic deception studies, High choice participants spoke with 2.3% less spatial Specificity, $F(1,49) = 7.80$, $p < .01$, and .07% less Imagery, $F(1,49) = 7.51$, $p < .01$. Imagery refers to words that provide a clear mental picture or concreteness to the message.

Table 18. Mean Linguistic Differences (High – Low Choice)

Category	Cues	Mean Difference
Quantity	Words	13.40
	Verb %	3.20*
Complexity	Word Lengths	-0.10
Certainty	Modal Verb %	0.60**
	Modifiers % (Adjectives + Adverbs)	0.01
Immediacy	Passive Verb %	0.01
	Personal Pronoun %	-0.01
Diversity	Lexical Diversity	-0.03
Specificity	Sensory %	0.11
	Spatial %	-2.30**
	Temporal %	0.47
	Imagery	-0.07**
Affect	Positive Emotion %	0.22
	Negative Emotion %	0.02
	Pleasantness	0.00
	Activation	0.00

* p < 0.05, one-tailed. ** p < 0.01, one-tailed.

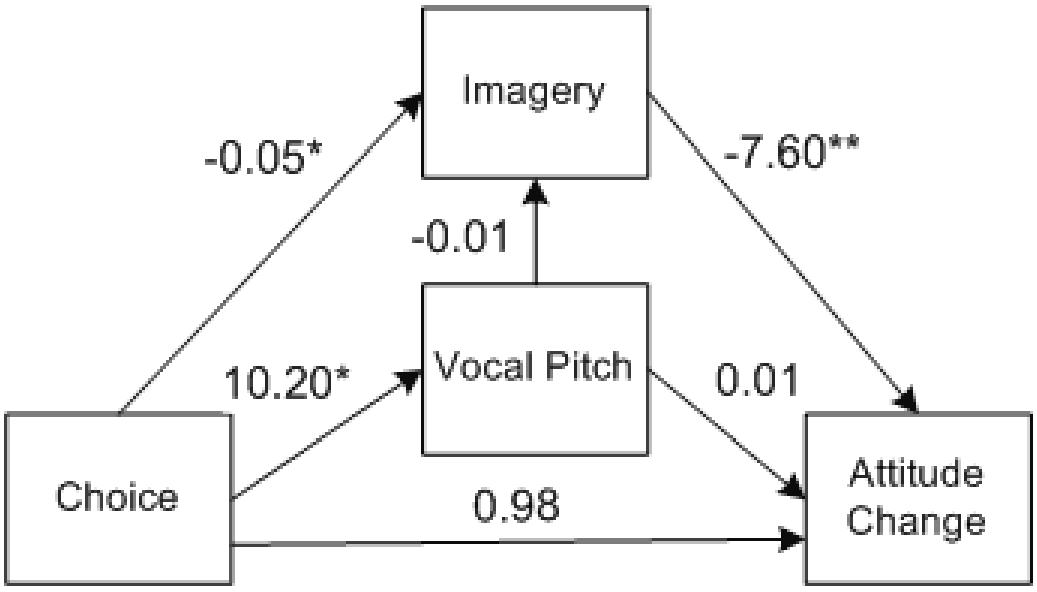
4.3.6 Mediation of Attitude Change

The path model depicted in Figure 18 was specified. The model includes Attitude Change on Pitch, Imagery, and Choice, Imagery on Pitch and Choice, and Pitch on Choice. To test mediation or indirect effects, the model effects and confidence intervals were estimated using Maximum Likelihood Estimation and Bias Corrected Bootstrap sampling (Frazier, Tix, & Barron, 2004, Mallinckrodt,

Abraham, Wei, & Russell, 2006, L. K. Muthén & Muthén, 1998, Shrout & Bolger, 2002) with 10,000 draws.

Vocal pitch, the measurement of arousal, did not mediate attitude change and choice, Indirect Effect = 0.13, 95% CI (-0.23, 0.86). Additionally, vocal pitch did not mediate the relationship between choice and imagery, Indirect Effect = -0.01, 95% CI (-0.3, 0.01).

Imagery significantly mediated choice and attitude change, Indirect Effect = 0.43, 95% CI (.06, 1.08). When participants experienced cognitive dissonance, they included more Specificity in their arguments and changed their attitude more.



Unstandardized path coefficients are shown. N = 51. *p < .05, **p < .01.

Figure 18. Imagery and Pitch Mediating Choice and Attitude Change Model

Imagery may reflect the Specificity of arguments. However, Specificity alone may be to general of a categorization. The word usage measured as

reflecting Imagery may actually correspond to more abstract language. To explore this interpretation the word cloud in Figure 19 displays the most frequent high Imagery words. The larger the word in the word cloud the more it contributed to the overall high Imagery average.

Figure 19. Word Cloud of High Imagery Words Used by High Choice Participants

The word cloud reveals words such as physical, need, service, certain, help, or opportunities as contributing to high Imagery. This could mean that High Choice participants were using more abstract ideas and concepts than Low Choice participants. Moreover, these words are very specific to the study and reflect concepts germane to cutting funding for the disabled. The more High choice participants considered the inconsistency between their personal values (e.g., help, service, need, opportunity) and their actions, the more motivated they became to change their attitude.

4.4 Layered Voice Analysis

The vocal analysis software (Nemesysco, 2009a) used in this study is advertised to provide measurements from the voice indicative of deception, emotion, cognitive effort, and stress. The increased arousal, cognitive effort, and emotions concomitant with cognitive dissonance should be reflected in the measurements calculated by the vocal analysis software. To evaluate the vocal analysis software, repeated measures ANOVA was conducted on each of the calculated variables to compare the effect of the choice manipulation.

To reduce the impact of individual differences in vocal characteristics, speech, and emotional range, the average vocal measurement from the introductory stem (e.g., "My name is Jane Doe and I represent the University…") was subtracted from the measurements taken during each argument. The vocal measurements analyzed reflect differences from their introductory statements, a more neutral value.

All of the reported and analyzed vocal measurements, excluding Lie Probability, were converted to their corresponding z-scores for ease of interpretability and comparison. A description of each of the variables calculated is detailed in Table 1.

4.4.1 Deception Detection

All of the participants in the study were lying; however, participants in the High choice condition were lying, ostensibly, of their own volition. The Lie Probability variable calculated by the vocal analysis software is a variable derived

from a statistical combination of vocal variables that indicate the likelihood that the speaker is being deceitful.

There was no difference in the mean lie probability between High and Low choice participants, $F(1,48) = 0.37$, $p < .54$. Despite the fact that all participants were lying, the grand mean Lie Probability was 36.51% (SD = 10.7). Consistent with the previous deception study, Lie Probability does not correspond with experimentally induced lie behavior. This motivates an analysis of the basic measurements provided by the system instead of the built-in or derived ones.

4.4.1.1 Emotion, Stress, and Cognitive Effort

Based on the earlier deception study, the variables FMain, SOS, AVJ, and JQ are predicted to reflect sensitivity to the cognitive dissonance induced in High choice participants. Table 24 reports the main effects of choice on each of the vocal measurements. Only SOS and FJQ significantly discriminated between High and Low choice participants. However, FJQ was unexpected and after a Bonferroni correction (.05/13=.0038), only SOS remained significant.

Table 19. Main Effect of Choice on Vocal Measurements

	d.f.	F	p
SPT	48	2.49	.12
SPJ	48	0.02	.88
JQ	48	0.01	.96
AVJ	48	3.15	.08
SOS	48	7.88*	<.01
FJQ	48	4.38*	.04
FMAIN	48	0.01	.99
FX	48	0.00	.99
FQ	48	0.07	.78
FFLIC	48	0.09	.31
ANTIC	48	1.41	.24
SUBCOG	48	0.89	.35
SUBEMO	48	0.05	.82

Participants experiencing cognitive dissonance (High choice) had lower SOS levels (M=0.75, SD=1.17) while making counter attitudinal arguments when compared to their neutral or introductory stem, $F(1,49) = 7.88$, $p = <.01$. Figure 20 below illustrates the difference in SOS for High and Low choice participants. Participants in the High condition reduced their SOS (M=-0.39, SD=0.89) while Low condition participants increased their SOS levels when making arguments.

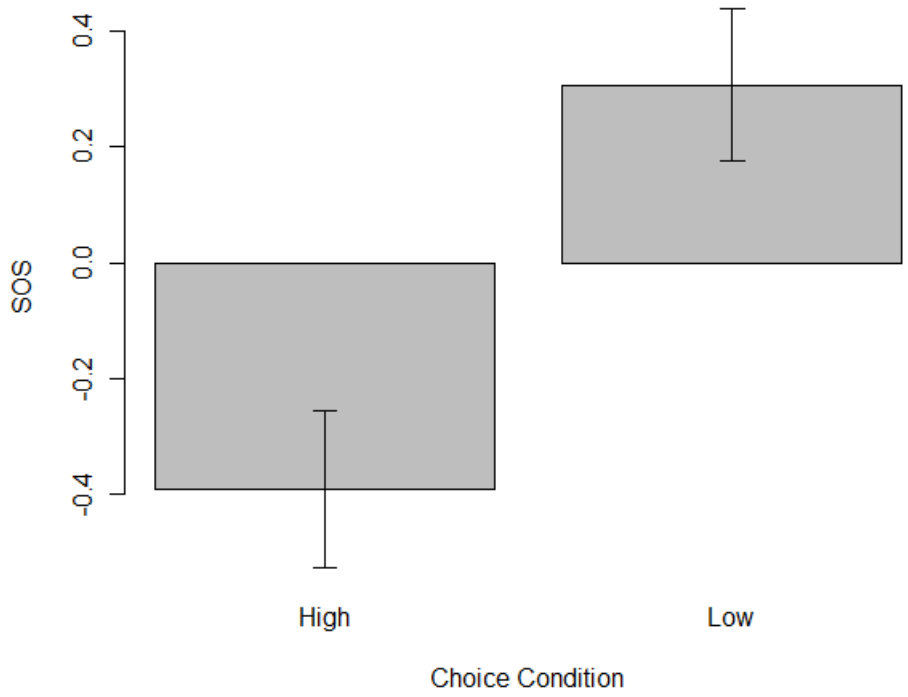


Figure 20. SOS Difference (Argument – Intro Stem) on High and Low Choice Conditions

SOS or "Say or Stop" is documented as being an indication of fear or unwillingness to speak. A white paper published by the software vendor states, “This parameter measures the willingness (excited to speak) or unwillingness (not excited to speak) of a person to discuss an issue and is often a signal to peek into other issues” (Nemesysco, 2009b). Additionally, the Level II Training manual for the LVA 6.50 software frequently cites high levels of SOS as indication of stress or fear (Voice Analysis Tech, 2009).

The reduction in SOS for High choice participants suggests that the utterances before the arguments are made may be the most diagnostic. Including the introduction level (Argument levels: Intro/Name, Argument One, and Argument 2) in a repeated measures ANOVA (Choice x Argument) reveals a significant interaction on Condition and Argument, $F(2,96)=6.21$, $p < .01$. This interaction is illustrated below in Figure 21.

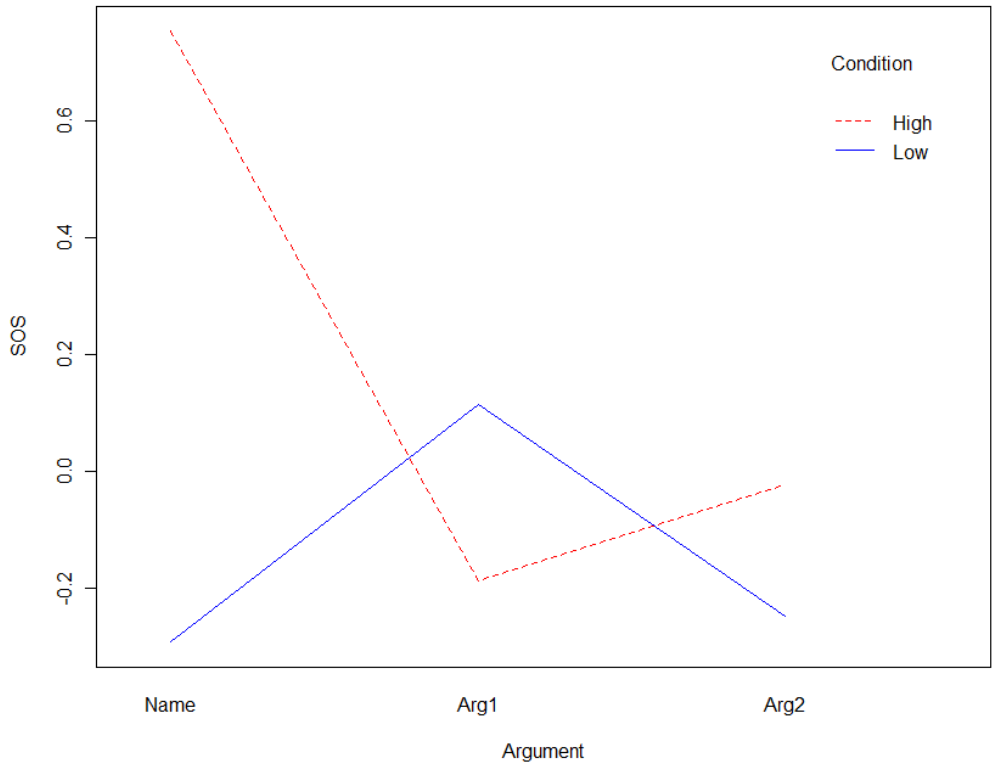


Figure 21. Interaction Between Argument and Condition on SOS

A Tukey HSD pairwise comparison of all Condition x Argument interactions revealed significant differences between High Choice x Argument One and High Choice x Name ($p=.01$), High Choice x Name and Low Choice x

Argument 2 ($p < .01$), and High Choice x Name and Low Choice x Name ($p < .01$). Table 20 below contains a full summary of the pairwise comparisons conducted and reveal that SOS levels were identical for both conditions across all arguments. This supports an interpretation that participants spoke with either elevated SOS (High Choice) or reduced SOS (Low Choice) prior to making their counter attitudinal arguments, but both conditions converged and maintained the same SOS level when actually making arguments.

If SOS measures fear, participants displayed the most (High Choice) or least (Low Choice) fear in anticipation to making their arguments. But, once the argumentation began, participants in both conditions acclimated to the task and displayed the same level of fear. The reduced cognitive demands of making the introductory statement may have left them opportunity to focus on the upcoming task and its consequences (Scher & Cooper, 1989).

Table 20. Full Summary Tukey HSD Pairwise Comparisons

Linear Hypotheses	Estimated Mean	Std. Error	Z-Value	p
Arg1.High - Name.High = 0	-0.94*	0.29	-3.27	0.01
Arg2.High - Name.High = 0	-0.78+	0.29	-2.69	0.08
Name.Low - Name.High = 0	-1.04**	0.27	-3.84	<.01
Arg1.Low - Name.High = 0	-0.64	0.27	-2.35	0.18
Arg2.Low - Name.High = 0	-1.00**	0.27	-3.68	<.01
Arg2.High - Arg1.High = 0	0.16	0.29	0.57	0.99
Name.Low - Arg1.High = 0	-0.10	0.27	-0.38	0.99
Arg1.Low - Arg1.High = 0	0.30	0.27	1.11	0.88
Arg2.Low - Arg1.High = 0	-0.06	0.27	-0.22	1.00
Name.Low - Arg2.High = 0	-0.27	0.27	-0.99	0.92
Arg1.Low - Arg2.High = 0	0.14	0.27	0.51	1.00
Arg2.Low - Arg2.High = 0	-0.22	0.27	-0.83	0.96
Arg1.Low - Name.Low = 0	0.41	0.26	1.59	0.60
Arg2.Low - Name.Low = 0	0.04	0.26	0.17	0.99
Arg2.Low - Arg1.Low = 0	-0.36	0.26	-1.42	0.72

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1

An illustration of the full set of Family-Wise Confidence Intervals is available in Figure 33 located in Appendix B. It is a convenient graphic that can be used to quickly interpret the relationship within the entire family of pairwise comparisons.

4.4.2 Mediation of Attitude Change

The extent to which levels of SOS may account for attitude change is investigated. Two methods of single variable mediation analysis were used, the causal steps approach (Baron & Kenny, 1986) and the Sobel test (Sobel, 1982, 1986).

Following the causal steps outlined by Baron and Kenny (1986), Step 1 revealed a significant relationship between the High Choice condition (Treatment) and Attitude Change (Dependent Variable), b=-1.10, t(48)=-2.29, p=.03. In Step 2, there was a significant relationship between SOS (Proposed Mediator) and the High Choice condition, b=-0.81, t(48)=-3.07, p<.01. However, in Step 3, SOS (Proposed Mediator) did not predict Attitude Change, b=-1.98, t(48)=-1.24, p=.22, when included as an independent variable with the High Choice Condition (Treatment).

To overcome distribution assumptions, a Bias Corrected Bootstrap sampling (10,000 draws) of the coefficients of Causal Step 3 further confirmed that the SOS (Proposed Mediator) coefficient was not significantly different from zero CI (-1.51, 0.12).

Figure 22 illustrates the relationship between the High Choice condition and the results of a Sobel test (Sobel, 1982, 1986) that yielded a z-value of 1.10, p=0.27. Consistent with the causal steps method, SOS appears to only be moderated by Choice condition (i.e., cognitive dissonance) and does not attenuate or reflect attitude change.

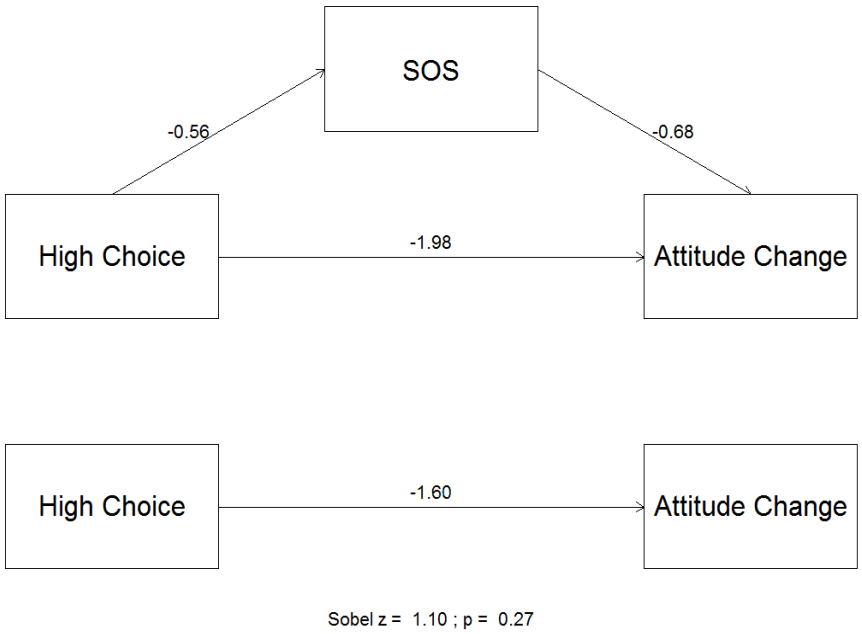


Figure 22. SOS Mediation of Attitude Change

4.5 Discussion

This research represents a first step towards addressing the current gap in our understanding of the arousal and affect process during cognitive dissonance.

Participants experiencing more cognitive dissonance spoke in a higher vocal pitch, an indicator of negative arousal. The existence of arousal was further supported by the reduction in performance, measured by response latency, when High choice participants delivered their second argument.

Cognitive dissonance caused participants to speak with higher linguistic Quantity and Certainty. This likely reflects their formation of cognitions that support the cut in funding. High choice participants spoke with less Specificity

when making arguments. The reason that dissonance caused less concrete language is unclear. Arousal, measured by pitch, did not have an effect on Specificity language.

Pitch, the primary measure of arousal in this study did not mediate attitude change. However, Imagery, the measurement of Specificity, did significantly mediate attitude change.

The relationship between Imagery and dissonance should be explored further. One interpretation of these data is that less Imagery reflects more abstract and high level language such as ideas or concepts instead of objects. Concepts or ideas have a greater likelihood of being more important to the self. More involvement of the self in the argument would have increased their motivation to reduce dissonance and the inconsistency between their actions and beliefs (Stone & Cooper, 2001).

In addition to exploring cognitive dissonance, this study further supported IDT's prediction that motivation and arousal matters when measuring vocal and linguistic lie behavior. Unmotivated liars or Low choice participants actually spoke with more Specificity and were less affected by cognitive effort because of the absence of arousal. Future research should investigate this phenomenon further and compare the vocalics and linguistics measurements with participants making both counter (lie) and congruent (truth) arguments.

4.5.1 Layered Voice Analysis

The Layered Voice Analysis measurement SOS or Say or Stop emerged as a significant moderator during the initial neutral (name) phase. These results correspond with the claim by the vendor of the software that a lower value indicates “indifference or arrogance” while higher levels reflect “fear” towards a speaking topic. Participants who felt they had a choice (dissonance induced) were more fearful and concerned with making the arguments and had higher SOS. In contrast, participants that felt they did not have a choice experienced lower levels of SOS, likely because they felt less concern or possibly indignant from being “forced” to comply.

The extent to which SOS can be used to predict deceptive speech remains a function of how fearful the speaker is. Moreover, the speaker might be fearful for other reasons that could mislead investigators that they are lying. As indicated in the software documentation, it should serve primarily as a cue to inform interview questioning.

For automated deception detection, the existence of fear would need to be contextualized through semantic analysis of the linguistic content. Once contextualized, the importance and relevance of the fearful topic would need to be submitted to an artificial intelligence engine. This potential for automated vocal processing using artificial intelligence is further explored in the next study.

4.5.2 Dynamics of Time

The temporal component of the Layered Voice Analysis variables was also investigated in this study. Specifically, the variable of SOS only revealed moderation during the initial phase of the interaction (Name Phase). Participants in the High Choice condition had elevated SOS levels that normalized during the first and second argument. Low Choice participants had reduced SOS levels that converged with the High Choice SOS levels during the first and second arguments. The initial levels of SOS did not predict actual dissonance reduction (mediation of attitude change), but did serve as a reliable measurement of arousal and possibly fear (as documented in the software manual).

This finding leads to another potential confounding issue, predicted by IDT, the importance of time. People change moods, feelings, and react to interactions dynamically overtime. As such, the same vocal cues could denote different meanings at different times.

The next study investigates the importance of time in measuring the vocalics of trust in addition to a closer to automated instantiation of the herein researched vocal technology.

5 STUDY THREE - VOCAL DYNAMICS OF TRUST OVER TIME DURING INTERACTIONS WITH AN EMBODIED CONVERSATIONAL AGENT

5.1 Introduction

Developing statistical and machine learning models for classifying and predicting emotion and deception in the voice is part of the larger research goal of imbuing computers with emotional intelligence. This area of research is termed Affective Computing (Picard, 2000, Scherer, Bänziger, & Roesch, 2010) and relies on sensors and computing technology to afford computers the ability to affect and perceive human emotions. This effort is complex and necessarily multidisciplinary.

To ensure that efforts in complementary research streams and disciplines are operating within a common framework, Nunamaker et al. (2011) introduced the Special Purpose Embodied Conversational Intelligence with Environmental Sensors (SPECIES) system model. The model in Figure 23 below illustrates the components needed and their interrelations to achieve real-time and automated computing with emotional intelligence.

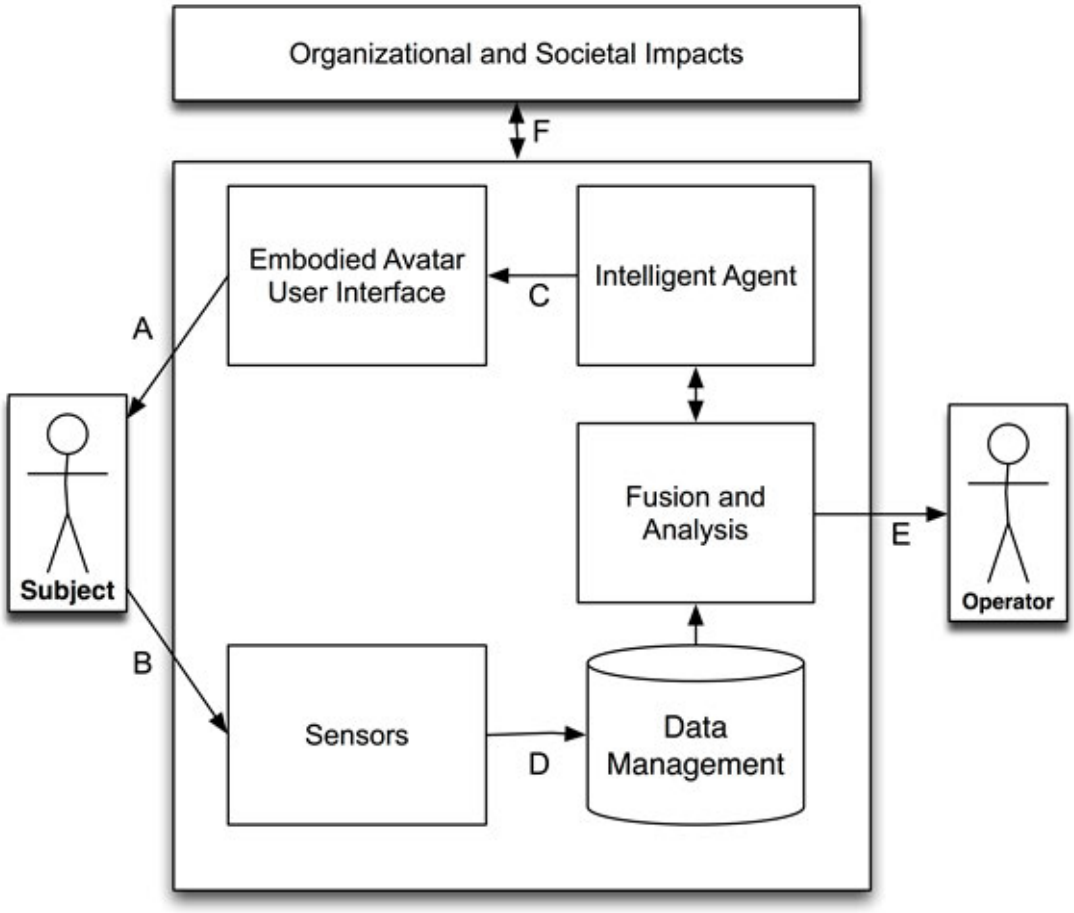


Figure 23. Special Purpose Emboded Conversational Intelligence with Environmental Sensors System Model

SPECIES is a conceptual model for supporting an ECA or automated emotion detection system. The components of the model prescribe interdisciplinary research in (A) Embodied Conversational Agent Signals and Messages, (B) Human Behavior and Psychophysiological Signals, (C) Agent Effectors that Change Appearance and Messages, (D) Data Storage and Segmentation, (E) System Recommendations to the Operator, and (F) Privacy, Ethical, and Policy Considerations.

5.2 Embodied Conversational Agent

The SPECIES model is the basis for the Embodied Conversational Agent developed and used in this study. The ECAs depicted in Figure 24 served the role of an interviewer in a screening scenario ostensibly possessing emotional and artificial intelligence. The human participants that interacted with the system were unaware that the ECA was scripted and believed they were interacting with an intelligent instantiation of the SPECIES model.

Figure 24. Embodied Conversation Agent Interviewer

During the interview the ECA changed its demeanor (Neutral or Smiling) and Gender (Male or Female) as it asked participants 16 questions arranged across four blocks. The question script used by the ECA is provided below in Table 21. Each of the participant's verbal responses to these questions was

recorded with an omnidirectional microphone with a 20-20,000 Hz Frequency Range, -35 dB Sensitivity, and SNR > 62 dB.

Table 21. Questions Asked of Participants by Embodied Conversational Agent

1st Block
1. Please describe in detail the contents of your backpack or purse.
2. I am detecting deception in your responses. Please explain why that is.
3. What will you do after you get through this checkpoint?
4. Please tell me how you have spent the last two hours before coming to this checkpoint.
2nd Block
5. Has anyone given you a prohibited substance to transport through this checkpoint?
6. Why should I believe you?
7. What should happen to a person that unlawfully takes prohibited substances through a checkpoint?
8. Please describe the last trip or vacation that you took.
3rd Block
9. Do any of the items in the bag not belong to you? If so, please describe which items those are.
10. How do you feel about passing through this checkpoint?
11. Please elaborate on why you feel that way.
12. Based on your responses, the previous screeners have detected that you are nervous. Please explain why that is.
4th Block
13. Are there any of your responses that you would like to change? If so, please describe what they are.
14. Is there anything that you should have told us but have not?
15. How do you think that our assessment of your credibility will work out for you today?
16. Why do you think that?

After each question block (i.e., every four questions) participants reported measures of Trust comprised of measurements of Integrity, Ability, and Benevolence adapted from Ohanian (1990) and Reysen (2005). These measurements were comprised of 12 semantic differential word pairs to measure Integrity (Undependable-Dependable, Dishonest-Honest) Ability (Unknowledgeable-Knowledgeable, Unqualified-Qualified, Unskilled-Skilled, Uninformed-Informed, Incompetent-Competent), and Benevolence (Unfriendly-Friendly, Uncheerful-Cheerful, Unkind-Kind, Unpleasant-Likeable).

5.3 Procedures

Upon arrival to the experimental facilities, participants completed consent forms and a demographic pre-survey. Before the interview with the ECA, participants were instructed to pack a duffle bag with innocuous items that someone may bring with them through a security checkpoint. Participants then took the packed bag with them to the ECA station to begin the automated interview.

The ECA began asking the questions and after each activated a surreptitious microphone and waited while the participant responded to the question. The participants then clicked a mouse connected to the system when they were done responding. At each question block, participants would rate their level of trust in the ECA. For each question block, the ECA randomly selected a different gender and demeanor. A Latin Square design was employed to ensure

all possible gender and demeanor combinations were experienced by each participant during the entire interview.

At the end of the interview the ECA informed the participant that they had passed the screening and to proceed through the checkpoint. A waiting experimenter then met and debriefed the participant on the study.

5.3.1 Sample

88 participants were recruited for the study. Most of the participants came from a medium-sized city in the southwestern United States. The mean age of the population was 25.45 years (SD = 8.44). Fifty-three of the participants were male and 35 were female.

5.3.2 Vocal Processing

All of the participants' responses to the ECA's questions were recorded digitally to 48kHz mono WAV files. The mean length of each vocal response was 7.5 seconds (SD = 6.15). All of the vocal recordings where resampled to 11.025 kHz and normalized to each recording's peak amplitude. The standard vocal measurement of pitch (F0) was then calculated using the Phonetics software Praat (Boersma, 2002).

Because of recording equipment error and poor audio quality, 28 participants had unusable audio. There were a total of 866 audio files processed and included in this study. Unlike studies one and two, this study investigates vocal pitch and response duration as independent variables to predict perceived

trust of an ECA over time. The measurements of vocal pitch and duration were averaged across each of the four question blocks.

5.3.3 Measurement of Perceived Trustworthiness

The factors of Ability, Benevolence, and Integrity were submitted to a multilevel confirmatory factor analysis following the protocol suggested by Muthén (1994), Dyer, Hanges, and Hall (2005). Each of these constructs specified in Mayer's model of Trust (1995) were modeled with paths to a latent variable of Trust.

Both between and within subject correlation matrices were simultaneously extracted and submitted to a confirmatory factor analysis using the Maximum Likelihood with full information method (N=352, Subjects=88). Intraclass correlations (ICC) measure how much variance in a variable is attributable to between subject variance (Muthén, 1991). ICC reflects the proportion of variance that is attributable to between subject variance (calculated as Between Subject Variance / Within Variance + Between Variance).

ICC for the item measurements ranged from .08 to .5, suggesting a high degree of between subject variance that could seriously impact the extraction of factors if subject clustering were ignored. An RMSEA of .05 and CFI of .974 ($\chi^2(83)=159.29$, p=<.001), indicated that the measurement of trust was a good fit to these data. The significant χ^2 test likely resulted from the over powered test because of the large sample size (Bollen 1989). Figure 25 illustrates the final measurement model of Trust and the factor loadings for the within subject

correlation matrix. A composite measurement of trust was calculated using the measurement model and used as the primary dependent measure for this study.

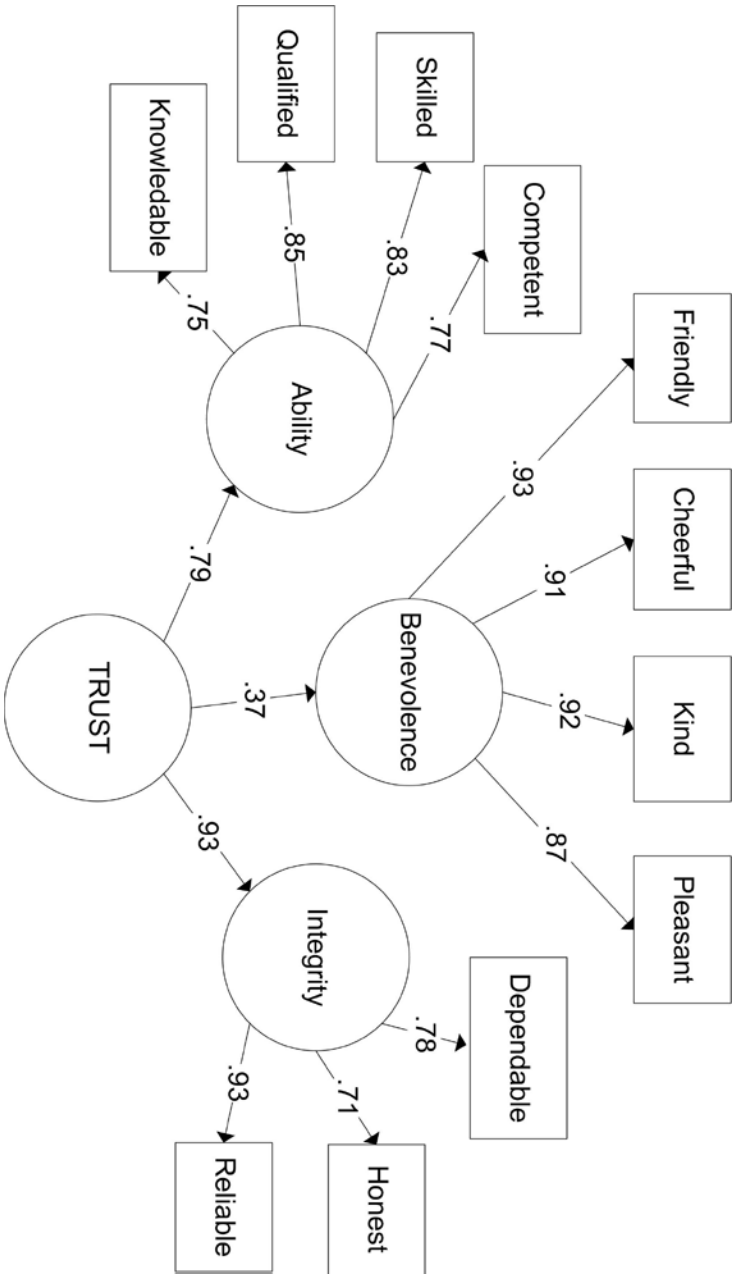


Figure 25. Confirmatory Factor Analysis of Trust based on Within Correlation Matrix

5.4 Results

5.4.1 Time and Trust

To assess the relationship between trust and time, a multilevel growth model was specified with trust as the response variable (N=218) regressed on completion time (in seconds) and average question response duration (in seconds) for each question block. To reflect the repeated measure experimental design over time, both time and the Intercept of trust were modeled to vary within Subject (N=60) as random effects.

To test the hypotheses that trust can be predicted by a linear change in time, the specified model was compared to the unconditional means model, which omits any fixed effects using deviance-based hypothesis tests. The difference in deviance statistics was $\chi^2(3,N=218) = 19.17$ and significant at the $p < .001$ level. This allows the rejection of the null hypothesis that time does not predict trust. Allowing random intercepts and time to correlate within subjects did not improve the fit to the data. This means that initial trust levels of participants did not affect the rate of trust change over time.

Examining the coefficients of Model 1 in Table 22 below reveals main effects of time and duration on trust. Participants had an average trust of 4.09 for the avatar at the beginning of the interaction. For every second of interaction with the avatar, trust increased by .04, $t(156)=2.67$, $p<.01$, But, for every second spent answering the avatar's question over the average, 7.6 seconds, reduced trust declined by -0.05, $t(156)= -4.11$, $p < .001$.

Figure 26 below illustrates the relationship between time and trust. The average participant increased his or her trust by .42 (4.51 - 4.09). When participants took 6.1 seconds over the average response time to answer the avatar's questions, their trust was reduced by 0.31. Figure 26 depicts a continuous change for illustration, but the change can occur discontinuously. For instance, a participant could initially respond to the ECA's first question under the average duration (i.e., < 6 sec), but then respond longer than the average duration for the third or fourth questions, which would result in a reduction in trust. The two trajectories illustrated are examples of participants always responding either at or above the average duration.

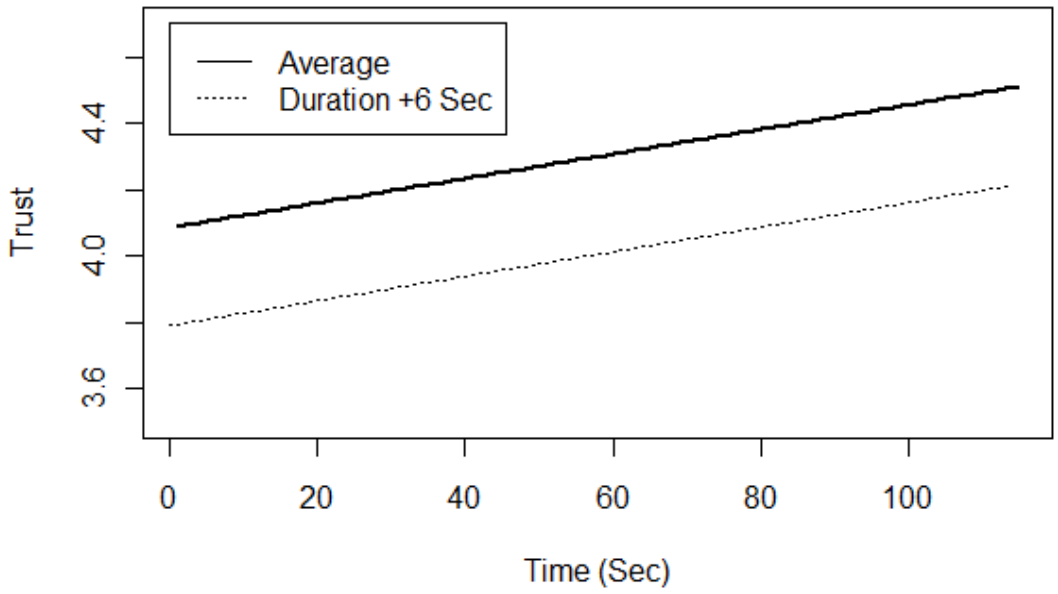


Figure 26. Main Effects of Duration and Time

This result sheds light on the important relationship between time and trust. Participants all increased their trust of the avatar over time, however, when

they spent an inordinate amount of time responding to a question, they lowered their trust of the avatar.

5.4.2 Time, Demeanor, and Gender

To test the hypothesis that the manipulation of avatar Demeanor and Gender affect human trust, the dummy coded variables Avatar Male (1=Male, 0 = Female) and Avatar Smile (1=Smile, 0=Neutral) were added to the growth model. These codes reflect the avatar gender and demeanor participants interacted with prior to reporting their trust levels for each question block.

A deviance hypothesis test comparing the specified model against the growth model reveals a significant improvement to fit, $\chi 2(3,N=218) = 10.79$, $p=.01$. Model 2 in Table 22 below reveal a significant main effect for smiling that increases trust by nearly half a point, $b=0.48$, $t(153)=2.97$, $p<.01$. There was no significant difference between trust of male or female avatars, $t(153)=0.53$, $p=59$, nor any interaction between avatar smiling and gender, $t(153)=-1.18$, $p=24$.

Figure 27 below illustrates the effect of Demeanor on trust. While all participants increase their trust of the avatar over time at the same rate, a smiling avatar increased their trust immediately. The figure displays a hypothetical situation of all smiling avatars versus neutral avatars. However, trust over time could be discontinuous if Smiling and Neutral demeanors were alternated throughout the interaction; trust would rise when the ECA smiled and fall when the ECA had a neutral expression by .48.

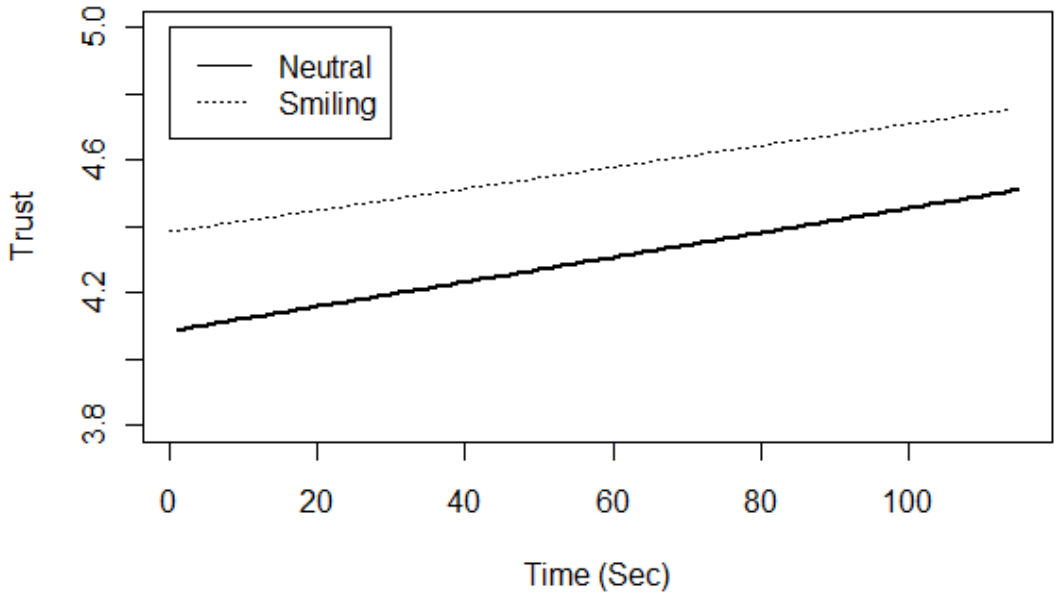


Figure 27. Main Effects of Demeanor and Time

5.4.3 Vocal Pitch, Time, and Trust

To test the hypothesis that vocal pitch predicts trust, the vocal pitch of participants while speaking to the avatar was added as a fixed effect to the growth model. The variable Human Male (Male =1, Female =0) was included to control for the difference in pitch between male and female participants. The deviance hypothesis test revealed a significant improvement of fit to the data, $\chi^2(3,N=218) =8.2$, p=.04. This allows us to reject the null hypothesis that vocal pitch is unrelated to trust.

Model 3 found below in Table 22 details the relationship between vocal pitch, time, and trust. For every 1Hz over the average vocal pitch (M=156Hz), trust drops by .01, t(154)=-2.47, p=.01. This is further qualified by the significant interaction of vocal pitch and time, b=9.3e-05, t(154)=2.19, p=.03. This

interaction implies that overtime the negative relationship between Pitch and trust attenuates. Higher vocal pitch earlier in the interaction is more predictive of lower trust levels.

Figure 28 below reflects two hypothetical trajectories over a 115 second interaction. The average participant speaking at 156Hz starts with an initial trust of 4.02 that increases at a rate of .005 per second up to approximately 4.6 at the end of the interaction. As an illustration, if participants spoke 50Hz above the average pitch (206 Hz), they would have a lower initial trust level of 3.42, but overtime the inverse relationship between vocal pitch and trust attenuates towards equilibrium of trust.

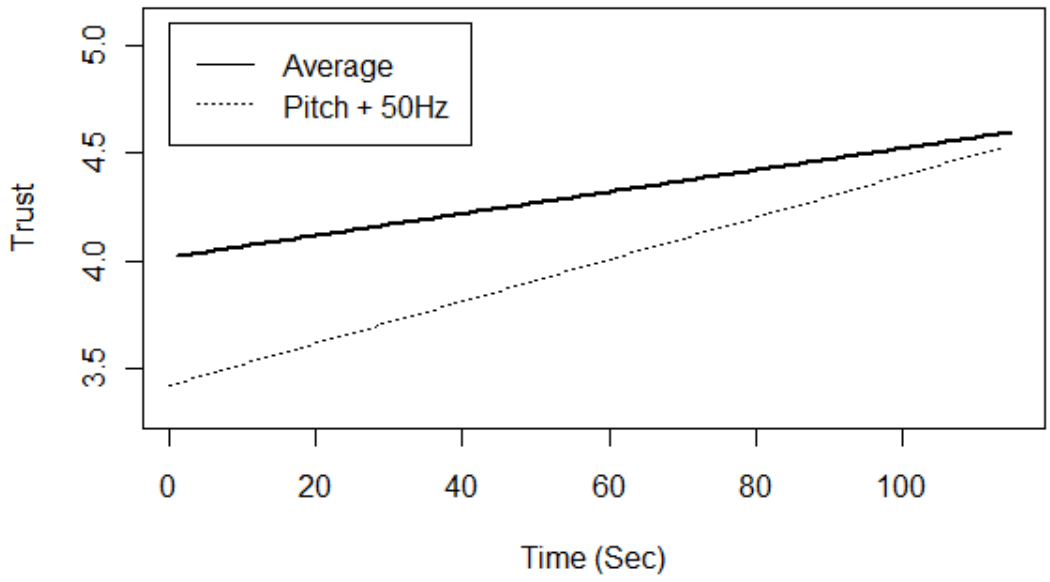


Figure 28. Main Effect and Interaction of Vocal Pitch and Time

5.4.4 Final Model of Trust

A final model was specified that includes avatar time, duration, demeanor, vocal pitch, participant gender, and the variable No College (No College = 1, At Least Some College = 0) to account for some of the participant variance in trust. Using Deviance based hypothesis tests this model provided a significantly better fit to the data than any of the earlier models and had the lowest model AIC of 568.94. Examining the coefficients below in Table 22, we see that the pattern of the predictors remains the same as discussed earlier, however, there was a significant main effect of No College, b=-.93, t(57)=-2.58, p=.01. Participants who did not have any college trusted the avatar less, but still increased their trust of the avatar at the same rate over time.

Table 22. Comparison of Models Predict Trust (N=218, 60 Subjects)

		Model 1	Model 2	Model 3	Final Model
Fixed Effects					
Initial Status	Intercept	4.09***	3.91***	4.02***	3.93***
Rate of Change	Time (Sec)	0.04**	0.003*	0.005**	0.005***
	Vocal Pitch *Time			9.3e-05*	1.18e-04**
	Duration (Sec)	-0.05***	-0.04***	-0.04***	-0.04***
	Avatar Smiling		0.48**		0.35**
	Avatar Male		0.08		
	Smiling*Male Avatar		-0.26		
	Vocal Pitch (Hz)			-0.01*	-0.01**
	Human Male			-0.59~	-0.47
	No College				-0.93**
Random Effects - Variance Components (Standard Deviation)					
Level-1:	Within-Subject	0.83	0.80	0.79	0.79
Level-2:	In initial status	0.54	0.55	0.57	0.50
	In rate of change	-	-	-	-
Goodness-of-fit					
	Deviance	593.79	583.00	585.58	568.94
	AIC	605.79	601.00	603.58	590.93
	BIC	626.10	631.47	634.04	628.16

~p<.10; *p<.05; **p<.01; ***p<.001. All predictors grand-mean centered except for Male/Smiling Avatar & No College. Intercept can be interpreted as average college educated participant speaking with a neutral female avatar. The average duration was 7.5 sec and vocal pitch was 156 Hz. No variance in rate of change because time points were measured in seconds of interaction not fixed events.

5.5 Discussion

The initial findings demonstrate that the human-to-human measures of trust transfer to human interactions with ECAs. This is consistent with other research that shows that many of the attributes of human-to-human interactions are the same when humans interact with these lifelike artificial agents (Nass & Steuer, 1993, Nass, Moon, Morkes, Kim, & Fogg, 1997). The participants ascribed, in varying degrees, the characteristics of integrity, ability, and benevolence to the computer system and these measures were consistent with the latent variable of trust.

Initially, the vocalic measures show that both vocal measures of pitch and the duration of responding reflected negative perceptions of trust. Participants who took longer to respond and answer questions posed by the agent may have felt obligated to explain themselves to the agent and answer the questions more elaborately. This may have led to increased distrust. Additionally, vocal pitch was inversely related to trust, however, this effect was strongest earlier in the interaction. Vocal measures of pitch reflect arousal that must be contextualized to interpret. Earlier in the interaction, participants were building trust with the agent. However, after a certain point, the arousal may have reflected excitement or another positive state. Alternatively, arousal itself may have declined.

Finally, of all of the individual differences only education-level was significant. Age, gender, and other differences were not. However, participants that did not have any college education had a systematically lower level of

perceived trust. This could be based on several factors including their lack of familiarity with technology, or that they did not view the system as benevolent, or able. This relationship deserves further examination in future studies.

This study investigates vocal measures of trust during an interaction with an ECA when there was no manipulation of trust or emotion. The next study investigates how manipulating the intentions and honesty of human participants affects their vocal responses during ECA security interview questions.

6 STUDY FOUR – VOCAL BEHAVIOR DURING AN AUTOMATED SECURITY SCREENING

Building on study three, a field study was conducted incorporating the ECA interviewer and a vocalic sensor. The ECA conducted rapid screening interviews with participants in a mock airport screening environment. All participants packed baggage prior to their interview, however, some also assembled and packed a bomb.

6.1 Sample

Twenty-nine European Union (EU) border guards participated in a trial of new border technologies. All of the participants spoke English during their interaction with the ECA kiosk, but English was not their first language. The participants were all experienced in primary screening on the border of their respective countries Austria, Belgium, Estonia, Finland, France, Germany, Greece, Latvia, Malta, Netherlands, Norway, Poland, Romania, and Sweden. Of the 29 participants, 22 were male and 7 were female.

6.2 Procedure

Participants were randomly assigned into either the Bomb Maker (N = 16) or Control (N = 13) condition. Participants in the Bomb Maker condition assembled a realistic, but not operational, improvised explosive device (IED) before packing the IED and an assortment of clothes in a travel bag. The IED

assembled and packed by participants is pictured below in Figure 29. Participants in the Control condition did not assemble an IED and only packed clothes in a travel bag. After packing their bags, all participants went directly to the ECA kiosk, in a separate room, for a rapid screening interview. All participants were instructed to successfully convince the ECA of their credibility.

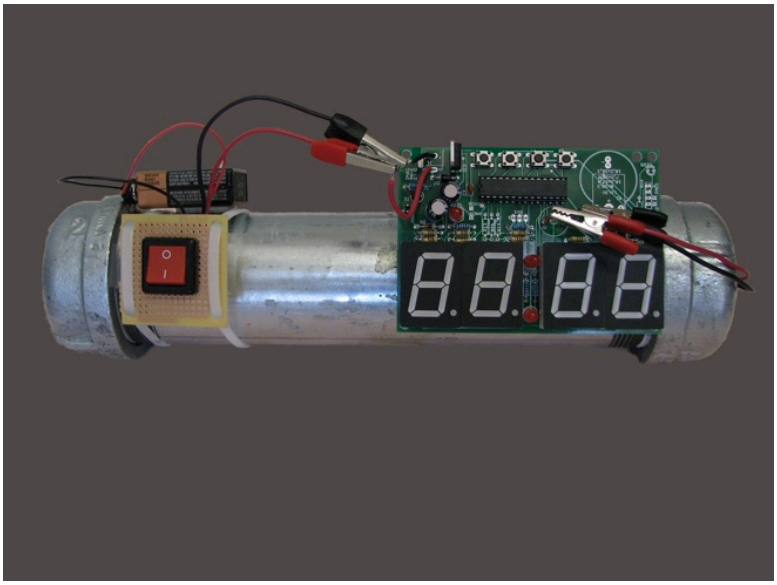


Figure 29. Improvised Explosive Device Carried by Bomb Maker Participants

The ECA conducted the same interview script with participants from the previous study. However, only the male ECA with the neutral demeanor, depicted in Figure 24, conducted interviews. This demeanor was selected because it is perceived as the most dominant and powerful.

Participants completed a realistic interview that did not include any breaks to report their perceptions of the ECA. The same question script from study 3 and detailed in Table 21 was used during the interview. Question five during the interview "Has anyone given you a prohibited item to transport through this

check point?" was of primary interest. Only participants in the Bomb Maker condition would be lying and experiencing the extra concomitant stress and arousal during this question.

The vocal sensor or microphone was integrated into the kiosk and recorded all responses from the participants. All of the recordings were automatically segmented by the ECA during the interview. The vocal recordings in response to question 5 had a mean response length of 2.68 seconds (SD = 1.66) and consisted brief denials such as "no" or "of course not." All of the recordings were processed with the Phonetics software Praat (Boersma, 2002) to calculate the vocal measurements for analysis.

6.3 Results

An ANCOVA with condition (Bomb Maker, Control) as the between-subjects factor and Voice Quality, Female, Intensity, and High Frequency Vowels as covariates, revealed no main effects. All participants had an elevated mean vocal pitch of 338.01 Hz (SD = 108.38), indicating arousal and high tension in the voice. However, there was no significant difference in vocal pitch between the Bomb Makers and Control conditions, $F(1,22) = 0.38$, $p = .54$.

In addition to mean vocal pitch, the variation of the pitch is also reflective of high stress or arousal. The standard deviation of vocal pitch provides a measurement of vocal pitch variation. Submitting the measurement of vocal pitch variation to an ANCOVA revealed a significant main effect of Bomb condition,

F(1,22) = 4.79, p = .04. Participants in the Bomb Maker condition had 25.34% more variation in their vocal pitch than the control condition participants.

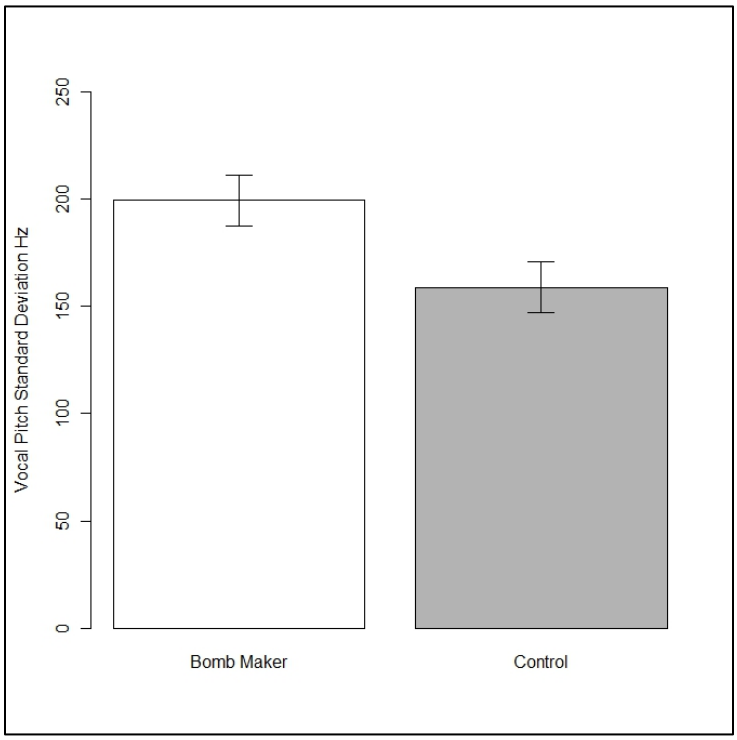


Figure 30. Main effect of Bomb Condition on Vocal Pitch Variation

Table 9 reports the summary of the analysis of covariance. Consistent with the results from study two, the covariates Voice Quality, F(1, 22) = 23.27, p < .01, Gender F(1, 22) = 7.85, p < .01, and Intensity, F(1, 22) = 12.16, p < .01, accounted for the additional variance in pitch variation due other factors such as linguistic content or word choice and accent.

Table 23. Analysis of Covariance Summary

Source	Sum of Squares	df	F
Voice Quality	36,887	1	23.27**
Gender	12,434	1	7.85*
Bomb Condition	7,591	1	4.79*
High Freq Vowels	4,412	1	2.78
Intensity	19,269	1	12.16**
Error	34,872	22	

* $p < 0.05$; ** $p < 0.01$

6.4 Discussion

This study establishes additional ecological validity for an ECA supported by vocalic analysis and investigated with professional border guards. The ECA used the neutral male demeanor for this study, but projecting power and dominance during the entire interaction may not be the best strategy.

The border guard participants admitted feeling nervous during their interaction with the serious ECA. This greater arousal across the whole interaction likely contributed to elevated mean pitch across all participants, leaving little room for variation between conditions. The variance of the pitch reflects both stress and uncertainty. Similar to when a question is posed, inclinations of pitch towards the end of a message connote uncertainty in the English language.

This interpretation of the variance in pitch is supported in Figure 31 that illustrates the pitch contours of two example Innocent and Bomb Maker participants saying "No" to the ECA. A pitch contour reflects the change in pitch over time when speaking. The pitch contour of the example Bomb Maker rises over 50 Hz from the onset to the end of their utterance. In contrast, the Innocent participant maintained a relatively stable pitch with a slight negative slope.

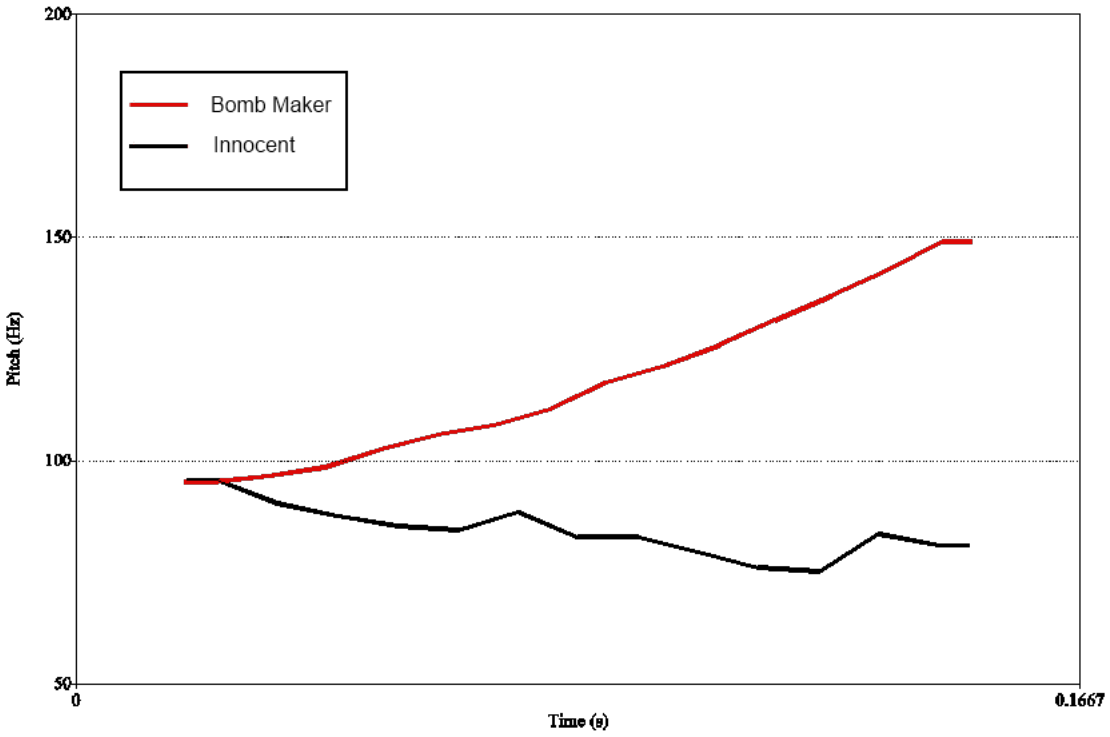


Figure 31. Pitch Contours of Example Bomb Maker and Innocent Participants Saying the Word "No"

Future research in emotional and deceptive vocalics must incorporate more dynamic representations of the voice than just descriptive statistics (e.g., Mean, SD) that obfuscate the trajectory of vocal frequency or amplitude.

7 LIMITATIONS OF THE RESEARCH

The results of the preceding studies are encouraging and support the possibility that computers will one day be capable of automatically detecting emotion and deception using the voice. However, there are limitations to the conducted research that will be summarized in this section.

7.1 Computation of Emotion

Across all the studies in this research arousal and stress predicted vocal behavior. However, arousal alone does not provide enough information about the emotion being experienced. For instance, was the valence of the arousal, negative or positive? If positive, increased vocal pitch could reflect joy. On the other hand, a negative valence could mean fear or anger. Further complicating the matter is a disagreement on how many dimensions of emotion exist, if any. In addition to arousal and valence, a third dimension of power has also been included in models of emotion (Gunes & Pantic, 2010b, Scherer, Schorr, & Johnstone, 2001, Scherer, Schorr, et al., 2001). This dimension allows the distinction between disgust and anger to be revealed. Both are high on arousal and negative in valence. However, someone who is disgusted feels restricted or with less power over a situation than when simply angry (Scherer, Schorr, et al., 2001).

Beyond dimensional explanations, Appraisal Theory predicts a more complex model of emotion, one that predicts that all emotions are first preceded by an evaluation or appraisal of an environmental stimulus that leads to an

emotion (Scherer, Schorr, et al., 2001). This accounts for people reacting emotionally different when experiencing the same stimulus (e.g., romantic breakup). In study three, participants had changes in their perceived trust overtime. Appraisal theory would explain this difference as a result of their different appraisals of the situation and ECA. Initially, they may have had limited experience with the system and were mistrustful. After interacting with the system, they may have found it familiar and similar to human-to-human interaction and evaluated it more trustfully.

Regardless of the emotional theory, emotion is still too complex and rich to fully predict with existing vocal measurements alone. Study two revealed a fuller picture of cognitive dissonance induced speech by including linguistic measurements to augment interpretation of the vocal signal. Future research will need to move towards multimodal behavior and physiological measurements to predict human emotion.

7.2 Use of Linguistics

The linguistic processing and analysis confined to study two was based on simple dictionary based word counts. This is problematic because each word is considered in isolation of the semantics or grammar of usage. This makes it impossible to determine if, for instance, the word “record” is used as a noun or verb. This misclassification made over many ambiguous words introduces non-random error that may have masked important effects.

One way to address word ambiguity is to employ part-of-speech tagging, which is a type of annotation of the linguistic data that disambiguates words according to their structure and grammar. This can be accomplished manually using human coders. However, using human coders is very time intensive and often leads to inter-coder discrepancies that can be difficult to identify and reconcile. There are automated tools for tagging part-of-speech, but most rely on probabilistic and machine learning models based on common written language usage.

Most people speak differently and more informally than they write. This makes existing part-of-speech models less calibrated for automated analysis of transcripts of spoken language. However, incorporating the vocal signal into part-of-speech classification may improve its accuracy significantly. Returning to the example of the ambiguity of the word "record", in English, we place emphasis on the first syllable of a two-syllable word when it is a noun and on the second syllable when it is a verb. For example, "I will re-**cord** your voice" is used as a verb, and "Let's check the **re-**cord" is used as a noun. Combining both the vocal and language information for automated part-of-speech tagging should lead to improved accuracy.

In study two, the pattern of linguistic findings could be partly a result of the experimental design. The increased spatial or imagery language could have been primed by the wording of the argument instructions, which included words such as facilities and physical. This limitation alone does not account for the

systematic differences between high and low choice participants, but caution should be used when trying to generalize the findings to other contexts.

7.3 Short Deceptive Responses

Study one analyzed the vocal behavior of short responses (i.e., one or two words) to facilitate comparison between participants. This one-way and rigid interaction may have artificially tempered cognitive demands or arousal. Since the questions were not open-ended, the participants had no fear of having to elaborate or support their lie to a suspicious interviewer.

The next round of analysis on these data should focus on the more complex open-ended question portion of the experiment. This introduces additional challenges and limitations. For instance, how can the deceptive portion of a message be identified? Deceptive participants might say that they spent last summer in Mexico with family. But, what if they only equivocated and actually went to Mexico two summers ago with friends? The distinction is subtle, but important. An equivocation is harder to detect and should require less cognitive effort to maintain. To partially address this limitation, deception should be analyzed on a scale of honesty and not simply as a binary state.

7.4 Uncertain Speech

The finding that Bomb Makers used more uncertain speech, as measured by increased variance in the pitch, should be investigated further. The Bomb Makers in the study faced no aversive consequences for their actions. Under

these less stressful conditions the uncertainty in their voice may reflect curiosity in the Embodied Conversational Agent and its ability to detect their deception.

The experimental conditions may not be representative of an actual hostile individual attempting to pass a security screening in a high stakes environment. This study should have included a more thorough debriefing or post-survey to ascertain participants' feelings and thoughts during the screening. The statistical model described in study four is likely more predictive of vocal curiosity rather than the hostile intent of a would-be bomber.

The primary motivation for the fourth study was to improve the ecological validity of the research. In comparison to laboratory experiments, this objective was met. However, the scenarios and experimental conditions investigated are still far from reflective of the real security screening environment of law enforcement and national security.

8 FUTURE DIRECTIONS AND RESEARCH

This section outlines recommendations for future directions and research based on the totality of the findings across the four studies.

8.1 Multi-Sensor Measurement of Emotion and Deception

In the introduction to this dissertation, the memory of a time when a parent was angry was invoked. The truth is that we could tell by the voice that they were angry, but it wasn't the voice alone that we used to determine this. We may have noticed they, uncharacteristically, used our full name (linguistic content), had a serious expression (facial gesture), were flushed (increased heart rate and body temperature), or stood in an aggressive posture (body gestures). Just like in our normal interactions, we will need to provide computers with at least as much information as we processes when evaluating emotions. To accomplish this, future research must fuse and analyze multiple behavioral and physiological sensors when modeling emotion and deception.

8.2 Commercial Vocal Analysis Software

This research revealed several Layered Voice Analysis measurements for predicting deception and cognitive dissonance-induced arousal. Despite the checkered past of previous vocal analysis software, current and emerging commercial software should not be immediately dismissed.

The prescription by vocal analysis software vendors that their software only predicts deception or emotion in realistic high stakes contexts should not be met with incredulity. Experimental designs and protocols such as those included in this study should be developed to further explore the validity and potential for predicting emotion on the primitive variables, not just the built-in classifications.

In addition to developing more inventive experimental tests, scientists should be more engaged with commercial software developers. This means transforming the currently adversarial relationship into a collaborative one to cross-pollenate ideas and technical and behavioral knowledge. Many of the vendors of the technologies work in the security industry and can serve as an invaluable resource for emotional behavior under real world conditions.

8.3 Vocal Deception

Future 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 (Buller & Burgoon, 1996). However, most deception experiments and even the polygraph exam focus on behavior difference scores over a set of questions (Vrij et al., 2008). 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 stimuli. Multilevel regression and latent growth curves using structural equation models can be used to model this behavioral change over time (Fitzmaurice, Laird, & Ware, 2004, Moskowitz & Hershberger, 2002, Singer & Willett, 2003). 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.

8.4 Conservative Deception Effect Reporting

Another aspect of this research is the need to account for Type II error when performing multiple simultaneous and post-hoc comparisons. Studies one and two incorporated both Bonferroni and Tukey post-hoc tests on all Layered Vocal Analysis comparisons to avoid Type-II errors and reveal significant effects that can be replicated.

Deception studies are replete with tables of significant effects and cues. Sometimes the directions of the effects are in opposite directions and directly contradict each other. When pressed for explanation, most researchers decry contradictory results as artifacts of experimental design or data collection

procedures. However, this could also be more easily explained by anti-conservatism in statistical tests.

At the $\alpha = .05$ significance level, one out of twenty hypothesis tests will be randomly significant. This, paired with the data driven role of current emotion and deception prediction research, makes the likelihood of Type II error or identifying an effect in the wrong direction a near certainty.

The need to report conservative effect sizes and significant statistics is great when dealing with software such as Layered Voice Analysis variables. The findings will be rightfully contested by an incredulous academy. However, by erring on the side of caution, replicable effects will be revealed that will guide and lead to further scientific discoveries. Deception researchers must be more cautious when accepting significant results and more inclined to disprove their predictions, particularly when testing multiple cues. This will provide greater certainty that an identified statistical effect is accurate, meaningful, and replicable.

8.5 Embodied Conversational Agent as Experimental Confederate

Studies three and four were motivated by the investigation into how people speak when they trust and deceive a fully automated screening system designed in accordance with the SPECIES system model. However, outside of the specific research on Affective Computing phenomenon, using an Embodied Conversational Agent (ECA) as a controlled experimental confederate is feasible and possibly preferable over human ones.

More research will need to be conducted on how much overlap exists between human-to-human and human-to-agent phenomena. For instance, Nunamaker et. al (2011) found that human participants project proscriptive and prescriptive stereotypes depending on the gender and demeanor of the ECA. If enough overlap is confirmed, all behavioral sciences could benefit from even more controlled experiments using tireless and 100% consistent ECA confederates. This would provide additional experimenters for under-resourced research laboratories while removing systematic experimenter effects.

8.6 Integration with Existing Affective Computing Frameworks

Projects such as SEMAINE (2010), an API and Standards-Based framework for Emotion-Oriented systems are technical implementations of automated ECAs with emotional intelligence. SEMAINE differs from the herein discussed SPECIES model in scope and concreteness. The SPECIES is conceptual and serves to organize research efforts at both the high and low level within a common systems model. This includes both technical developments and research on human behavior. In contrast, SEMAINE is primarily technical and provides a modular, cross-platform framework that prescribes data storage standards, communication and messaging protocols, and a middleware platform for integrating new functionality for emotion based systems.

The SEMAINE and SPECIES frameworks reflect solutions to the same complex challenge. Increased collaboration between researchers working on parallel investigations in Affective Computing and emotion detection needs to

occur. This will prevent redoubled efforts, encourage shared data standards, and avoid technological fragmentation.

An analysis module used by SEMAINE for emotion detection uses the OpenSMILE toolkit (2010) developed to consolidate audio input, signal processing, extraction of features, and classification and data mining capabilities. This toolkit provides the functionality of multiple software packages with an API for developing custom applications. One such application is the OpenEAR (2009) toolkit, which relies on the OpenSMILE core to specifically analyze and recognize emotion and affective speech.

A toolkit such as OpenSMILE will be needed to translate the findings from this research into a real-time system for detecting emotion. In order to analyze and classify vocal recordings from experimental data, this research required multiple software packages and lengthy and separate post-processing procedures. However, toolkits such as OpenSMILE and OpenEAR need to go even further and be accessible to non-technical behavioral scientists who cannot write software or process complex vocal datasets. Future research and developments should emphasize development of tools that remove the technical barrier to entry for vocal behavior research. This will lead to a rapid increase in vocal behavior science by extending its investigation beyond a small subset of technically savvy social scientists.

9 CONCLUSION

This dissertation investigated vocal behavior as it occurs naturally while lying, experiencing cognitive dissonance, or receiving a security interview conducted by an Embodied Conversational Agent (ECA). In contrast, the majority of research on vocal behavior has been conducted on acted or performed emotional speech (Banse & Scherer, 1996, Gunes & Pantic, 2010a, Juslin & Laukka, 2003, Pittam & Scherer, 1993, Scherer, Banse, & Wallbott, 2001). Moreover, subsequent vocal research has relied on common emotional behavior databases meant to improve inter-study reliability. These approaches have questionable ecological validity when attempting to translate performed or acted profiles of vocal behavior to the real world.

This research investigated standard acoustic measurements of vocal behavior in addition to commercial vocal analysis software. Specifically, commercial software advertised with the ability to detect deception. Prior research has investigated the lie detection claims of commercial vocal analysis software, but no research to date has delved into the validity of its claims that its primitive vocal measurements reflect emotion and stress.

In study one, to investigate vocal deception and stress, a deception experiment was conducted to identify the standard acoustic and vocal analysis software measurements that predict deceptive speech. 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. Using standard acoustic measurements, both vocal pitch and voice quality were found to be sensitive and predictive of deception and stress.

The results of a multilevel factor analysis and exploratory lasso regression on the commercial vocal analysis software measurements suggest the existence of latent variables measuring Conflicting Thoughts, Thinking, Emotional Cognitive Effort, and Emotional Fear. A logistic regression model using the vocal measurements for predicting deception outperformed the Support Vector Machine and Decision Tree approaches with a prediction accuracy ranging from 46% to 62%. The results of study one suggest that the claim that the vocal analysis software measures stress, cognitive effort, or emotion cannot be completely dismissed.

A common criticism is that low stakes, experimentally sanctioned lying is not representative of the real world. To address this concern, a second study was conducted, that incorporated a novel variation of the Induced-Compliance paradigm to manipulate cognitive dissonance experienced by participants when lying. Cognitive dissonance theory predicts that lying about an issue of importance creates an inconsistency that triggers a negative drive state, manifested physiologically (Elkin & Leippe, 1986, Festinger, 1957, J. M. Olson & Stone, 2005). No research to date has investigated what the vocal markers of

cognitive dissonance are and how this theory can inform current deception theories and thinking.

Participants experiencing more cognitive dissonance spoke with higher vocal pitch, response latency, linguistic Quantity, and Certainty and lower Specificity. Linguistic Imagery mediated the dissonance and attitude change. Imagery was found to be reflective of abstract language usage, which may correspond to more important and self-relevant concepts. Submitting vocal arguments to commercial vocal analysis software revealed that cognitive dissonance induced participants to exhibit higher initial levels of Say or Stop (SOS), a measurement of fear.

The third study explored the vocal behavior of participants while being interviewed by an ECA in a screening scenario. In this study, participants spoke naturally to the ECA and vocal measurements of their voice were modeled to predict trust as a function of time during the course of the interaction. This study was unique from existing vocal behavior research because it employs an extremely structured interview performed by an ECA. Studies that incorporate human interviewers suffer from high variability in questioning technique, delivery, and personality. Moreover, this study represents an important end goal and application of this body of research. Specifically, a real-time processing and analysis of vocal behavior for automated security screenings.

A statistical model was developed that could predict human trust during the interaction using the voice, time, and demographics. All participants

increased their trust of the ECA over time, but they trusted the smiling ECA most. Participants without a college education trusted the ECA the least, but increased their trust of the ECA at the same rate as college educated participants. Vocal pitch only predicted participant trust of the ECA during the early stage of the interview. This suggests that vocal pitch may be primarily measuring arousal, which reflects trust early on, but may have predicted another emotion (e.g., excitement) after trust was established. Future research should expand the self-report measurements collected during the interaction to reflect a greater variety of emotions.

The fourth study contained in this dissertation is a variation of the third ECA study, but investigated in the field. This study was built on the earlier three studies, but emphasized the ecological validity and potential for automated vocal analysis to support law enforcement. In this study an ECA conducted a security screening with actual border guards from the European Union (EU). Some of the border guards were randomly selected to build and attempt to smuggle a bomb past the ECA interviewer. Participants who had built the bomb had 25.3% more variation in their vocal pitch than the control condition participants. The statistical model developed for this study was prototypical of a model suitable for integration into an automated screening system.

In sum, this dissertation makes multiple contributions to our understanding of vocal emotion, dissonance, and deception. There is still much to learn and investigate as we improve our methods, technology, and understanding

of emotion and deception detection. However, this research provides support that the voice is potentially a reliable and valid measurement of emotion and deception suitable for integration into future technologies such as automated security screenings and advanced human-computer interactions.

APPENDIX A – EXTENDED VOCAL ANALYSIS SOFTWARE FIGURES AND TABLES

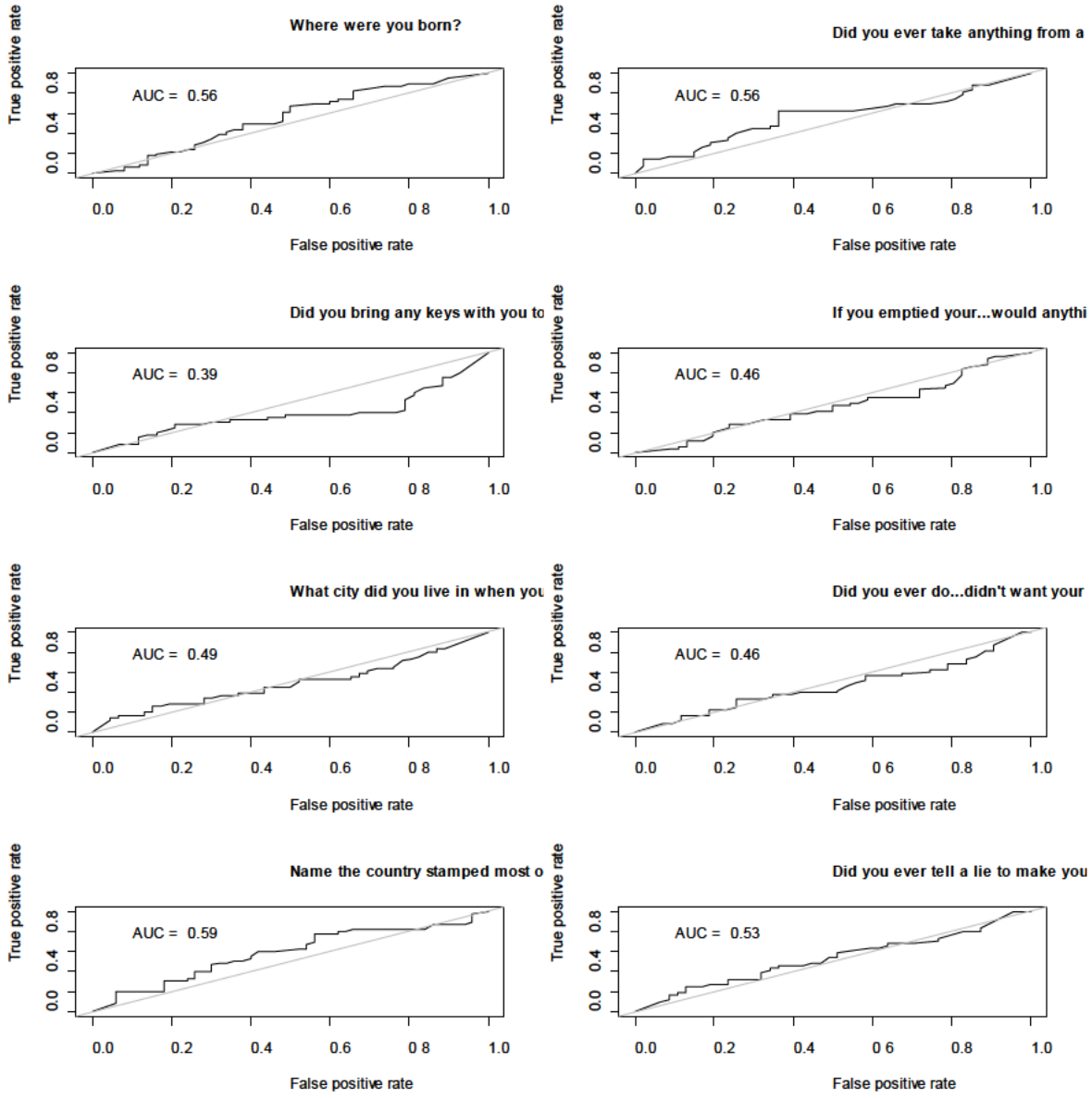


Figure 32. ROC Curves for Vocal Analysis Software Built-In Lie Detection for Each Question

Table 24. Total, Within, and Between Correlation Matrices

	SPT	SPJ	JQ	AVJ	SOS	FJQ	FMAIN	FX	FQ	FFLIC	ANTIC	SUBCOG	SUBEMO
Total correlation matrix													
SPT	1.00												
SPJ	-0.30	1.00											
JQ	-0.11	0.30	1.00										
AVJ	-0.04	0.21	0.56	1.00									
SOS	0.11	0.21	0.27	0.10	1.00								
FJQ	-0.06	0.21	0.43	0.74	0.04	1.00							
FMAIN	0.08	0.08	0.13	0.06	0.15	0.03	1.00						
FX	-0.04	0.09	-0.04	0.01	0.10	0.01	-0.54	1.00					
FQ	-0.04	0.01	-0.04	0.05	0.00	0.02	-0.38	0.31	1.00				
FFLIC	-0.10	0.12	-0.02	0.02	0.09	0.04	-0.52	0.77	0.32	1.00			
ANTIC	0.12	0.02	0.08	0.02	0.08	-0.03	-0.12	0.19	0.19	0.19	1.00		
SUBCOG	-0.10	0.23	0.23	0.33	0.03	0.32	-0.13	0.12	0.05	0.11	-0.02	1.00	
SUBEMO	0.07	0.04	-0.01	-0.04	0.09	-0.05	0.03	0.11	0.12	0.12	-0.02	0.02	1.00
Pooled within-sample correlation matrix													
SPT	1.00												
SPJ	-0.18	1.00											
JQ	-0.17	0.27	1.00										
AVJ	-0.02	0.07	0.52	1.00									
SOS	0.11	0.24	0.25	0.12	1.00								
FJQ	-0.07	0.08	0.36	0.65	0.04	1.00							
FMAIN	0.05	0.12	0.17	0.15	0.15	0.10	1.00						
FX	0.00	0.06	-0.08	-0.01	0.09	0.01	-0.50	1.00					
FQ	-0.04	-0.03	-0.09	-0.02	0.00	-0.03	-0.37	0.29	1.00				
FFLIC	-0.04	0.08	-0.07	-0.03	0.08	0.01	-0.45	0.74	0.29	1.00			
ANTIC	0.01	0.06	0.02	-0.02	0.07	-0.03	-0.07	0.16	0.18	0.15	1.00		
SUBCOG	-0.02	0.04	0.14	0.08	0.04	0.04	-0.04	0.12	-0.07	0.10	-0.02	1.00	
SUBEMO	0.09	0.08	0.03	0.03	0.11	0.01	0.10	0.13	0.12	0.15	-0.04	-0.01	1.00
Estimated between-sample correlation matrix													
SPT	1.00												
SPJ	-0.52	1.00											
JQ	-0.01	0.37	1.00										
AVJ	-0.04	0.56	0.67	1.00									
SOS	0.22	0.01	0.59	0.06	1.00								
FJQ	-0.05	0.58	0.63	0.96	0.03	1.00							
FMAIN	0.13	-0.08	-0.01	-0.20	0.16	-0.19	1.00						
FX	-0.16	0.22	0.19	0.11	0.24	0.07	-0.77	1.00					
FQ	-0.04	0.22	0.25	0.38	-0.14	0.29	-0.51	0.49	1.00				
FFLIC	-0.24	0.27	0.17	0.20	0.18	0.15	-0.81	0.97	0.50	1.00			
ANTIC	0.33	-0.11	0.25	0.13	0.21	-0.02	-0.30	0.39	0.28	0.33	1.00		
SUBCOG	-0.17	0.52	0.41	0.63	0.03	0.69	-0.29	0.23	0.37	0.18	0.00	1.00	
SUBEMO	0.03	-0.08	-0.12	-0.17	0.05	-0.22	-0.20	0.06	0.12	0.04	0.04	0.07	1.00
Intraclass correlation	0.46	0.26	0.25	0.34	0.04	0.29	0.21	0.10	0.09	0.17	0.23	0.61	0.28

Table 25. Comparison of Models Accounting for the Within-Subject Variance in FMain

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed Effects						
Intercept	-0.24*	-0.36*	-0.67*	-0.55*	0.09	0.20
	-(0.08)	-(0.08)	-(0.13)	-(0.14)	-(0.33)	-(0.34)
Female	0.55*	0.54*	0.50*	0.50*	0.52*	0.54*
	-(0.09)	-(0.10)	-(0.09)	-(0.09)	-(0.09)	-(0.09)
Response Length		0.22*	0.30*	0.28*	0.28*	0.28*
		-(0.08)	-(0.09)	-(0.09)	-(0.09)	-(0.09)
Stress			0.07*	0.07*	0.07*	0.07*
			-(0.03)	-(0.03)	-(0.03)	-(0.03)
Truth				-0.23*	-0.21*	-0.21*
				-(0.07)	-(0.07)	-(0.07)
Motivation					-0.07*	-0.06
					-(0.03)	-(0.03)
Social Control						-0.06
						-(0.05)
Random Effects - Variance Components						
Within-Subject	0.11	0.11	0.06	0.06	0.05	0.05
Within-Question	0.01	0.00	0.00	0.00	0.00	0.00
Residual	0.81	0.80	0.82	0.80	0.80	0.80
Goodness-of-fit						
Loglikelihood	-1003.43	-1000.61	-834.06	-829.19	-808.96	-808.27
AIC	2016.86	2013.21	1682.11	1674.38	1635.92	1636.54
BIC	2039.87	2040.83	1713.12	1709.81	1675.58	1680.61

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

APPENDIX B – EXTENDED VOCAL DISSONANCE

FIGURES

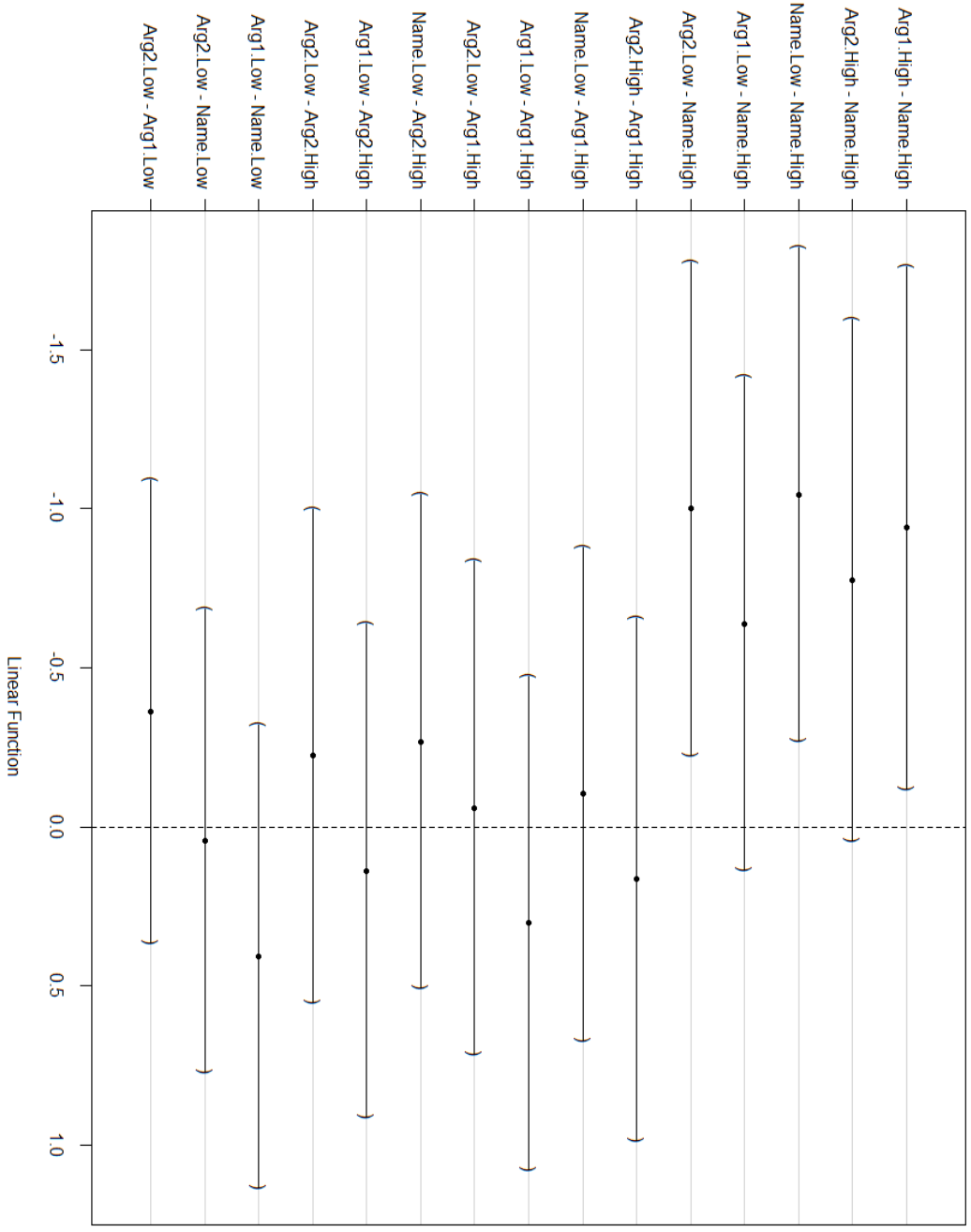


Figure 33. 95% Family-Wise Confidence Intervals of All Interactions

APPENDIX C – VOCAL PROCESSING SCRIPTS AND CODE

Praat Script for Processing Folders of Experimental Audio Files

```
#Normalize amplitude and calculte F0,F1-f4, harmonicity, intensity statistics on folder full of wavs
#This scripts loops through a directory of  directories with each directory a subjectid and wavs inside correspond to question numbers or segments
#The output is a csv file with identifier subjectid and segment/question number
#and the full set of vocal measurments for each corresponding recording

#replace with directory of waves and text segment files (also change output directory below in header and results appending
directory$ = "D:\Vocal Data\ "

#get list of directory names in folder directory$
Create Strings as directory list... directoryList 'directory$'\*
numberOfDirs = Get number of strings

#create header row in results file
fileappend "D:\Vocal Data\AvatarPerception-F0IntensHarm-Results.csv"
SubjectID,QuestionNo,Duration,
F0Mean,F0Median,F0Range,F0Min,F0Max,F0SD,F1Mean,F1Median,F1Range,F
1Min,F1Max,F1SD,F2Mean,F2Median,F2Range,F2Min,F2Max,F2SD,F3Mean,F3
```

```
Median,F3Range,F3Min,F3Max,F3SD,F4Mean,F4Median,F4Range,F4Min,F4Ma
x,F4SD,HarmMean,HarmSD,HarmMin,HarmMax,HarmRange,IntensityMean,In
tensityMedian,IntensityRange,IntensityMin,IntensityMax,IntensitySD'newline$'

#Start looping through directory list starting at index 3 where directories (., ..)
are not included
for idir from 3 to numberOfDirs

  #Select directorylist variable
  select Strings directoryList

  #Retrieve current directory name for list
  dirName$ = Get string... idir

  #Loop through each directory and list file contents
  Create Strings as file list... list 'directory$'\'dirName$'\*.wav

  numberOfFiles = Get number of strings

##########LOOP inside each directory in outer loop and process each AUDIO
FILE inside####### BEGIN
  for ifile to numberOfFiles
```

```
select Strings list

#get current filename based on index in internal loop
fileName$ = Get string... ifile

#subjectid is foldername
subjectid$ = dirName$

#Determine the number of digits in question by location of period in filename
periodIndex = index(fileName$, ".")

#Get current question number from filename by grabbing left most characters before period
questionno$ = left$(fileName$, periodIndex - 1)

#Read in wav
Read from file... 'directory$'\'dirName$'\'fileName$'

#get name of new sound object
soundobname$ = selected$("Sound")

#Normalize peak amplitude to .95 to avoid clipping
```

```
Scale peak... 0.95

#Create pitch object with guasian window (accurate = yes)
To Pitch (ac)... 0 75 15 yes 0.03 0.45 0.01 0.35 0.14 600

#Reselect Sound
select Sound 'soundobname$'

#Create Formant Object
To Formant (burg)... 0 5 5500 0.025 50

#Reselect Sound
select Sound 'soundobname$'

#Create Harmonicity Object harmonics to noise ratio
To Harmonicity (cc)... 0.01 75 0.1 1

#Reselect Sound
select Sound 'soundobname$'

#Create intensity object, default 100 hz and 0 timestep and subtract mean sound
pressure
```

```
To Intensity... 100 0 yes

#get total duration
select Sound 'soundobname$'
duration  = Get total duration

 #Select pitch object
 select Pitch 'soundobname$'

#Set start and end time for statistics (0 0) is whole file
istime = 0
ietime = 0

 #Calculate stats of f0 for selected pitch object and interval
 fomean = Get mean... istime ietime Hertz
 fomin = Get minimum... istime ietime Hertz Parabolic
 fomax = Get maximum... istime ietime Hertz Parabolic
 fosd = Get standard deviation... istime ietime Hertz
 fomed = Get quantile... istime ietime 0.5 Hertz
 forng = fomax- fomin

 #Select formant object
```

```
select Formant 'soundobname$'

#Calculate stats of f1 for selected pitch object and interval (different median function then pitch object)
f1mean = Get mean... 1 istime ietime Hertz
f1min = Get minimum... 1 istime ietime Hertz Parabolic
f1max = Get maximum... 1 istime ietime Hertz Parabolic
f1sd = Get standard deviation... 1 istime ietime Hertz
f1med = Get quantile... 1 istime ietime Hertz 0.5
f1rng = f1max- f1min

#Calculate stats of f2 for selected pitch object and interval (different median function then pitch object)
f2mean = Get mean... 2 istime ietime Hertz
f2min = Get minimum... 2 istime ietime Hertz Parabolic
f2max = Get maximum... 2 istime ietime Hertz Parabolic
f2sd = Get standard deviation... 2 istime ietime Hertz
f2med = Get quantile... 2 istime ietime Hertz 0.5
f2rng = f2max- f2min

#Calculate stats of f3 for selected pitch object and interval (different median function then pitch object)
```

```
f3mean = Get mean... 3 istime ietime Hertz
f3min = Get minimum... 3 istime ietime Hertz Parabolic
f3max = Get maximum... 3 istime ietime Hertz Parabolic
f3sd = Get standard deviation... 3 istime ietime Hertz
f3med = Get quantile... 3 istime ietime Hertz 0.5
f3rng = f3max- f3min

#Calculate stats of f4 for selected pitch object and interval (different median function then pitch object)
f4mean = Get mean... 4 istime ietime Hertz
f4min = Get minimum... 4 istime ietime Hertz Parabolic
f4max = Get maximum... 4 istime ietime Hertz Parabolic
f4sd = Get standard deviation... 4 istime ietime Hertz
f4med = Get quantile... 4 istime ietime Hertz 0.5
f4rng = f4max- f4min

#Select Spectrum object (NEED TO get correct chunks)
#select Spectrum 'soundobname$'

#Calculate Spectrum stats
#speccog = Get centre of gravity... 2.0
#specsd = Get standard deviation... 2.0
```

```
#specskew = Get skewness... 2.0
#speckurt = Get kurtosis... 2.0
#speccentralmoment = Get central moment... 3 2

#Select Harmonicity Object
select Harmonicity 'soundobname$'

#Calculate Harmonicity Stats
harmmean = Get mean...  istime ietime
harmsd = Get standard deviation... istime ietime
harmmin = Get minimum... istime ietime Parabolic
harmmax = Get maximum... istime ietime Parabolic
harmrng = harmmax - harmmin

#Select intensityobject
select Intensity 'soundobname$'

#Calculate stats of intensity for selected regions
intensemean = Get mean... istime ietime
intensemin = Get minimum... istime ietime Parabolic
intensemax = Get maximum... istime ietime Parabolic
intensesd = Get standard deviation... istime ietime
```

```
intensemed = Get quantile... istime ietime 0.5
intenserng = intensemax - intensemin

#append each F0 value to results file
fileappend "D:\Vocal Data\AvatarPerception-F0IntensHarm-Results.csv"
'subjectid$','questionno$',
'duration','f0mean','f0med','forng','f0min','f0max','f0sd','f1mean','f1med','f1rng','f
1min','f1max','f1sd','f2mean','f2med','f2rng','f2min','f2max','f2sd','f3mean','f3med'
,'f3rng','f3min','f3max','f3sd','f4mean','f4med','f4rng','f4min','f4max','f4sd','harm
mean','harmsd','harmmin','harmmax','harmrng','intensemean','intensemed','inte
nserng','intensemin','intensemax','intensesd''newline$'

#Clean up all created objects, leve original directory string, and filename list
select all
minus Strings directoryList
minus Strings list
Remove

endfor #END OF File processing loop (INSIDE)######
```

```
#Clean up all created objects, leve original directory string, filename string will be recreated in above loop
select all
minus Strings directoryList
Remove
endfor
```

REFERENCES

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

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. *Advances in experimental social psychology*, *4*, 1-34.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 219–224.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*(3), 614.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences* (Vol. 17, pp. 97-110).

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Bollen, K. (1989). *Structural equations with latent variables*. Wiley New York.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. Guilford Press New York.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*, 203-242.

Burgoon, J. K. (1983). Nonverbal violations of expectations. *Nonverbal interaction*, *11*, 11–77.

Burgoon, J. K., Ebesu, A., White, C., Koch, P., Alvaro, E., & Kikuchi, T. (1998). The multiple faces of interaction adaptation. *Progress in communication sciences*, 191-220.

Chang, C., & Lin, C. (2001). LIBSVM: a library for support vector machines. Citeseer.

Chen, Y., & Lin, C. (2006). Combining SVMs with various feature selection strategies. *Studies in Fuzziness and Soft Computing*, *207*, 315.

Clark, L. A., & Pregibon, D. (1992). Tree-based models. *Statistical models in S*, 377–419.

Cleveland, W. (1993). *Visualizing data*. Hobart Press.

Damphousse, K., Pointon, L., Upchurch, D., & Moore, R. (2007). *Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting*. U.S. Department of Justice.

DePaulo, B. M., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*, 74-118.

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

deTurck, M., & Miller, G. (1985). Isolating the Behavioral Correlates of Deception. *Human Communication Research*, *12*, 181–201.

deTurck, M., & Miller, G. (2006). Deception and arousal. *Human Communication Research*, *12*, 181-201.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407-451.

Ekman, P., & Friesen, W. V. (1969). Nonverbal Leakage and Clues to Deception.

Elkin, R. A., & Leippe, M. R. (1986). Physiological Arousal, Dissonance, and Attitude Change:: Evidence for a Dissonance-Arousal Link and a. *Journal of Personality and Social Psychology*, *51*(1), 55-65.

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

Eyben, F., Wollmer, M., & Schuller, B. (2009). OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. *Affective*

Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (pp. 1–6).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the international conference on Multimedia* (pp. 1459–1462).

Festinger, L., & Carlsmith, J. (1959). Cognitive consequences of forced compliance. *Journal of abnormal and social psychology*, *58*, 203-210.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis*. Wiley-IEEE.

Francis, M. E., & Pennebaker, J. W. (1993). LIWC: Linguistic inquiry and word count. *Dallas, Texas: Southern Methodist University*.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of counseling psychology*, *51*(1), 115-134.

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press New York.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press New York.

Green, D., & Swets, J. (1966). Signal detection theory and psychophysics.

Gudykunst, W., & Lee, C. (2003). Assessing the Validity of Self Construal Scales. *Human Communication Research*, *29*, 253-274.

Gunes, H., & Pantic, M. (2010a). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99. doi:10.4018/jse.2010101605

Gunes, H., & Pantic, M. (2010b). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99.

Haddad, D., Walter, S., Ratley, R., & Smith, M. (2001). Investigation and evaluation of voice stress analysis technology. Air Force Research Lab Rome Ny Information Directorate.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

Harnsberger, J., Hollien, H., Martin, C., & Hollien, K. (2009). Stress and deception in speech: evaluating layered voice analysis. *Journal of forensic sciences*, *54*(3), 642–50. doi:10.1111/j.1556-4029.2009.01026.x

Holguin, R. (2008, December 12). L.A. Co. gets cutting edge lie detector. Retrieved from http://abclocal.go.com/kabc/story?section=news/bizarre&id=6554064

Hollien, H., Harnsberger, J., & Institute for Advanced Study of Communication Processes. (2006). *FINAL REPORT CIFA CONTRACT – FA 4814-04-0011 Voice Stress Analyzer Instrumentation Evaluation* (p. 62).

Hollien, H., Harnsberger, J., Martin, C., & Hollien, K. (2008). Evaluation of the NITV CVSA. *Journal of forensic sciences*, *53*(1), 183–93. doi:10.1111/j.1556-4029.2007.00596.x

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Inbau, F. E. (1948). *Lie detection and criminal interrogation*. Williams & Wilkins Co.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological Bulletin*, *129*(5), 770.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, 65–135.

Kean, T. H., Hamilton, L. H., Ben-Veniste, R., Kerrey, B., Fielding, F. F., Lehman, J. F., Gorelick, J. S., et al. (2004). *The 9/11 Commission Report*. NATIONAL COMMISSION ON TERRORIST ATTACKS UPON THE UNITED STATES WASHINGTON DC.

Koziol, J., Zhang, J., Casiano, C., Peng, X., Shi, F., Feng, A., Chan, E., et al. (2003). Recursive partitioning as an approach to selection of immune markers for tumor diagnosis. *Clinical Cancer Research*, *9*, 5120.

Ladefoged, P. (2001). *A course in phonetics*. Boston, MA: Wadsworth Publishing.

Lippold, O. (1971). Physiological tremor. *Scientific American*, *224*, 65–73.

Lippold, O., Redfearn, J. W. T., & Vučo, J. (1957). The rhythmical activity of groups of motor units in the voluntary contraction of muscle. *The Journal of physiology*, *137*, 473.

Losch, M. E., & Cacioppo, J. T. (1990). Cognitive dissonance may enhance sympathetic tonus, but attitudes are changed to reduce negative affect rather than arousal. *Journal of Experimental Social Psychology*, *26*(4), 289-304.

Macht, M. L., & Buschke, H. (1983). Age Differences in Cognitive Effort in Recall. *J Gerontol*, *38*, 695-700.

Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology*, *53*(3), 372.

Moffitt, K. (2010). *Structured Programming for Linguistic Cue Extraction*. The Center for the Management of Information. Retrieved from http://splice.cmi.arizona.edu/

Moskowitz, D., & Hershberger, S. (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications*. Lawrence Erlbaum Associates.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, *22*(3), 376.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology*, *25*, 267-316.

Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide. *Los Angeles, CA: Muthén & Muthén*, *2007*.

Nass, C., & Steuer, J. (1993). Voices, Boxes, and Sources of Messages. *Human Communication Research*, *19*(4), 504-527.

Nass, C., Moon, Y., Morkes, J., Kim, E.-Y., & Fogg, B. J. (1997). Computers are Social Actors: A Review of Current Research. *Human values and the design of computer technology*, Center for the Study of Language and Information Lecture Notes (pp. 137-161). Stanford, CA: Cambridge University Press.

National Institute for Truth Verficiation Federal Services. (2011). *CVSA*. Retrieved from http://www.cvsa1.com/CVSA.htm

Nemesysco. (2009a). *Layered Voice Analysis (LVA) 6.50*. Retrieved from http://www.lva650.com/

Nemesysco. (2009b). *Layered Voice Analysis (LVA) Technology White Paper Introduction*. Netania, Israel.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665.

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., & Patton, M. (2011). A System Model for Human Interactions with Intelligent, Embodied Conversational Agents. *Journal of Management Information Systems*.

Olson, J. M., & Stone, J. (2005). The influence of behavior on attitudes. *The handbook of attitudes*, 223–271.

Pallak, M. S., & Pittman, T. S. (1972). General motivational effects of dissonance arousal. *Journal of Personality and Social Psychology*, *21*(3), 349-358.

Park, H. S., & Guan, X. (2006). Cultural Differences in judgement of truthful and deceptive messages. *Journal of Intercultural Communication Research*, *3*, 201.

Picard, R. W. (2000). *Affective computing*. The MIT press.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. *Handbook of emotions*, 185–197.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Reid, J. E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law and Criminology (1931-1951)*, 542–547.

Reise, S., Ventura, J., Nuechterlein, K., & Kim, K. (2005). An illustration of multilevel factor analysis. *Journal of personality assessment*, *84*, 126-136.

Reysen, S. (2005). Construction of new scale: The Reysen Likability Scale. *Social Behavior & Personality: An International Journal*, *33*(2), 201-208.

Rice, W. (1989). Analyzing tables of statistical tests. *Evolution*, *43*, 223-225.

Riggio, R. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology*, *51*, 649-660.

Rockwell, P., Buller, D. B., & Burgoon, J. K. (1997a). The voice of deceit: Refining and expanding vocal cues to deception. *Communication Research Reports*, *14*, 451-459.

Rockwell, P., Buller, D. B., & Burgoon, J. K. (1997b). Measurement of deceptive voices: Comparing acoustic and perceptual data. *Applied Psycholinguistics*, *18*, 471-484.

Rummel, R. (1970). *Applied factor analysis*. Northwestern Univ Pr.

Sagan, C. (1980, December 14). Encyclopedia Galactica. *Cosmos*. 02:24 minutes in: PBS.

Scher, S. J., & Cooper, J. (1989). Motivational basis of dissonance: The singular role of behavioral consequences. *Journal of Personality and Social Psychology*, *56*(6), 899.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76.

Scherer, K. R., Bänziger, T., & Roesch, E. (2010). *Blueprint for affective computing: a sourcebook*. Oxford Univ Press.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, USA.

Scherer, K. R. (1985). Methods of research on vocal communication: paradigms and parameters. *Handbook of methods in nonverbal behavior research* (pp. 136-198). New York: Cambridge University Press.

Schröder, M. (2010). The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems. *Advances in Human-Computer Interaction*, *2010*, 1–21. doi:10.1155/2010/319406

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological methods*, *7*(4), 422-445.

Singelis, T., Triandis, H., Bhawuk, D., & Gelfand, M. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-Cultural Research*, *29*, 240.

Singer, J., & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA.

Skinner, B. F. (1953). *Science and human behavior*. Free Press.

Skolnick, J. H. (1960). Scientific Theory and Scientific Evidence: An Analysis of Lie-Detection. *The Yale Law Journal*, *70*, 694.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25-25.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, *13*(1982), 290–312.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological methodology*, *16*, 159–186.

Stone, J., & Cooper, J. (2001). A Self-Standards Model of Cognitive Dissonance. *Journal of Experimental Social Psychology*, *37*(3), 228–243.

Streeter, L. A., Krauss, R. M., Geller, V., Olson, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, *35*(5), 345–350.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Titze, I. R., & Martin, D. W. (1998a). Principles of voice production. *Acoustical Society of America Journal*, *104*, 1148.

Titze, I. R., & Martin, D. W. (1998b). Principles of voice production. *Acoustical Society of America Journal*, *104*, 1148.

Titze, I. R., & Martin, D. (1998c). Principles of voice production.

Voice Analysis Tech. (2009). *Layered Voice Analysis 6.50 Level II Training*. Madison, WI.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Wiley-Interscience.

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

Walker, J. (2008, May 8). Phone lie detector led to 160 Birmingham benefit cheat investigations. *Birmingham Post*. Birmingham,UK.

Waterman, C. K. (1969). The facilitating and interfering effects of cognitive dissonance on Simple and Complex paired associates learning tasks. *Journal of Experimental Social Psychology*, *5*(1), 31-42.

Whissell, C. M. (1989). The dictionary of affect in language. *The measurement of emotion*, 113-131.

Wright, D. B., & London, K. (2009). *Modern Regression Techniques Using R: A Practical Guide for Students and Researchers*. London: Sage.

X13-VSA Ltd. (2011). *X13-VSA Pro*. Retrieved from http://www.lie-detection.com/

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, *71*(6), 1544-1550.

Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, *29*(5), 703-709.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., & Nunamaker, J. F. (2003). An exploratory study into deception detection in text-based computer-mediated communication. *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on* (p. 10). IEEE.

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