ISJ investigative sciences journal

Volume 8, Number 1, January 2016

Assessing Deception by Voice Analysis: Part II: The LVA

James D. Harnsberger, PhD[1] and Harry Hollien, PhD[2]

Abstract

This overview is the second of a two-part series designed to brief law enforcement personnel, attorneys and members of the intelligence services about the ability of voice analyzers to detect deception and stress from speech. Such analyses are important as devices of this type are proliferating and their use is creating problems. Thus, the question was asked: Can these devices actually provide valid information about deceptive behaviors? The effectiveness of NITV's Computer Voice Stress Analyzer (CVSA) was reviewed in the Part I article (Hollien and Harnsberger, 2013) where it was reported that the system was found to be ineffective. This second of the two reviews will focus on a somewhat different device – Nemesysco's Layered Voice Analyzer (LVA). Here the summary will provide information about 1) its background, 2) relevant research, 3) a large laboratory experiment and 4) an extensive field study. The highly controlled laboratory experiment, assessing LVA, employed speech samples of individuals who systemically varied their utterances from normal to those which were intensely deceptive. To create the latter, subjects had to hold very strong views about some issue and were required to make sharply derogatory statements about them while believing that they would be observed by colleagues and friends. A double-blind evaluation of these utterances was carried out by two teams of qualified LVA operators. The field experiment (Horvath et al 2013) focused on a group of suspects split in two groups -- a group that had produced truthful utterances and another whose members produced deceptive speech. The veracity of these utterances was validated by several procedures based primarily on polygraph evaluations. The detection rates provided by LVA operators were both assessed directly and then contrasted with the accuracy of experienced auditors. The results obtained by both studies demonstrated that the LVA system operates at only about chance levels.

Keywords: Forensic sciences; deception; detection of lying; voice stress; speech analysis; speech and deception; Phonetics.

[1] Dr. James D. Harnsberger, Associate Dean, College of Criminal Justice, University of New Haven, New Haven, CT 06516; Email: jharnsberger@newhaven.edu

[2] Dr. Harry Hollien, Institute for Advanced Study of the Communicative Processes, University of Florida, Gainesville, FL 32607; Email: hollien@ufl.edu

Introduction

As has been indicated, this presentation continues the review of the effectiveness of psychological stress evaluators when they are employed in an attempt to detect deception (see Hollien and Harnsberger 2013 for Part I). First, it should be stressed that all of the devices cited were developed in an attempt to solve the problems facing members of the judicial, criminal justice and intelligence communities when they attempt to determine if a speaker is telling the truth. That is, law enforcement and intelligence agents must endlessly assess statements made by suspects or informants for evidence of truth and/or deception. Attorneys face a similar problem with clients and witnesses. This situation is so pervasive that any system which promises to identify truthful utterances when they occur – and separate them from falsehoods – would be of substantial value. As a result, a number of such devices exist which are advertised as capable of meeting the problem. Indeed, they are vigorously promoted as both effective and easy to use. But are they?

The materials to follow constitute an effort to provide useful responses to these questions for judges, attorneys and agents of all types. As stated, this, Part II effort, will focus on the LVA and its validity as a lie detector. For technical information about the laboratory research on the LVA, see Harnsberger et al 2009.

As was also noted in Part I, research of this type closely parallels investigations which are conducted by attorneys when they analyze a client's case. Law enforcement professionals also operate in a similar manner when investigating crimes. They too, use observations and analyses when attempting to determine if a suspect is guilty or not. Thus, the process of gathering and interpreting information carried out by agents and attorneys pretty much parallels the kind of research that has to be conducted in assessing devices and systems such as these. Admittedly, the jargon here may differ a little but the procedures are quite similar. Hence, the remarks to follow should be intelligible to the reader with but little tweaking.

Background

It is not at all difficult to imagine the impact upon society which would occur if it were possible to reliably (and quickly) detect when an individual was lying. Consider the effect that it would have on family relationships – or, on a larger scale, when advertisements are heard. Better yet, imagine the societal improvement which would occur if it were possible to determine a politician's beliefs, and (especially) his or her intent, simply by somehow assessing their speech. Accordingly, there is but little question, but that the availability of an infallible lie detector would be of great value to the courts, to law enforcement, and to counter-intelligence operations. For one thing, there would be little need for criminal trials. The guilt or innocence of a defendant could be determined by simply asking the question: "Did you do it?" -- and then analyzing the response.

On the other hand, it has been well established that, to identify a "lie", the behavior observed would have to be measurable and that all speakers would have to exhibit that specific feature (or cluster of features) *whenever* a lie was spoken. In addition, these same behaviors could not also be indicative of other types of behavior. Of course, basic research on the detection of lying is being carried out (examples, DePaulo et al 2003, Rockwell et al 1997, Burgoon and Qin, 2006). Yet all of the investigators in this area are limited by the problem described by Lykken (1981). He effectively articulated the key concept here; that is, for deception to be detected, a lie response must exist. In addition, some measurable physiological or psychological reaction must *always* occur when a person lies. Lykken insists that, "Until a lie response has been identified and its validity and reliability established, no one can claim to be able to detect or measure falsehood on anything remotely approaching an *absolute* level." His statement is a little strong perhaps, but it does provide a reasonable set of standards for detecting falsehoods.

The System to be Assessed

But what do the manufacturers of the layered voice analyzer claim it can do? Well, among other things, they insist that it can detect stress and deception and do so accurately. As a consequence, a number of these systems are currently being used by law enforcement, security, military and intelligence organizations. Of course, they are constrained by the fact that testimony based on the resulting "analyses" is not accepted in courts-of-law. That is, such testimony is not found there because, at present, it does not come close to being admissible with respect to either Daubert (1993) or Frye (1923), much less that it satisfies any reasonable level of "certainty" (Marco, 2000). Indeed, a large number of state legislatures and courts have voted, or ruled, against acceptance of testimony of (or reports based on) this type of analysis. On the other hand, their presence cannot be ignored as their very use forces it (at least tangentially) into both investigations and trials. Indeed, the present authors have been consulted a number of times about their nature and validity in both criminal and civil cases (examples: Florida vs Joyner, 2008, Commonwealth vs Perrier, 2013).

On the other hand, is it at all possible that systems such as the LVA, are *potentially* acceptable? Indeed, it is well known that human speech and voice *does* contain features which can be used to provide information about a person (see Hollien, 1990). Examples include speaker identification – an area based on analysis of speaker-specific vocal properties (Stevens, 1971; Nolan, 1983; Hollien, 2002). Another involves the detection of alcohol intoxication. Here too, substantial research is available (Pisoni and Martin, 1989; Chin and Pisoni, 1997; Klingholtz et al, 1998; Hollien et al, 2001a, 2001b). In addition, human emotions (including psychological stress) can often be detected in voice (Williams and Stevens, 1972; Hollien, 1980; Hicks and Hollien, 1981; Scherer, 1986; Cummings and Clements, 1994). However, as an unfortunate consequence of the relationships cited above – and even though such is not the case – the manufacturers of the LVA and similar devices argue that research on these issues provides an operational basis for *their* products. A second unintended consequence – also resulting from this situation – is that the voice-stress/deception controversy has overshadowed (and, to some extent, stifled) legitimate research from being carried out on the basic speech-deception relationships.

Also, and as was pointed out in the Part I article, it must be stressed that the psychological and neurological substructure for all of these behaviors are quite complex. The oral production of any spoken language involves the use of multiple sensory modalities, high level cognitive functioning, complex cortical processing and a large series of motor acts (Fant, 1973; Netsell, 1983; Abbs and Gracco, 1984). Thus, while the resulting theories would predict that even subtle operations – ones such as the detection of deception and/or truth – *should* be possible, the relativity simple (and somewhat primitive) mechanisms supporting the devices being reported on here probably cannot be expected to be effective. These limitations are especially relevant since these systems were constructed, and put on the market, before any serious research was carried out on them.

Research on the Various Voice/Stress Approaches

The question must now be asked: Has any research at all been carried out on the ability of these devices to operate effectively? The answer is "yes" but unfortunately nearly all of it came after they were already in use. Moreover, most of the older research was focused on those systems related to the PSE (Psychological Stress Evaluator). As will be remembered from Hollien and Harnsberger (2013), the proponents of that type apparatus insist that their devices identify lying or stress by measuring the *microtremors* which occur in the laryngeal muscles of human beings. Unfortunately, however, microtremors of this type appear to be found only in the long muscles of the body (Brumlik and Yap, 1970; Lippold, 1971) and not in the laryngeal muscles (Inbar and Edin, 1976; Shipp and Izdebski, 1986). Nevertheless, the manufacturers ignore this information and argue that their apparatus does measure "shifts" (in these "microtremors") when stress and deception occur. Of course, even if true, this operation would only lead to yet additional problems. First, there is no evidence at all that the microtremors would affect voice and speech in any manner even if they did exist. Second, it has not been demonstrated that stress always (or even usually) accompanies deception.

On the other hand, the proponents offered little (Heisse, 1976) to no direct scientific evidence to support their claims. As a result, independent investigators have researched the issue. The results are as follows. While some investigators suggest that the voice stress devices of the earlier type *might* detect stress – at least under certain circumstances (McGlone, 1975; Brockway et al, 1976; Vandercar et al, 1980) – the link between stress and deception has not been established (see Horvath, 1979; Cestaro, 1996; Jangiro and Cestaro, 1997).

However, it can also be said that, to date anyway, none of the relevant systems had been provided with an extensive and comprehensive assessment. Indeed, spokesmen for the companies involved in system production have complained that investigators either 1) have not controlled their procedures to a level which would permit robust data to be generated (Lynch and Henry, 1979; O'Hair et al, 1985), 2) have assessed or controlled too few variables (McGlone et al, 1974; Greanor, 1976; Leith et al, 1983), 3) have carried out research that was somewhat narrow in scope (Horvath, 1978; Haddad et al, 2002), or 4) focused only on field studies (Kubis, 1973; Barland, 1975). On the other hand, it also must be pointed out that almost all of these

projects have exhibited rather good research protocol and none of them provided data which supports the use of the devices in question.

Moreover, even our own research (on the original PSE system) was not as definitive as would have been desirable (Hollien et al, 1987). In that investigation, subjects were used who were committed anti-vivisectionists; they were required to vigorously attack the position that animals are abused, or worse, when used in research and related activities. As expected, the task cited above was shown by self-reports and standard tests of stress to produce conditions of very high stress and/or significant anger when the subject uttered the required "lies". Yet the system being tested was unable to identify these conditions. In short, even though limited (no control group was included and only the "hit" rate was assessed), the obtained data demonstrated that the device only operated at about chance levels.

Nemesysco's Layered Voice Analysis

What is the LVA all about? What does Nemesysco claim it can do? This device, it is claimed, is capable of detecting not only deception in speech, but also a large variety of other behaviors. Included are conditions of emotional stress, cognitive effort, fear of discussing a particular topic, stress due to deception, anxiety, arousal, speech intent, condescending attitudes, physical attraction, and many others. For perspective, it should also be noted that it is a recent iteration of a whole line of products. They include devices such as Truster, Truster-Pro, and Vericator) currently marketed in conjunction with related products such as SENSE and VoiceSum (also by Nemesysco and its US company V LLC). For the present review, however, only the device's sensitivity to the presence of deception and stress will be considered.

Nemesysco also indicates that LVA's sensitivity to the cited emotional and cognitive states is based on methods that are distinctively different from other previously discussed voice stress analyzers -- ones such as the Psychological Stress Evaluator (PSE) and the Computer Voice Stress Analyzer (CVSA). As will be remembered, those systems are supposed to measure the acoustic consequences of hypothetical ''microtremors''. In contrast, the LVA manufacturer states (in the LVA manual) that it ''performs a wide-spectrum analysis, uses an automatic calibration and filters through emotion levels.'' In training materials provided by V LLC, their device is also said to rely upon a ''voice frequency'' analysis involving the application of ''8000 mathematical algorithms'' to ''129 voice frequencies'' that are affected by ''psychological versus physiological body reactions to the stress of telling lies.'' Of course, it is not possible from descriptions such as these, to identify any of the information that actually is extracted from the speech signal, or, especially, how it might be processed by LVA. For instance, the phrase ''wide spectrum analysis'' is not one typically used by Phoneticians and Acousticians as there are many types of spectra -- all with their own operational definitions. Further, ''129 voice frequencies'' might -- or might not -- refer to 1) the product of bandpass filters which span the range of audible frequencies, or 2) vocal fundamental frequencies or 3) the frequency of vowel formants. Since the manufacturers provide little to no meaningful information upon which

operational descriptions might be based, it is most difficult to determine just what may be happening inside an "LVA". In short, it must be dealt with as being only a "black box".

One of the few investigations that focused directly on the LVA was carried out by Damphousse and his associates (2007, 2009). They employed this device (and also the NITV's CVSA) in field research that may be considered to parallel a "real world setting". The "subjects" for their study were a number of individuals who had been arrested for various types of crimes and were being incarcerated at a county jail. The investigators asked a number of them pertinent questions about their recent drug use. Their answers were processed for "evaluation" by the LVA system as a part of the Federal Arrestee Drug Abuse Monitoring (ADAM) program. Their responses were also evaluated by several experienced operators who had been trained and certified at the LVA school (V LLC). The "ground truth" here (i.e., veracity of the data) was established by standard urine tests (also mandated by the ADAM program) for five drugs (cocaine, opiates, marijuana, etc.).

The investigators reported that "the results were not very promising" with the LVA consistently failing "to correctly identify respondents who were being deceptive." They specified that "only about 15% of the respondents who actually had used drugs" but indicated that they had not, "were identified as being deceptive." This work essentially demonstrated that the device being tested was functionally unable to detect deception.

One Approach to LVA Testing

Consideration of the cited research leads to the inescapable conclusion that it is necessary to carry out extensive and rigorous – but balanced and fair – research if the effectiveness of the LVA is to be understood. Accordingly, a model was designed for this purpose (Hollien and Harnsberger, 2006, and Hollien et al 2008). It was employed as a basis for this research program and it was initiated by a large, rigorously controlled laboratory experiment. For this purpose, utterances involving truthfulness, deception, psychological stress -- produced at various levels of deception, truth and stress -- were obtained from a variety of speakers. These experimentally induced behaviors were both relevant and verifiable by independent assessment. The model's other levels specified field-grade experiments or ''real life'' studies, but ones where attempts can be made to achieve reasonable levels of control. Two studies of this (second) type have been carried out by other groups. The first (Damphousse 2007) has been discussed; the second (Horvath et al 2013) will provide material for an extended review in the final section of this presentation.

The Laboratory Experiment

Before the reader can be expected to evaluate the validity of the findings provided by investigations such as those to be described -- or what they tell us – it is necessary to understand just how the procedures they used were structured and carried out. Accordingly, attempts will be

made to provide enough detail to permit decisions to be made about the level of accuracy and validity of our procedures.

The Speakers

Seventy-eight adult volunteers, half each male and female, were screened for inclusion in the project; their ages ranged from 18 to 63 years. The group's makeup roughly paralleled the demographics of the U.S. population (represented were salesmen, school teachers, the military, students, housewives, clergy, nurses, police, and so on). To be included, they also had to be shown to hold very strong personal views about some subject (e.g., politics, religion, sexual orientation, gun ownership, or some such). Those selected were recorded in a quiet room with laboratory quality microphones coupled to audio recorders and a computer. Additionally, digital video recordings were made of them during all experimental runs.

Determining Subjects' Stress Levels

Mindful of the problems associated with the external determination of internal emotional states of humans (Weinstein et al, 1968; Bradley and Stocia, 2004), four methods appropriate for the assessment of psychological stress were administered. They included 1) two tests of anxiety/stress based on self-reports and 2) two continual body response evaluations consisting of galvanic skin response (GSR) and pulse rate (PR). The anxiety/stress tests consisted of an "emotion felt" checklist and a modified version of the Hamilton test (Maier et al, 1988). The two physiological measures were selected to be both minimally invasive and to permit quick tracking of the fast paced experimental procedures. The results were combined; subjects had to demonstrate at least double their basic values for the stress/lie conditions. (see below).

The Speech Samples

The nature of the speech samples employed in this project was especially important. Among the several different types of utterances produced by the subjects for the various sub-experiments were 1) low stress truth and 2) high stress lie. All of these samples consisted of speech passages structured with 5-7 content sentences (45-57 words) and a 17-26 word 'content neutral' sentence of related context embedded near its center of the passage. The neutral sentence (example: "I have considered these experiences many times and have not, or probably will not ever change my mind) was inserted so that speech could be analyzed which contained no revealing language cues but was at the same stress or deception level as the full passage. The use of this control was most important as the LVA operators could hear the utterances during processing. Hence, these 'content neutral' phrases prevented operator exposure to language-based clues relative to the judgments they were asked to make. All speech samples were uttered a number of times but only the one meeting all *experimental* criteria was used. The two classes of samples employed were:

Low-Stress Truth Each subject uttered a truthful passage; the content of which was on a pre-designated unemotional topic. Examples: 1) descriptions of some particularly pleasant feature about where they lived or 2) where they went on vacation.

High Stress Lie Samples of this type consisted of untruths produced under conditions of high jeopardy. As stated, all subjects selected were required to hold very strong personal views about some issue (as based on admissions, interviews and observations by our Psychiatrist) For the experimental trials, they were compelled to utter statements that sharply contradicted these views and to do so while under the impression that their friends and peers would hear (and see) their performance. They also were instructed to produce them in a speaking style that strongly suggested that they actually believed the lie(s). Third, the psychological intensity of their utterances was enhanced by selected use of an electric shock procedure added to what was an already high stress condition. The equipment employed for this purpose was a standard (medical) electro-stimulus conditioning unit. The electric shock was administered during the initial run and any subsequent run where the subject failed to demonstrate highly significant signs of intense stress. As a result, subject's lies could be shown to reflect extremely high jeopardy.

It was by this rigorous procedure that samples of low stress truth could be directly contrasted to high stress/high jeopardy lies.

The Procedure

As would be expected, the procedures employed met all subject safety requirements as well as anonymity. Moreover, they were screened by the project's Psychiatrist. He asked them a series of questions about their 1) history of psychiatric disorders and psychological trauma, 2) history of heart conditions and other physical disorders, 3) current medication regimen, drug use and alcohol use, and so on. In addition, he conducted the interview in a manner that induced tension and heightened the selected subjects' arousal for the high stress/deception condition which was carried out first.

After a trial, each of the four stress-level checks were averaged and converted to a common scale. Only those subjects were included whose mean stress level, when lying with jeopardy, was always more than double their baseline stress level. Specifically, the mean *overall* stress shift observed for all subjects was 141% -- with a mean rise of 129% for male speakers and 152% for females. The speech samples for the 48 individuals (24 each: males, females), who met all selection criteria, were used in the analysis. Their speech materials were organized (by gender) into randomized sets of the truth and lie samples for evaluation by the LVA operators.

The LVA Operators

Two teams of examiners operated the LVA equipment. The first was a team of two evaluators provided by the University of Florida's Institute for Advanced Study of the Communication Processes. They attended (and completed) the V LCC school where they were tested and certified by the LVA instructors as "competent to conduct LVA analyses". The second evaluation team consisted of two highly experienced LVA instructors, chosen by the manufacturer from the V staff, who traveled to the University of Florida to participate in the study. Both teams (IASCP and V) classified all samples as either deceptive or nondeceptive and either stressed or nonstressed.

Evaluation Task

The LVA system requires a minimum of sentence-length speech materials be submitted for testing plus a ''balanced'' portion of an individual's normal speech for calibration purposes. Thus, to prepare for the LVA assessment the V company required all of the test samples to be individually paired with a calibration passage—provided by the subject. In turn, they and the experimental samples were inputted for assessment as single digital audio files following the manufacturer's instructions. Subsequently, they were assigned random file names to ensure that no stress or deception information about the sample would be available to any LVA operator. Indeed, this is why only the sections of the experimental samples, where no linguistic cues to the emotion being felt, were used.

The LVA analysis itself was conducted differently by the two teams of evaluators. The IASCP team employed the (recommended) protocol that did not require judgments by human operators. Rather the analysis was conducted automatically without any potential operator ''bias'' or other effects possible. The manufacturer's team did not follow the same procedure; rather they did not use the same protocol with all samples. Apparently, they employed the protocol they judged most effective for each sample. In any event, their operation was confidential and could not be documented. However, these operators were both highly experienced examiner/instructors specifically selected by V LLC. Thus, it is reasonable to assume that their evaluations fully met their employer's specifications.

The Findings for LVA

As would be expected, the resulting data were evaluated by means of a number of techniques designed to explore the possibility that the LVA system might be sensitive to stress, truth, and/or deception. In all cases, four values were calculated: true positive, false positive, false negative, and true negative. The true positive rate (or ''hit rate'' in Signal Detection Theory), refers to the percentage of the time that deception (or high stress) is said to be present when it actually is present. That is, true positive rates measure how often a device accurately classifies a deceptive utterance as deceptive, high stress as high stress, etc. Equally important is the calculation of the false positive rates (also known as the false alarm rate in Signal Detection

Theory). They correspond to the percentage of times the target signal is said to be present when in fact it is absent. False positive rates *must* be compared with true positive rates to determine the device's ability to correctly discern deception. An examination of the true positive rate alone does not provide system accuracy or validity as a high true positive rate can be the product of either its actual accuracy or simply its bias (or the operator's bias), regardless of the actual presence or absence of the behavior being tested. An accurate device would show true positive rates that are both higher than (and significantly different from) the false positive ones. On the other hand, a device that performs at chance would show equal, or close to equal, true and false positive rates. Finally, the product of both the IASCP and V teams, working separately, were given to an independent investigator who had them collated for the statistical analyses. In short, the experiment was conducted in a *double blind* manner.

It would appear that the best way to present the LVA results would be to focus on the most important contrast from among the many assessed. As stated, the one selected here was that between the high jeopardy *deceptive* utterances and low stress truth; these relationships can be observed by consideration of Table 1. That figure is organized with the collated data for the IASCP evaluation team in the left-hand portion of the chart and those for the V team (i.e., the manufacturer's instructors) on the right. Both of these sub-tables are structured identically with the scores for spoken deception, which was actually judged as deception, found in the upper left cell and that for the truth-truth judgment in the lower right.

Table 1:

Identification of deception and truth in the speech samples by both certified LVA evaluation teams (left side: IASCP; right: V team). The top left value (judged as deception when deception actually exists) and the lower right cell (actual and judged truth) are among the contrasts. Note the high values in the "false positive" category top right. (Data drawn from Harnsberger et al, 2009).

	Actual Condition	
Judged Condition	Deception	Truth
Deception	50% *(True Positive)*	60% *(False Positive)*
Truth	50% *(False Negative)*	40% *(True Negative)*

IASCP Team

	Actual Condition	
Judged Condition	Deception	Truth
Deception	52% *(True Positive)*	40% *(False Positive)*
Truth	48% *(False Negative)*	60% *(True Negative)*

LVA Team

As can be observed, at 50% correct, the IASCP team was able to identify at least half of the study's subjects when they were lying with high levels of jeopardy. At 52%, the level for the V operators was but little different. If these scores are taken alone, it can be argued that the LVA system can at least be used to identify lying about half of the time that it actually exists. Yet, to accurately assess the device's ability to detect lying, you must also look at the upper right hand values found on both sub-charts. Here it can be seen that the IASCP team identified low stress truthful statements as lying slightly more often than they did the ones that actually were deceptive (i.e., at 60%). While the V team did a little better, they still identified 40% of the low stress truthful utterances as being lies! These patterns strongly suggest that the system operates at but chance levels.

Seen in another dimension, it also was clear that the device, as used by either team, was pretty much unable to recognize low stress *truthful* statements as not being deceptive when they occurred. Since these two sets of data demonstrated that neither of the teams was able to correctly identify either deception or truth at much beyond chance, it only can be said that the system employed simply was unable to provide them with appropriate information.

As it turns out, application of several statistical analyses (initially ANOVA) validated the above statement. Moreover, an even more powerful index of sensitivity -- d' or d prime (MacMillan and Creelman, 2005) – also was employed. In this case also, the values for both teams were found to lie in regions far below even that of *minimal* sensitivity.

Of course, the LVA defenders might attempt to advance alternate interpretations of these data. For example, they could argue that the results might reflect experimental inadequacies -- ones where the stress shifts for the speech samples were not actually of the magnitude reported, or they actually were not comparable to those educed in situations outside of the laboratory (i.e., from 'real-world' interrogations by police or the military). This interpretation *might* appear cogent (marginally anyway) if only the true positive rates were assessed. However, and has been pointed out a number of times, the evaluation of LVA's performance on truthful and unstressed speech samples served as an important control. That is, they permitted the examination of that device's potential bias to identify speech samples as deceptive in either the presence or *absence* of that state. If the speech samples contained inadequate levels of 'real-world' deception, then false positive rates near zero would be expected. Such was not the case. Rather, the system misclassified the truthful samples as lying at nearly the same frequency as those for lying under jeopardy. These high false positive rates simply cannot be explained away by arguing that inherent limitations within laboratory protocol existed. Indeed, quite the contrary is true.

An LVA Field Study

The work described above was followed a few years later by field research conducted by Horvath et al (2013). They obtained actual samples of lies and truthful statements made by suspects who were being interrogated by experienced agents of the Michigan State Police in "real life" situations. These behaviors were validated by a series of procedures (including

convictions, confessions and polygraph examinations) which were conducted by personnel who did not participate further in the project. Project protocol was to carry out LVA analyses (certified operators) and then, to contrast its effectiveness against perceptual judgments made by human auditors, by experienced interviewers. As would be expected, the goal of both groups was to determine which of the suspects had uttered falsehoods and which of the others had spoken the truth.

Basic Material for Project

As stated, the data-base for this research project consisted of high-quality audio recordings extracted from the pre-test interview portion of polygraph examinations of suspects accused of committing a variety of crimes. Those employed were drawn from a large number of polygraph examinations which had been conducted by the MSP (Michigan State Police) examiners in actual cases. It should be noted that, while all of the interrogators agreed to permit their work to be used in the project, they were not involved further in it. In addition, only those recordings were included where the examiner was able to determine if the suspect had been either deceptive or truthful. All other outcomes, (i.e., judgments which were inconclusive, incomplete, etc.) were excluded. Second, the cited lie/truth judgments made by the examiner had to be further confirmed by at least one of the two statistical classification algorithms used in polygraphy (i.e., either the Polyscore, version 1.0.0.1, or the objective scoring system [OSS] version 2). Both of these validations are widely applied by the polygraph testing community (Webb et al 2008). It should be noted that only these tests were employed for sample selection. Other classes of confirmation (the confessions cited above for example) were not so used but were included in a sub-project conducted independently.

At this juncture, the research group had an experienced facilitator review the material obtained from 210 (then available) interviews. He, in turn, chose the first 74 from those that met all selection criteria and placed them in the data-base. The audio recordings for all 74 "subjects" first provided the material for LVA testing. Later, they were used for the alternate evaluation -- an assessment by auditors. However, only those portions involved in the testing for truth and deception were employed.

LVA Evaluation

Horvath and his associates stress that this project was carried out in a *double-blind* fashion with the LVA operators knowing only of the general nature of the experimental passages and how/why they were obtained. These operators were two MSP personnel who had completed the training course provided by the LVA company (V LLC). Subsequently, they were tested (successfully) and were certified as competent to conduct LVA examinations. All LVA analyses were carried out with the operators working independently. In all cases, they employed the recommended LVA algorithm in order to render their decisions as being 1) truthful, 2) deceptive, or 3) inconclusive. Again, it must be stressed that they followed all LVA protocol exactly and as specified by V LLC.

The authors of this project report that, of the 74 samples in the suspect data-base, 31 had been judged by the relevant polygraph examiner as truthful and 43 had been reported to be deceptive. As stated, these judgments then were confirmed by both the polygraph examinations and either one or the other of the computer scoring assessment algorithms.

The Auditors

The next step in the project was to recruit auditors for the aural perceptual procedure. A number of trained and experienced interviewers were evaluated and three were selected for this procedure. Two were highly experienced (each, over 20 years) in the use and evaluation of the Behavioral Analysis Interview (BAI) and related procedures; the third auditor was less experienced (6 years), but still was very familiar with these processes (see Horvath et al 1993 and 2008). At this juncture the three auditors were instructed to listen serially to each of the 74 audio files and judge them on the basis of being either truthful or deceptive. They, further, were required to provide a quantitative degree of confidence for each decision they made. Statistical analyses were carried out using the scores generated by the auditors' truth/deception decisions and their confidence scores.

The Findings

The authors first reported the scores for both the auditors and the LVA operators; they then contrast them with each other. The auditors' correct decisions for the 31 interviewees who were **truthful** ranged between 39% and 84%, with a mean of 68%. Moreover, these three individuals were able to make a decision on all 31 truthful samples i.e., they had no inconclusive judgments. The scores for the LVA operators were somewhat poorer than those of the auditors. One exhibited an accuracy of 52% -- but only for those judgments he was able to make -- as 35% of them were inconclusive. At 46% *correct*, the other LVA operator would seem to be nearly as accurate. Not so, as half of his judgments (i.e., 52%) were inconclusive. As a team, the LVA operators were correct 48% of the time. However, nearly as many of their decisions (i.e., 44%) were inconclusive. In short, there is little question but that the auditors were more accurate in identifying truthful statements than were the LVA operators.

The above observation was confirmed by the deceptive judgments. That is, when the scores for the auditors and LVA operators were contrasted with respect to the 43 individuals who were **deceptive,** the group differences were yet greater. The auditors' correct judgments ranged between 58% and 81% (with no inconclusive judgments) and their mean correct was 71%. Indeed, the accuracy of both the second and third auditors statistically exceeded chance. The LVA operators were not as accurate. The first of the two was correct only 28% of the time and second weighed in at but 21%. As a result, neither exhibited an accuracy for the deceptive utterances that was greater than chance. Moreover, their precision was further degraded by an inconclusive decision rate of 31%.

A Sub-project

Horvath and his associates also carried out a sub-study where they added a new level of "ground truth" to their original criteria. In this case they compared the judgments of the LVA operators to those of the auditors where the deceptive utterances studied were *further confirmed* by confessions. It turned out that 18 of the 43 suspects showing deceptive during interrogations could be further identified in that manner. Of course, the authors were aware that *false* confessions can exist -- even with respect to very serious crimes. Nonetheless, they felt justified in making observations of that type as part of the project. In doing so, they would provide an interesting second level test for the LVA. Of course, when the data from this sub-procedure are considered, false confessions cannot be ignored. That is, even though there is evidence that, under certain conditions, they will occur, substantial dispute and uncertainty still surround the issue (Cassell, 1998; Vrij, 2000). Accordingly, it would appear that these authors could legitimately use the procedure in research of this type.

In any event, there were 18 instances where the interviewee's deception had been "confirmed" by a confession. It was of interest therefore to examine the results for these 18 cases and contrast them to those of the 25 selections based only on the prior indicators (i.e., the polygraph results, the statistical classifier, and interrogator's judgments).

Table 2 presents the findings resulting from this procedure; in this instance, inconclusive outcomes are treated as errors. As can be seen, the mean accuracy of the three auditors in the 18 confession-confirmed instances was greater than those for either of the two LVA operators. The auditors' mean accuracy was 70% whereas this value for the LVA operators was only 42%. In those cases that were not confirmed by confession, the means were 72% and 48% respectively. It also is of interest to note that the auditor who exhibited the lowest accuracy here paralleled that of the LVA operator, who had the highest (i.e., 56% for both).

Table 2:

Summary table of the means contrasting the judgments of deception made by LVA operators as opposed to the auditors. Data are from the 18 evaluations where deception was supported by confessions and the 25 instances where the deceptions occurred but were not externally verified.

Evaluation group	N	Confirmed by confession		Not confirmed by confession	
		N=18		N=25	
		Correct	Incorrect	Correct	Incorrect
LVA	2	42%	58%	48%	52%
Auditors	3	70%	30%	72%	28%

Adapted from Horvath, McCloughan, Weatherman and Slowik, 2013

In sum, the findings reported by Horvath, McCloughan, Weatherman and Slowik (2013), for their field assessment of LVA's value in detecting deception, do not provide any reason for optimism. Here, the LVA operators produced correct calls for deception, on average, only 25% of the time when deception was verified by polygraph examination. When deception was not present (i.e., the suspects/subjects were truthful) the LVA operators were correct only 49% of the time. Worse yet, when the "guilty" persons had confessed their involvement in the matter under investigation and thus had acknowledged their deception, the accuracy of the LVA analysis still only averaged 48%. In short, there was no instance in which the LVA produced correct decisions at a level greater than chance. These results are remarkably similar to those reported in the field-based study by Damphousse et al (2007). Moreover, they are also consistent with the large laboratory experiment described in this review. A device that is, in fact, sensitive to these states should not falsely detect them if the procedures employed actually failed to elicit them. In other words, whether in the field or in the laboratory, the available research unfortunately does not show the LVA as having the ability to detect deception or, on the other hand, truthfulness.

To Conclude

In conclusion, it must be noted that, even though the LVA has not been shown capable of validly detecting the cited behavioral states from voice, it might be possible that its occasional success (as reported by its operators) could actually occur (see Whitworth, 1993 as an example). If so, however, it probably was due to some situational factor or relationship. One such investigational component could result from the skill of the interrogator (rather than efficiency of the equipment). As a matter of fact, Horvath et al (2013) supported this position by demonstrating that trained/experienced interrogators can detect the presence of truth or deception – at significantly high accuracy levels – simply by listening to (appropriate) utterances (it is doubtful that information of this type would surprise Fischer and Thompson, 2014). Indeed, these findings are important in-and-of themselves; they also are consistent with other (related) research findings (see for example Burgoon and Qin, 2006, Chin and Pisoni, 1997, Hollien, 1990, and Scherer, 1986). It also can be said that these findings effectively validate the position cited. On the other hand, it may be that devices of this type are used as a prop to intimidate interviewees into confessions. In any event, it would appear unwise for attorneys or law enforcement personnel to depend on the capabilities of this type of device. If it is encountered, the best course of action probably would be to either avoid it or challenge it.

Acknowledgements

The laboratory research project featured in this report, was supported by the U.S. Department of Defense – specifically, by the Counterintelligence Field Activity (CIFA) contract FA-4814-04-0011.

The authors also recognize the fine cooperation afforded by both V LLC and Nemesysco. We also are grateful for the excellent support provided by the V LLC staff/operators – especially with respect to chart reading

References

Abbs, J.H. and Gracco, V.L. (1984) Control of Complex Motor Gestures: Orofacial Muscle Responses to Load Perturbations of the Lip During Speech. Neuropsysiology, 51: 705-723.

Anolli L, Ciceri R. (1997) The Voice of Deception: Vocal Strategies of Naive and Able Liars. J Nonverbal Behav;21:259–84.

Barland, G. (1975) Detection of Deception in Criminal Suspects [unpublished Ph.D dissertation] Univ. Utah, Salt Lake City, UT: Univ. of Utah.

Bradley, M.T. and Stocia, G. (2004) Diagnosing Estimate Distortions Due to Significance Testing in Literature on Detection of Deception. Perception and Motor Skills, 98: 827-839.

Bransletter, L. and Brunette, L. (2006) The Truth and Voice Stress Analysis, National Academy Associates, 8: 24-27, 32.

Brockway, B.F., Plummer, O.B. and Lowe, B.M. (1976) The Effects of Two Types of Nursing Reassurance Upon Patient Vocal Stress Levels as Measured by a New Tool, the PSE. Nursing Research, 24: 440-446.

Brumlik, J. and Yap, C. (1970) Normal Tremor: A Comparitive Study, Springfield, IL, C.C. Thomas.

Burgoon, J. and Qin T. (2006) The Dynamic Nature of Deceptive Verbal Communication, J. Lang. Soc. Psychol. 25: 76-96.

Cassell, P. (1998) Protecting the Innocent from False Confessions and Lost Confessions from Miranda. J. Crim. Law and Crimin. 88: 497-556

Cestaro, V.L. (1996) A Comparison of Accuracy Rates Between Detection of Deception Examinations using the Polygraph and the Computer Voice Stress Analyzer in a Mock Crime Scenario. Ft. McClellan, AL: US Department of Defense Polygraph Institute Report No. DoDPI95-R-0004.

Chin, S.B. and Pisoni, D. (1997) Alcohol and Speech. San Diego: Academic Press.

Commonwealth of Massachusetts vs. Christopher Perrier, C-501-58684 Superior Court of Hampton County, Indictment No. 121421 (2013).

Cummings, K. and Clements, M. (1994) Analysis of Glotal Exitation of Emotionally Styled and Stressed Speech. J. Acoust. Soc. Amer., 98: 88-98.

Damphousse, K.R., Pointon, L., Upchurch, D. and Moore, R.K. (2007) Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting. Report No. 219031 to U.S. Department of Justice.

Damphousse, K. (2009) Voice Stress Analysis; Only 15 Percent of Lies About Drug Use Detected in Field Test. Washington, DC: National Institute of Justice, U.S. Department of Justice, NIJ Journal; 259.

Daubert vs. Merrel Dow Pharms. Inc 509 US 579, 113 S. CT 2786 (1993).

DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H. (2003) Cues to Deception. Psychol Bull;29:74–118.

Fant, G. (1973) Speech Sounds and Features. Cambridge, MA: The MIT Press.

Fischer, C and Thompson C.R. (2014) The Power of Intuition in Deception and Detection. Academy News, American Academy of Forensic Sciences, 44: 5, 25-28.

Florida vs. Joyner, 2008-CF-21253; Court of the Twelfth Judicial Circuit, Sarasota County, Florida (2008).

Frick RW. (1986) The Prosodic Expression of Anger: Differentiating Threat and Frustration. Aggress Behav;12:121–8.

Frye vs. United States. 293 F. 1013 (DC 1923).

Greaner, J. (1976) Validation of the PSE, [unpublished M.A. Thesis]. Tallahassee, FL, Florida State University.

Haddad, D. S., Walter, S., Ratley, R. and Smith, M. (2002) Investigating and Evaluation of Voice Stress Analysis Technology. Rome AFB, US Dept Justice Report, Grant 98-LB-VX-A103.

Harnsberger, J.D., Hollien, H., Martin, C. and Hollien, K.A. (2009) Stress and Deception in Speech: Evaluating Layered Voice Analysis, J. Forensic Sci., 54: 642-650.

Heisse, J.W. (1976) Audio Stress Analysis - A Validation and Reliability Study of the Psychological Stress Evaluator (PSE). Proceed., Carn. Conf. Crime Counter Measures, Lexington, KY, May 5-6; 5-18.

Hicks, J.W. and Hollien, H. (1981) The Reflection of Stress in Voice - 1: Understanding the Basic Correlates. Proceed., Carn. Conf. Crime Counter Measures, Lexington, KY, May 13-15; 189-194.

Hollien, H. (1980) Vocal Indicators of Psychological Stress. In: E. Wright, C. Bahn and R.W. Rieber, Editors. Forensic Psychology and Psychiatry, New York: New York Academy of Sciences, 47-72.

Hollien, H. (1990) Acoustics of Crime. New York: Plenum Press.

Hollien H, Saletto JA, Miller SK. (1993) Psychological Stress in Voice: New Approach. Studia Phonet Posnan.; 4:5–17.

Hollien, H. (2002) Forensic Voice Identification. London: Academic Press.

Hollien, H., DeJong, G., Martin, C.A., Schwartz, R. and Liljegren, K.J. (2001a) Effects of Ethanol Intoxication on Speech Supra-Segmentals. J. Acoust. Soc. Amer., 110: 3198-3206.

Hollien, H., Geison, L.L. and Hicks, J.W. Jr. (1987) Data on Psychological Stress Evaluators and Voice Lie Detection. J. Forensic Sciences, 32: 405-418.

Hollien H, Harnsberger JD. (2006) The Use of Voice in Security Evaluations. J Cred Asses Witness Psych; 7:74–8.

Hollien, H. and Harnsberger, J.D. (2006) Voice Stress Analyzer Instrumental Evaluation, Final Report. Gainesville, FL: CIFA Contract: FA-4814-04-0011.

Hollien, H. and Harnsberger, J.D. (2013) Assessing Deception by Voice Analysis: Part I, the CVSA. Investigative Sciences Journal. 5: 1-18

Hollien, H. Harnsberger, J.D., Martin, C.A. and Hollien, K.A. (2008) Evaluation of the CVSA Voice Stress Analyzer. J. Forensic Sciences, 53: 183-193.

Hollien, H., Liljegren, K. and Martin, C.A. (2001b) Production of Intoxication States by Actors: Acoustic and Temporal Characteristics. J. Forensic Sciences, 46: 68-73.

Horvath, F. (1978) An Experimental Comparison of the Psychological Stress Evaluator and the Galvanic Skin Response in Detection of Deception. Applied Psychology, 63: 338-344.

Horvath, F. (1979) The Effects of Differential Motivation on Detection of Deception with the Psychological Stress Evaluator and Galvanic Skin Test Response. Applied Psychology, 64: 323-330.

Horvath F. (1982) Detecting Deception: The Promise and the Reality of Voice Stress Analysis. J Forensic Sci.; 27:340–51.

Horvath, F., Jayne, B., Buckley, J., (1993) Differentiation of Truthful and Deceptive Criminal Suspects in Behavior Analysis Interviews. J. Forensic Sci. 39: 793-807.

Horvath, F., Blair, J., Buckley, J., (2008) The Behavioral Analysis Interview: Clarifying the Practice, Theory and Understanding of Its Use and Effectiveness. Int. J. Pol. Sci. and Mgt.; 10: 101-118.

Horvath, F., McCloughan, J., Weatherman, D. and Slowik, S. (2013) The Accuracy Auditors and Layered Voice Analysis (LVA) Operators' Judgments of Truth and Deception During Police Questioning, J. Forensic Scis., 58: 385-392.

Inbar, G.F. and Eden, G. (1976) Psychological Stress Evaluators: EMG Correlations with Voice Tremor. Biological Cybernetics, 24:165-167.

Janniro, M.J. and Cestaro, V.L. (1997) Effectiveness of Detection of Deception Examinations Using the Computer Voice Stress Analyzer. Ft. McClellan, AL: US Department of Defense Polygraph Institute Report No. DoDI96-R-0005.

Klingholz, F., Penning, R. and Liebhardt, E. (1998) Recognition of Low-Level Alcohol Intoxication from the Speech Signal. J. Acoust. Soc. Amer., 84: 929-935.

Kubis, J. (1973) Comparison of Voice Analysis and Polygraph as Lie Detection Procedures. Aberdeen Proving Ground, MD: US Army Land Warfare Laboratory, Technical Report LWL-CR-U3B70.

Leith, W.R., Timmons, J.L. and Sugarman, M.D. (1983) The Use of the Psychological Stress Evaluator with Stutterers. Fluency Disorders, 8: 207-213.

Lippold, O. (1971) Physiological Tremor, Scientific American, 224: 65-73.

Lykken, D., (1981) Tremor in the Blood, New York, McGraw Hill.

Lynch, B.E. and Henry, D.R. (1979) A Validity Study of the Psychological Stess Evaluator. Canadian J. Behavioral Sciences, 11: 89-94.

MacMillian, N.A. and Creelman, C.D. (2005) Detection Theory: A User's Guide (Second Ed.). Lawrence, New Jersey: Erlbaum Associates.

Maier, W., Buller, R. Phillip, M. and Heuser, J. (1988) The Hamilton Anxiety Scale: Reliability, Validity and Sensitivity to Changing Anxiety and Depressive Disorders. Defective Disorders, 14: 61-68.

Marco, C.H. (2001) Certainty of Expert Opinion, Chapter 2 in Forensic Sciences (Wecht, C.H., Ed). New York, Matthew Bender.

McGlone, R.E., Petrie, C. and Frye, J. (1974) Acoustic Analysis of Low-Risk Lies. J. Acoust. Soc. Amer., 55: 520 (A).

McGlone, R.E. (1975) Tests of Psychological Stress Evaluator (PSE) as a Lie and Stress Detector. Proceed., Carn. Conf. Crime Counter Measures, Lexington, KY, May 7-9; 83-86.

National Research Council (2003) The Polygraph and Lie Detection, Washington DC: The National Academic Press.

Netsell, R. (1983) Speech Motor Control: Theoretical Issues with Clinical Impact. In: Clinical Dysarthria. San Diego: College Hill Press, 1-19.

Nolan, J.F. (1983) The Phonetic Bases of Speaker Recognition. Cambridge, UK: University Press.

O'Hair, D., Cody, M.J. and Behnke, R.R. (1985) Communication Apprehension, Vocal Stress and Indices of Deception. Western Speech Comm., 49: 286-300.

Pisoni, D. and Martin, C.S. (1989) Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analysis. Alcohol: Clinical and Exp. Res., 13: 577-587.

Rockwell, P., Buller, D., Burgoon, J. (1997) The Voice of Deceit: Refining and Expanding Vocal Cues to Deception. Commun. Res. Rpts.; 14: 451-459.

Rosenfield JP, Soskins M, Bosh G, Ryan A. (2004) Simple, Effective Countermeasures to P300-based Tests of the Detection of Concealed Information. Psychophysiology; 41:205–219.

Scherer, K.R. (1986) Voice, Stress and Emotion: In: H. Appley and R. Trumbull, Editors. Dynamics of Stress: Physiological and Psychological Social Perspective. New York: Plenum Press, 157-179.

Shipp, T. and Izdebski, K. (1981) Current Evidence for the Existence of Laryngeal Macrotremor and Microtremor. J. Forensic Sciences, 26:501-505.

Siegman AW, Boyle S. (1993) Voices of Fear and Anxiety and Sadness and Depression: the Effects of Speech Rate and Loudness on Fear and Anxiety and Sadness and Depression. J Abnorm Psychol; 102:430–437.

Stevens, K.N. (1971) Sources of Inter- and Intra- Speaker Variability in the Acoustic Properties of Speech Sounds. Proceed., Seventh Internat. Congr. Phonetic Sci. Montreal, Aug 22-28: 206-232.

VanderCar, D.H., Greaner, J., Hibler, N., Speelberger, C.D. and Bloch, S. (1980) A Description and Analysis of the Operation and Validity of the Psychological Stress Evaluator. J. Forensic Sciences, 25: 174-188.

Vrij, A. (2000) Detecting Lies and Deceit: the Psychology of Lying and the Implications for Professional Practice. New York, NY: John Wiley & Sons, Ltd.

Vrij A, Edward K, Roberts KP, Bull R. (2001) Stereotypical Verbal and Nonverbal Responses While Deceiving Others. Pers Soc Psychol Bull; 27:899–909.

Vrij, A., Mann, S., Fisher, R. (2006) An Empirical Test of the Behavioral Analysis Interview. Law Hum. Behav. 30: 329-45.

Warden, R., Drizin, S., editors (2009) True stories of false condessions. Evanston, IL: Northwestern University Press.

Webb, A., Handler, M., Krapohl, D., Kritcher, J., (2008) A Comparison of the Objective Scoring System and Probability Analysis. Polygraph; 37: 250-255

Weinstein, J., Averill, J.R., Option, E.M. and Lazarus, R.S. (1968) Defensive Style and Discrepancy Between Self-Report and Psychological Indexes of Stress. J. Personal and Social Psych., 10: 406-413.

Whitworth, A.W. (1993) Polygraph or CVSA: What's the Truth About Deception Analysis, Law and Order, 29-31.

Williams, C.E. and Stevens, K.N. (1972) Emotions and Speech: Some Acoustical Correlates. J. Acoust. Soc. Amer., 1238-1250.