

Evaluation of Voice Stress Analysis Technology

Clifford S. Hopkins
Law Enforcement Analysis Facility
Lockheed Martin IT
Clifford.Hopkins@rl.af.mil

Daniel S. Benincasa
SUNYIT
Utica, NY
Daniel.Benincasa@sunyit.edu

Roy J. Ratley
Law Enforcement Analysis Facility
Lockheed Martin IT
Roy.Ratley@rl.af.mil

John J. Grieco
Air Force Research Laboratory
Rome, NY
John.Grieco@rl.af.mil

Abstract

The Air Force Research Laboratory (AFRL) has been tasked by the National Institute of Justice to investigate voice stress analysis (VSA) technology and evaluate its effectiveness for both military and law enforcement applications. This technology has been marketed as commercially available in computer-based form, and marketed as being capable of measuring stress and in some systems deception. This technology is reported as being easier to use, less invasive, and less constrained in their operation than standard polygraph technology. This study has found that VSA technology can identify stress better than chance with performance approaching that of current polygraph systems. However, it is not a technology that is mature enough to be used in a court of law. We also found that experience and training improves the accuracy over less trained individuals. Lastly, we explored how this technology may become an effective interrogation tool, when combined with polygraph technology.

1. Introduction

Law enforcement agencies are repeatedly exposed to offers of technological solutions that promise to reduce their officers' workload, improve their effectiveness, and/or save lives. While the thought of having the latest and greatest technology might seem enticing for many agencies, due to the constraints resulting from decreasing budgets, most agencies cannot afford the risk associated with the high cost of experimenting with different technology.

Military and law enforcement have in recent years, shown a strong interest in voice stress analysis. One such interest by law enforcement and military personnel is that commercial VSA systems, make claims to being capable of measuring stress, thereby promoting the ability to detect deception or lying. Military organizations, involved in the research of speech and audio technology have long considered the possibility that detection of voice stress can

be used as an adjunct to speech recognition technology in hopes to improve speech recognition capabilities.

Numerous police officers and agencies have been approached in recent years by vendors touting computer-based systems capable of measuring stress in a person's voice as an indication of deception. These systems have been advertised as being cheaper, easier to use, less invasive, and less constrained in their operation than polygraph technology [1]. Statements made by the developers have made claims indicating that a VSA examiner can conduct up to six (6) exams per day while a typical polygraph examiner can only conduct two (2) per day. This is, along with the fact that the systems require less training time for proficiency and are easier to administer than standard polygraph technology make this technology potentially attractive for Law Enforcement. They claim the use of their system is not affected by a speaker's medical condition, age, or consumption of drugs. Voice stress analysis does not require physical attachment of the system to the speaker's body and does not require that answers be restricted to "yes" and "no". Purportedly, according to some vendors, any spoken word or even a groan, whether recorded, videotaped, or spoken in person, with or without the speaker's knowledge, can be used as acceptable inputs to voice stress analysis systems. In this paper, we attempt to perform a realistic evaluation of the systems and address the overall effectiveness of VSA technology.

The significance of VSA technology in military applications is extensive. During military field interrogations of potential informants, VSA can be applied in a manner similar to the application administered by law enforcement: to detect deception. Also, VSA can be used to study the impact stressed speech may have on the performance of speech processing technology, such as speaker identification and language identification. If stress can be detected in voice, speech recognition algorithms can then be modified to adapt to stress speech and improves the recognition performance.

2. Theory Behind Voice Stress Analysis

While most vendors of VSA technology omit specific details on how their systems work, by studying the basic literature, key information can be extracted on the theory behind the technology.

Voice stress analysis originated from the concept that when a person is under stress, micro-muscle tremors (MMT) occur in the muscles that make up the vocal tract which are transmitted through the speech. VSA literature [2] points to a descriptor as the physiological basis for the MMT. This paper describes "a slight oscillation at approximately 10 cycles per second" (i.e. physiological tremors) during the normal contraction of the voluntary muscle. All muscles in the body, including the vocal chords, vibrate in the 8 to 12 Hz range. It is these MMT that the VSA vendors claim to be the sole source of detecting if an individual is lying.

In moments of stress, especially if a person is exposed to jeopardy, the body prepares for fight or flight by increasing the readiness of its muscles to spring into action. This in turn causes the muscle vibrations to increase. According to the Merck Manual [3], "enhanced physiologic tremors may be produced by anxiety, stress, fatigue, or metabolic derangements or by certain drugs. VSA systems claim to measure these tremors transmitted through the speech.

VSA systems can be broken into two separate categories, energy-based systems and frequency-based systems. Here we will discuss both technologies. The majority of the systems evaluated, as reported by the vendors, are based on the detection of the MMT. From the tests and research conducted during Phase I of our study, it was discovered that it was not the measurement of the MMTs that the signal processing was detecting, but a change in the energy of the spectrum envelope between 20 Hz and 40 Hz [4]. When a

voice response is processed through a series of filters, a waveform is displayed that represents the level of stress in the voice. In the event of a non-stress response, the waveform takes on the shape of a Christmas tree (see Figure 1). As stress increases in the subject, the processed waveform takes on a flatter shape. This waveform will flatten out until extreme (hard) stress occurs. See Figure 1(c). Medium stress levels can take on forms that can be a combination of the two extremes. See Figure 1(b). A waveform that shows signs of significant stress is labeled as deceptive when the response corresponds to a relevant question as compared to responses from control questions. This type of technology is known as energy-based VSA.

Frequency based VSA systems derive their results from multiple features, primarily identifying changes within frequency bands and the distribution of the frequencies within those bands. Amir Lieberman of Israel developed one such system evaluated during this study. Lieberman identifies the underlying technology as Layered Voice Analysis (LVA). Dr. J. H. L. Hansen, of the Robust Speech Processing Laboratory, Center for Spoken Language Understanding, University of Colorado at Boulder, also identified a multi-faceted approach to VSA in his research during the initial phase of this study that included the examination of frequencies and their distribution [5]. This system provides textual responses ranging from truth, through confusion to false statement, with various intervening responses. When examining the data a continuum of stress responses was identified and a comparison of the position of the relevant responses on this continuum was made to the position of the control response to determine if the answers were truthful or deceptive.

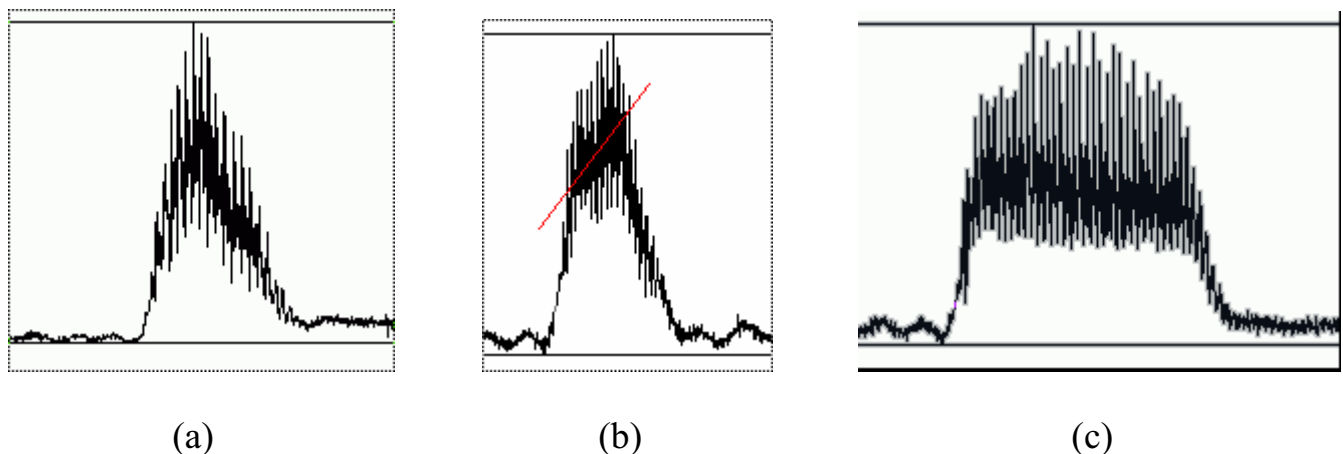


Figure 1. Waveform used to measure stress based on energy in the waveform, (a) little to no stress, (b) medium stress, (c) hard stress

3. Deception Detection

David T Lykken stated in his work that the VSA systems “do not claim to be able to detect deception directly”, but as is done in polygraph, the systems “measure variations in emotional stress”[6]. This theme is consistently affirmed in the open literature, in discussions with operators of Psychophysiological Deception Detection (PDD) systems, and in the training of PDD systems; polygraph and voice stress analyzers, are not “lie detectors”. It is the combination of the results displayed by the PDD system and the evaluation of these results by the trained examiner that determines whether the response is deceptive or non-deceptive. Through a careful pre-test interview and question formulation, an examination is given that will only reflect stress in those areas that are relevant to the matter being tested. Couple this with a thorough post-test interview and the examiner should be able to convince the subject to reveal their deception. Given that the subjects of these examinations are human, there can be some variance expected in their reactions to the questions and in the amount of stress displayed.

The VSA process of detecting deception is similar to that used in polygraph testing. VSA examiners use pre-test interviews to prepare the subject for the forthcoming test. The purpose of this process is to establish the credibility of the test procedures, develop a rapport with the subject, make observations of the subject’s behavior, develop questions for testing and breakdown any internalized barriers that may exist in the subject making admissions.

Pre-Test: It is important for the examiner to establish a presence as an expert in the field and to establish the credibility of the instrument employed to the examinee. This will reassure the innocent that the truth of their innocence will be revealed without causing undo stress that may be reflected in the examination. Conversely, the guilty party will become more concerned that their deception will be revealed which will then enhance the stress displayed in their responses.

A person being deceitful will unconsciously display visual and verbal clues to the examiner. These “leakages” can include crossing of legs or folding of arms as if to protect oneself, inappropriate laughing or remarks when presented with relevant issues, scratching of the nose, covering of mouth, etc. It is during the pre-test interview that the examiner identifies these characteristics as either being a normal behavior or an indication of lying.

Also, during the pre-test interview, the examiner opens the doorway for a confession by removing barriers the subject may have to admitting guilt. Concerns about being ostracized could be reduced, providing the subject with a means of rationalizing their actions. Just by reducing their fear of the consequences will cause some subjects to open up and admit to their guilt. Instructors and operators interviewed during this study reported that well-trained and experienced examiners were able to elicit a confession in the presence of the VSA or polygraph system, without ever turning it on.

It is during the pre-test interview that the examiner reviews the case with the subject, determines what question protocol to use and develops a set of meaningful questions. If the questions do not directly identify the issue at hand, a subject could lie in response to an improperly formulated question and not present stress to the examiner. This is because they may have a picture in their mind of what occurred and some other description may not trigger a stress response. Questions may also be created that bring to mind issues outside of the ones under investigation. This too will cloud their responses. By being able to focus the subject on the relevant points, the operator will improve their ability to obtain tests that accurately identify deceptive or truthful statements.

Test: The process for conducting the actual VSA test is also similar to those used for the polygraph. A question protocol is selected. This protocol is a series of questions arranged in a certain order, sometimes designed for a specific situation or prescribed to by individual VSA system designers. The test protocols can vary from 7 questions to 31 questions in length. The question type relates to the type of question asked and/or it’s purpose for being used.

The three basic question types are Irrelevant (Irr), which is a known truth that should not create any stress in the subject, Control (Con), a known lie, is used to gauge the subjects’ reactions and the Relevant (Rel), which is the issue under investigation. Other question types are extensions of the above and are used for specific functions. An example would be a Sacrifice Relevant (SR) question, where the subjects are asked if they intend to tell the truth during the testing and the question includes a description of the incident. This is done to focus the attention of the subject on the matter at hand.

During the test, the examiner will ask the questions and collect the subjects’ responses. The examiner will then examine the responses for indications of stress, comparing the irrelevant, control and relevant responses to make the determination of deception.

Post-Test: The examination does not stop at the end of the test. Since this testing is not admissible in court and the goal of the examination is to obtain a confession if the subject is guilty, then a post-test interview/interrogation is conducted. In the event that the examiner believes that the subject was deceptive during the test they will begin to probe deeper, becoming more interrogative in nature with the goal of getting an admission from the subject. In the event that the examiner does not believe the subject was deceptive, they will continue discussing the matter, looking for reactions to confirm their findings. Most VSA systems suggest that you not tell the subject that they either passed or failed, but that you tell them that you are going to have another examiner review the charts.

4. VSA Systems Tested

In our study we evaluated the following VSA systems:

VSA-2000 – is a hardware-based system that uses a digital readout to evaluate the stress levels of the speaker. Tests are recorded and the audio input is processed.

Computerized Voice Stress Analyzer (CVSA) - is an energy-based detection system that produces a filtered waveform for evaluation and is computer based. Testing is live and multiple waveforms can be displayed on a single page. The audio is passed through the sound card and is automatically directed to the CVSA system.

Digital Voice Stress Analyzer (DVSA) - is an energy-based system that uses waveforms and is capable of displaying multiple patterns per page. Testing with this system is accomplished from data that is recorded, digitized, and then stored in the computer.

TiPi 6.40 - is a frequency-based system and is the most recent iteration of the TrusterPro. This system automatically segments narrative responses and analyzes each phrase for stress. The system processes live audio in real-time mode and can digitize audio in off-line mode.

Digoenes Lantern - is an energy-based system that utilizes either live audio feeds or digitally stored audio. The waveforms are displayed individually

5. Testing Methodology

In 2003, a report was published by the National Research Council (NRC) of the National Academies with regard to the accuracy of the polygraph test and its use in the screening process [7]. In this report, the NRC identified three main areas for consideration when examining the value of a PDD system: reliability, accuracy and validity. We will use these same criteria for evaluating VSA technology.

Reliability describes the ability of the testing to be replicated under similar conditions. Similar conditions relate to obtaining the same results when administered at different times, places, and involving different examiners and subjects. As a psychophysiological detection tool, VSA does not measure the psychological state of the examinee, but rather the physiological reaction to a perceived threat. Because each human is different and each situation is different, each response to the same problem may be different each time it is presented. This creates difficulty when judging the reliability of VSA (or any other PDD) testing, since reliability is considered to be the ability to replicate the results under similar conditions. From initial testing with the same subject, being asked the same questions, by the same examiner, on different days can and will produce different results. Inter-rater reliability (obtaining the same results from different raters) is often cited in studies; however this is more a function of the reliability of the training to produce operators who can recognize the deceptive patterns rather than the reliability of the instrument to consistently produce the same pattern.

Accuracy and validity are somewhat related when measuring performance. Without validity, accuracy measurements become a weak performance measure. Accuracy, also referred to as criterion validity, is the measure of how correctly the test identifies the truth or

deception. Criterion validity is essential for determining the overall performance of a system, but without construct validity the procedure is nearly useless. Construct validity is the explanatory theories and concepts that are used to support the technology and serves to validate the results. If a technology is producing accurate results, and there is not enough sufficient explanation to back the results, then the results cannot be validated. In this instance, it is the psychophysiological reaction to the stress from being deceitful and the changes in vocalizations resulting from that stress. This reaction follows a chain of events from perception of the threat through the nervous system, to the triggering of the bodies reaction and then returning of the system to homeostasis. This process has been identified and verified through years of research [7].

Prior to performing any tests, the evaluators that were chosen received training on each of the systems, collected data with known results, and then tested the data using the systems evaluated.

Test files were prepared primarily by Law Enforcement Analysis Facility (LEAF) personnel, although, later in the study outside analysts (individuals trained in using the VSA systems that were participating in the study) were enlisted to evaluate the data for Deception Indicated (DI)/ No Deception Indicated (NDI) decisions. Table 1 displays the experience levels of the evaluators and number of test files evaluated.

Table 1. Experience of evaluators

Evaluator	Hours of Training	Status	Law Enforcement Experience In Years	Number of Tests Evaluated
1	192	In-house	26	335
2	64	In-house	N/A	164
3	180	In-house	N/A	18
4	88	Outside agency	30+	17
5	48	Outside agency	30+	17

The audio data collected consisted of recorded truth verification examinations, where bipolar (Yes/No) responses were given. Ground truth was required to identify deceptive/non-deceptive results. This ground truth typically consisted of a confession and some form of corroborating evidence. In the case of the non-deceptive individuals, a confession or arrest of another person or clearing by other means of investigation was sufficient. Of the 135 files collected, ground truth was available for 56 cases. Of this number, 25 were reported as DI and 31 were listed as NDI.

Most of the computer-based VSA systems tested operated on a sampling rate of 11.025 kHz and a bit rate of 8 bits per sample. Audio data was received in a variety of formats. Several files were received via audiotape, which required digitizing. This data was initially digitized at 48 kHz at 16-bit resolution and recorded onto a computer. The

data was then re-sampled to the desired rate (specific to the system being evaluated) and saved as audio wave files. These audio files were then transferred to a laptop computer for use in the evaluation process. Most audio data files were received on CD ROMs or ZIP disks as .wav files. These files were reviewed and re-sampled as necessary.

The CVSA and VSA2000 required direct audio input. The audio files were first recorded to a digital audio recorder (DAT or Mini-Disc format). Based on the results of Phase I of this study, it was found that this type of recorder allows for the best audio replication possible [4]. These recordings were then played into the VSA systems and the vendors' procedures were followed to perform the examinations. Later in the study, using commercial audio processing software as an audio player, one computer was attached directly by cable to the testing computer via the sound ports. It was found that this method provided a more efficient means of visually locating the responses and replaying them as needed. It also eliminated the time required for re-recording the data.

6. Test Objectives and Results

The testing focused on two objectives. The first objective was to test the VSA systems on the audio data files collected where ground truth was available. The objective was not to compare the different systems with each other, but to evaluate the technology. The second test objective was to determine if experience in reading the output of the VSA systems played a significant role in the operator's ability to determine deception.

For the first objective, we tested and evaluated the VSA technology by utilizing the five different commercial VSA systems (previously identified) to process audio files for which the ground truth was available. The analysts were required to make a Deception Indicated (DI) or a No Deception Indicated (NDI) decision on each file. It was decided to not allow for inconclusive results because of the presence of ground truth for all files tested and the procedures followed by National Institute of Truth Verification (NITV). The NITV training requires that examiners make a DI or NDI decision on each voice stress test conducted. Test 1 determined how effective the technology was at correctly identifying whether the statements being evaluated were truthful. When evaluating the output, each analyst was directed to use those techniques taught in the training class for each of the individual VSA system employed.

The test data consisted of the 56 audio files for which ground truth was known and was examined in various sets and subsets. Each audio file was tested multiple times using different systems. Tests were also repeated on some audio files using systems already tested. Some of the systems allowed for processing of data stored in the computer, while other systems required data to be fed into the computer from an external source. A trained analyst evaluated the files and logged their results. With the exception of the processing for the Truster/LVA tests, the primary evaluator processed the files, printed the charts and provided the charts to the

other evaluators for scoring. Once all examinations on the files were complete, the results were compared to the ground truth and scored as a positive (P), false positive (FP), negative (N) or false negative (FN). The combined results from each of the VSA systems are reported in Table 2 below.

A test was labeled *positive* when the test result determined the subject was deceptive and the true condition was also deceptive. A test was labeled as *false positive* when the test result determined the subject was deceptive and the true condition was truthful. A test was labeled *negative* when the test result determined the subject was truthful and the true condition was also truthful. Finally, a test result was labeled as *false negative* when the test result determined the subject was truthful when the true condition was deceptive.

Table 2. Raw results from VSA testing for all systems

Analyst	Positive	False Positive	Negative	False Negative
1	128	76	78	36
2	48	31	38	29
3	5	3	5	5
4	8	2	4	3
5	9	6	2	0
Overall	198	118	127	73

The following results employ similar approaches identified by the NRC in their report [7].

The test results indicate that the VSA systems produced results that allow for the detection of deception at a rate better than chance.

Sensitivity and specificity are used as indicators, as applied in deception detection, which defines a systems ability to accurately identify deception when it is present. Sensitivity is defined as the conditional probability of a system identifying deception when deception is present. Specificity is the conditional probability of a system identifying truthfulness when truthfulness is present. One must also consider the probability of a false alarm (when you identify something as being true, but in reality it is false). We can identify two other indicators to measure these probabilities. False negative probability is defined as the conditional probability of a system identifying truthfulness when deception is present. A false positive probability is defined as the conditional probability of system identifying deception when truthfulness is present. These indicators are given in Table 3, which are derived from the results provided in Table 2.

Since deceptive responses are considered a positive response, a perfectly sensitive system (100%) would identify all deceptive responses. The formula for this

indicator is the number of correctly identified positive responses divided by the number of true positive responses. The more sensitive the test is the more likely it will correctly identify deceptive individuals.

Table 3. Indicator values

Indicator	% Probability
Sensitivity	73%
Specificity	52%
False Negative Probability	27%
False Positive Probability	48%

Specificity indicates how accurately the system will correctly identify non-deceptive individuals. This indicator is the proportion of truly negative cases to the cases that produce negative results in the testing process. It is calculated by dividing the correctly identified negative responses by the actual number of truthful individuals tested. If an innocent subject tests positive due to other sources of stress, then the result is said not to be specific to the issue being tested for.

To provide an overall accuracy estimate this study followed the methods employed by the NRC in estimating polygraph accuracy [7]. An “equi-variance” binormal model is employed and shows six theoretical Receiver Operating Characteristic (ROC) curves, representing the different accuracy levels. The following graph (Figure 2) illustrates the use of this model. The single dot plotted on this graph is the average accuracy score for all systems evaluated. This point was plotted by locating the sensitivity on the left side scale and the specificity on the top edge scale. The diagonal line labeled “A=0.50” represents what is considered as the “chance” level. The reverse diagonal represents the equal error rate. The intersection of our measured Sensitivity and Specificity results falls just below the .70 accuracy curve.

In addition to sensitivity and specificity this study looked at the predictive values, both positive and negative of the results. The positive predictive value (PPV) is the percentage of those responses with test results indicating deception that were actually found to be deceptive after the investigation. Conversely, negative predicative value (NPV) is a predictor of the percentage of negative (truthful) tests that were correctly identified as truthful responses. Table 4 is a display of the average predictive values of the systems used in the tests. These probabilities were calculated from the values presented in Table 2.

The second test, conducted near the end of the two-year study, sought to investigate how experience and background of an analyst impacted the results obtained from the VSA testing. Two of the analyst came from backgrounds with no law enforcement experience and three analyst had over 25 years of experience each (See Table 1). The experience

these analysts had with VSA systems varied from over 5 years to less than 6 months.

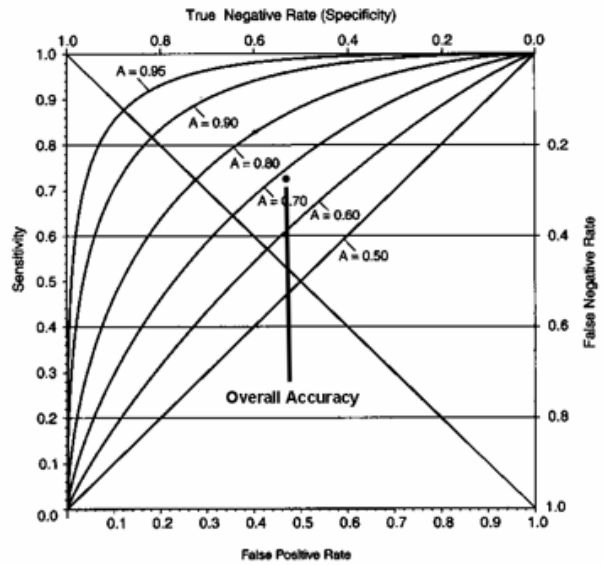


Figure 2. Equi-variance binormal model for accuracy estimation

Table 4. Predictive values

Predictor	% Probability
Positive Predictive Value	62.6%
Negative Predictive Value	63.5%

The primary analyst had 26 years in law enforcement, had never worked with any PDD technology prior to this study, but did receive training in all systems tested, totaling 192 training hours. The secondary analyst had no law enforcement background and received training in one energy based system and one frequency based VSA (64 training hours).

A subset of the above test files was used in the evaluation, which involved the following systems: Diogenes Lantern, TrusterPro (LVA) and DVSA. The waveform charts were created, provided to an independent member of the LEAF staff, who then placed new subject identifiers on the charts and cross referenced the original ID to the new ID. The primary analyst examined the data and made DI/NDI calls. The data was then scored and the results were compared to similar tests conducted at the start of the study. Results are provided in Table 5.

Test results show that an analyst ability to correctly identify deceptive subjects improves with practice and feedback. In the tests where the primary evaluator ran previously examined files, there was a rise in sensitivity scores as indicated in Table 6. However, this increase was at the expense of a lower level of specificity.

Table 5. Results based on experience of an evaluator

Analyst	Positive	False Positive	Negative	False Negative
Primary (start of study)	49	26	34	17
Primary (end of study)	42	25	26	8

Table 6. Accuracy improvements relating to experience of an analyst

Analyst	Sensitivity	Specificity
Primary (start of study)	74 %	57%
Primary (end of study)	84%	51%

An additional examination and comparison of the results logged between the primary evaluator and the secondary evaluator was conducted. The primary evaluator had considerable experience with VSA technology, having spent the majority of the previous two years researching and working with VSA systems. The secondary evaluator had less experience, having only been assigned to the program for a year and during which was spending only 33% of his time on the VSA project. In Table 7 the accuracy calculations for this comparison are displayed.

Table 7. Accuracy improvements relating to experience between analysts

Analyst	Sensitivity	Specificity
Primary	78 %	51%
Secondary	62%	55%

From these results we can see there is a significant increase in accuracy for the analyst with more experience with VSA systems (primary) as compared to the analyst with less VSA experience (secondary).

7. Testing and Considerations

Throughout this study, various data sets were collected and considered. These recordings, made under various conditions and settings, were collected from disparate sources. Different examiners and their styles were observed and various systems were employed to evaluate the audio data. Due to these variances, several problem areas were identified and have been addressed here for consideration of their effect on the outcome of the test results.

Many of the audio files received for study contained background corruptions that made analysis difficult, or even impossible, due to the poor quality resulting from a low

signal to noise ratio (SNR). Those audio files that were so badly corrupted were discarded for the purpose of this study. No attempt to measure the SNR was made during this study, but the corruption in some audio files was observed in the waveforms produced using the energy-based VSA systems. This noise came from a variety of sources, both intentional and unintentional. One source reported using a fan in the room to create white noise to distract the subject from focusing on noise in an adjoining hallway. Other noise sources were observed from ventilation systems, alternating current (AC) hum from electrical devices and other miscellaneous noise (voice, vehicles, etc.) from sources outside of the interview rooms. In some instances the microphone used to record the tests was separate from the microphone used to collect the voice for input into the VSA system. These secondary microphones were often placed a distance from the subject and resulted in a weak recording of the responses.

The LEAF has demonstrated Air Force developed technology designed to reduce noise in audio signals. Informal experiments were performed to examine what would happen to the resultant waveforms if the original signal were enhanced through this type of noise reduction. The Air Force Research Laboratory (AFRL) Speech Enhancement Unit (SEU) software was utilized and a two-question segment from a corrupted file was selected. A second area of silence was sampled to create a noise model. The model was then used to reduce the noise level by 80%. Once the noise had been reduced, the two questions were analyzed with two of the energy-based systems.

From this experiment, we found that the targeted portion of the waveform becomes more defined and easier to evaluate when the noise is reduced. What is not known is what affect the noise has on identifying deception from corrupted data. Likewise, it is also unknown what affect removing the noise may have on the systems ability to correctly display and identify the stress. A more defined waveform was displayed as a result of noise reduction, although, it was unclear if these results improved the performance of the VSA systems.

In the case of frequency-based VSAs, it was found that by applying noise reduction to the corrupted audio, the program was able to more clearly segment the responses. This system also accepted more samples as speech rather than discarding them as background noise. In this type of testing the more samples available to set the baseline, and test for deception, the better. During an interview with Amir Lieberman, designer of Layered Voice Analysis (LVA), he claimed that because the software sets a baseline level and movements from that baseline are what are calculated, and that the inclusion of noise should not affect the results. Our concern is that the noise levels do not remain constant across the data collection and it is unclear what affect this may have on the results.

8. Fusion with Polygraph

The fusion of similar or supporting technologies and information sources is a common theme in the intelligence

arena. Data collected from surveillance systems, human intelligence and signal interceptions are all used to create a more credible picture of the situation, rather than relying on one sole source to base important decisions. By fusing multiple PDD techniques, examiners have an enhanced probability of correctly identifying deceptive individuals.

Fusion of PDD technologies is likely to add value to any testing process as long as the performance of each individual technology is better than chance (50%). Having background information on a subject assists in creating appropriate test questions that won't cause inappropriate stress. Also, having other means to confirm the findings of one PDD system is also very useful.

Initially, fusion was considered as a mathematical process to provide statistical support to PDD as a whole by combining scores from each of the fields to support the credibility of the other. Later fusion took on the form of running a voice stress system in parallel to view the waveform of an audio response together with the tracings of a polygraph. This was in hopes that it would provide additional data in determining whether deception was present.

With the inclusion of the LVA system, which provides real-time analysis, a serial fusion concept was developed whereby the VSA could be used in a real-time mode against narrative responses during the pre-test interview. LVA can segment and evaluate narrative audio data automatically and display a message regarding the subject's mental state. The ability to see what areas are causing the subject difficulty would allow the operator to fashion questions that either created greater focus on the incident or avoid questions that may have reflected stress from outside issues. It is believed that this ability to focus the questions more tightly would

reduce the number of inconclusive results and improve the accuracy of the tests. See Figure 3.

VSA and polygraph follow the same examination procedures. They are as follows: pre-test interview, testing and post-test interview. Although there are differences in the amount of baseline stress an individual experiences during polygraph testing as compared to voice stress testing. The primary source of stress in polygraph testing is the physical and uncomfortable attachment of the polygraph sensors. This may become a factor for VSA if it is found that the amount of stress in the subject's voice increases the longer the subject is attached to the polygraph sensors. If the voice stress level does increase throughout the test, the testing would then require a comparison between known truths, known lies and unknown responses for the determination of deception.

Voice stress is a physiological reaction caused by a psychological response to a stimulus. The reaction and recovery time from the point of stimulus to the appearance of the reaction within the voice is nearly instantaneous, as is the reaction of the galvanic skin resistance (GSR) measure used in the polygraph. The remaining polygraph channels, pulmonary and cardio react and recover at a slower pace. In mechanical polygraphs the length of the pen arm adjusts for this difference in the GSR, while computerized polygraph adjust for the difference on the screen. Using various data sources recorded under stress, Dr. Hansen reported, "when a speaker is under stress, their voice characteristics change. Changes in pitch, glottal source factors, duration, intensity and spectral structure from the vocal tract are all influenced in different ways by the presence of speaker stress" [5].

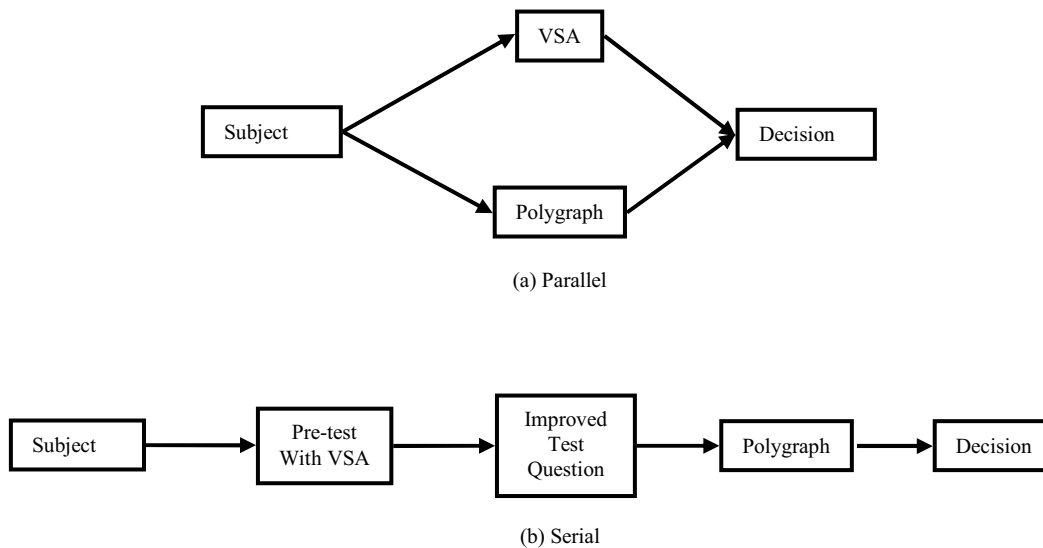


Figure 3. Different options for fusing VSA with polygraph

Interference between polygraph and voice stress analysis is commonly attributed to the physical attachments. The collection of an audio signal during a VSA examination can be accomplished without attaching a microphone to the person. If a desk microphone is employed it may be desirable to have the subject speaking directly into it and since polygraph testing requires the person to sit still, once the microphone has been positioned there would be no need to adjust it. If a clip on the microphone is used, care should be given not to catch the wire under the other sensors. Placing the microphone on the subject should be the last item to be installed. Attaching the microphone to a neck lanyard and hanging the lanyard on the subject aids in positioning the microphone the same distance from the mouth.

Another area of concern is the questioning style compatibility. This is not so much a concern of the formatting of the questions, as it is the timing of the questions. With the exception of CVSA, most VSA systems employ the same questioning formats as those used in polygraph testing. Test protocols such as General Question Tests, General Series Tests, Modified General Question Tests, various forms of Zone of Comparison and Peak of Tension tests are employed in both. CVSA uses a General Series, a Zone of Comparison, a Modified Zone of Comparison and Peak of Tension, but changes the question protocols to include an Irrelevant question after every Control and Relevant question. The insertion of an Irrelevant question is to give the examiner an opportunity to collect the carry-over stress and score it with the associated question prior to it. The NITV teaches that the second (Irrelevant) question needs to be asked immediately after the first question (Control or Relevant) in order to collect the carry-over stress. It is this difference in the timing of the questions that creates the greatest incompatibility between the polygraph and voice stress. Polygraph questions are separated by 20 to 25 seconds to allow the system to record any delayed stress and return to a homeostatic level.

Further examination of the test procedures of both VSA and polygraph will have to be conducted to determine if and what changes need to be made to ensure the best results when fusing these technologies.

9. Conclusion

It was not the objective of this study to recommend one Psychophysiological Deception Detection technology over another. Rather, it was our intent to provide users with an unbiased evaluation of VSA technology along with enough information to assist them in making decisions on what type of system to employ.

Our results indicated that, given the VSA systems tested, results indicate that this technology, with a trained and experienced examiner is capable of detecting

deception or truthfulness in a subject at a rate better than chance. The experience an examiner has with VSA technology plays a key role in their ability to detect deception.

These instruments alone are not “lie detectors”. The decision as to whether a subject is being truthful or lying should only be made by a trained examiner. This decision should be based upon reviewing the data presented by the instrument, the demeanor of the subject, and other evidence from the case. VSA systems are capable of providing an examiner with a waveform or other response that may be a reasonable reflection of the stress level being experienced by the subject, in a majority of the cases. The correct interpretation of this indicator is the responsibility of the examiner.

The goal in using a VSA system or polygraph should be to convince the subject that they cannot deceive the operator, and that the instrument will detect their deception and their best avenue is to confess to the crime. This study has shown that VSA systems will produce results that trained operators can employ with confidence to obtain confessions.

The results of these examinations should not be considered “proof positive” of innocence or guilt. While some consider these tools as an aid in focusing an investigation on proving someone guilty, officers should not lose sight of other suspects or evidence that may indicate otherwise.

This study has also shown that when training and experience of an examiner are taken into consideration, test results indicate that over time there is a marked improvement in an examiners ability to correctly identify deceptive subjects. Likewise, the comparison between the two analysts of different skill levels also indicates that experience may be a factor in improving accuracy. Observations made during the study of other analysts seems to indicate that the more opportunities one is given to run tests, examine charts and receive feedback (ground truth), the better the examiner becomes.

10. References

- [1] National Institute For Truth Verification, <http://www.cvsal.com/cost.php>.
- [2] O. Lippold, *Physiological Tremor*, Scientific American, Volume 224, Number 3, March 1971.
- [3] M. H. Beers and Robert Berkow, *The Merck Manual of Diagnosis and Therapy*, 17th Edition, John Wiley & Sons 1999.
- [4] D. Haddad, et. al, *Investigation and Evaluation of Voice Stress Analysis Technology*. Final Report for National Institute of Justice, Interagency Agreement 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

[5] J. H. I. Hansen, et. al., *Methods for Voice Stress Analysis and Classification*, as appendix to *Investigation and Evaluation of Voice Stress Analysis Technology*, Final Report for National Institute of Justice, Interagency Agreement 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

[6] D. Lykken, *A Tremor in the Blood, Uses and Abuses of the Lie Detectors*, New York, McGraw-Hill, 1981.

[7] National Research Council of the National Academies, *The Polygraph and Lie Detection*, Washington DC, National Academies Press.