ISJ investigative sciences journal

Volume 5, Number 2, July 2013

- CORRECTED COPY –

ASSESSING DECEPTION BY VOICE ANALYSIS PART I: THE CVSA

Harry Hollien, PhD.[1]

James D. Harnsberger, PhD[2]

Abstract

Devices that allegedly can detect lying from "analysis" of a person's speech are proliferating. Their presence is creating problems for the courts, as well as for criminal justice and intelligence agencies. But, can assessments of this type actually provide valid information about deceptive behaviors even though past research suggests their ability to do so is suspect? As stated, this report will review prior research, provide information about a laboratory experiment, and analyze recent field research. The experiment was a large, well-controlled, laboratory study of the National Institute of Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA). It employed speech samples of individuals who systematically varied normal with intensely deceptive and stressed speech. To create the latter, they 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 by two teams – one trained/certified by NITV; the other made up of NITV senior examiners – was conducted. The results demonstrated that the CVSA system operated only at about chance levels. This finding is consistent with those reported by other investigators; it is confirmed by field research.

Keywords: Detection of lying; voice stress; deception; speech analysis; speech and deception; Phonetics.

[1] Institute for Advanced Study of the Communicative Process, 68 Dauer Hall, University of Florida, Gainesville, FL 32607

[2] Institute for Advanced Study of the Communicative Process, 68 Dauer Hall, University of Florida, Gainseville, FL 32607

Forward

A number of the authors' professional acquaintances, ones who specialize in law enforcement, work within the judicial system or are in the military, have indicated that a substantial number of their colleagues are not familiar with the *science* associated with lie detection. And even if they are, their familiarity tends to be limited to the bases for polygraphy. It appears that about all they know about voice stress/deception devices are articles/information found in the media (example: Bransletter and Brunette, 2006) or stories told by agents about their "experiences" (example: Whitworth, 1993). Moreover, we personally encountered the confusion expressed by many of these professionals when we conduct field research or engage in forensic activities. In addition, we have recently received a number of inquiries from agents and attorneys asking about the "use" of the PSE, CVSA and LVA; these inquiries have led to participation in several relevant cases (see below). Accordingly, it would appear useful to provide a review – aimed specifically at these professionals – that focuses on an interpretation of the research literature and its impact on everyday investigations, criminal/civil cases and counter-intelligence needs. We will include several of our own studies (Hollien et al, 1987; Hollien et al, 2008; Harnsberger et al, 2009) for illustrative purposes but will start with a discussion of research by others and conclude by consideration of certain unpublished studies and some new research.

Prologue

One of the more serious problems currently facing members of the judicial, criminal justice and intelligence communities involves their ability to determine when a speaker is telling the truth. Law enforcement and intelligence agents must endlessly assess statements made by suspects or informants for truth or deception. Attorneys face the same problem with clients and witnesses. This situation is so pervasive that any procedure or system which promises to identify truthful utterances – and separate them from falsehoods – would be of both great interest and substantial value.

Of the many proposed approaches are devices which are said to be capable of analyzing speech in order to detect stress and/or deception. Indeed, they are promoted and advertised extensively. Best yet, they appear to be easy to obtain and use. But, do they perform as advertised?

The essay to follow has been structured in an effort to provide a useful response to this question. In this case, the first of two major approaches will be described and evaluated. It is the National Institute for Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA) which is a modernization of the Psychological Stress Evaluator (PSE); the second article will review Nemesysco's Layered Voice Analyzer (LVA). Relevant research about the CVSA (some of which was carried out by the present authors) will be summarized and its import discussed. In this regard, it is important to note that the processes involved in researching these devices actually parallels investigations conducted by attorneys when they analyze a client's case. So too do law enforcement professionals operate in this manner. They investigate crimes, using observations and analyses, in order to determine if a suspect is guilty or not. Other analogies could be cited but, as can be seen, the process of gathering and interpreting information carried out by agents and attorneys is similar to the research that has to be conducted in assessing these

voice truth detectors. Admittedly, the jargon here may differ a little from theirs, but the present authors will attempt to keep these differences to a minimum. Let us start with a brief review.

Introduction

It is not difficult to imagine the societal changes which would occur if we were able to reliably detect when an individual is lying. Consider the effect it would have on advertising – or on family relationships. Even more important, consider what would happen if we could determine politicians' beliefs, and (especially) their intent, simply by listening to them speak?

Without question, the availability of a valid system of this type would also be of inestimable value to the courts and criminal justice systems – and to operations carried out by counter-intelligence agents. For one thing, there would be no need for trials by jury. The guilt or innocence of anyone accused of a crime *actually* could be discovered by asking them: "Did you do it?" and analyzing their response.

But, for this to qualify as a "Lie", the observed behavior would have to be *measureable* and everyone would have to exhibit that specific feature (or a cluster of them) *whenever* they lied. In addition, these same traits could not be indicators of any other type of behavior. It was Lykken (1981) who seems to have best articulated the key concept here. He argues that if a lie is to be detected, a lie response – i.e., a measureable physiological or psychological reaction – must always occur. He then said: "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."

But what do the manufacturers of "voice stress analyzers" claim their devices are able to do? Well, for some years now they have asserted that they actually *can* detect stress and deception by means of their particular type of voice analysis. As a consequence, the use of Psychological Stress Evaluators (PSE), Voice Stress Analyzers (VSA) and similar systems has become rather widespread. Indeed, they, and the more up-to-date ones like the CVSA, are presently employed by a number of law enforcement, security, military and intelligence organizations. Of course, testimony based on their "analyses" has not yet been accepted in Courts-of-Law. The reason such testimony is not found there is that it simply does not come close to being admissible with respect to Daubert or Frye (1993, 1923), much less satisfy 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 this type. On the other hand, their presence cannot be ignored as their use is extensive and it forces itself (at least tangentially) onto both investigations and trials. Indeed, the present authors have been consulted a number of times (two examples: Florida vs. Joiner, 2012; Massachusetts vs. Perrier, 2013) about the nature and validity of "stress analysis" for use in both criminal and civil cases.

But is there any chance at all that these voice-stress evaluators are but acceptable evaluation procedures "in waiting"? In that regard, it must be conceded that it is well known that human speech and voice **does** contain features which can be useful in providing 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). Finally, human emotions (including psychological stress) can 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 – even though such is not actually the case – the manufacturers of "voice stress" devices argue that the relationships found in the stress research can provide a basis for the use of their products. A second unintended consequence, resulting from this situation, is that the voice/stress controversy has overshadowed (and, to some extent, stifled) the legitimate research being carried out in that area.

Moreover, it must also be stressed that the psychological and neurological substructure, for the behaviors reviewed above, are not all that simple; indeed, they 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 tend to predict that even subtle operations – ones such as the detection of deception and/or truth – *should* be possible, the simple (and primitive) mechanisms of the "stress evaluators" probably will not be effective.

In this regard, please consider the following. The CVSA proponents argue that their devices identify lying or stress by measuring the *microtremors* which occur in the laryngeal muscles. Unfortunately, microtremors 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). Yet, they further argue that they measure the "shifts" which occur (in these "microtremors") when the presence of stress and deception inhibit them. On the other hand, and as a result of these statements, they are faced with two problems. The first is that it has not been demonstrated that stress always (or even usually) accompanies deception. The second is that there is no evidence that the microtremors would affect voice and speech even if they did exist. In any event, the CVSA manufacturers offer no scientific evidence to support their claims, nor have they published any research (except Heisse's, 1976) about their method's validity.

Is There Any Research on Physiological Stress Evaluators?

Yes, there has been. As you might expect, the lack of information here has led independent investigators to study the CVSA system and its predecessors. These efforts have ranged from occasionally mixed reviews to mostly negative ones. That is, while some investigators have suggested that these devices *might* detect stress – at least under certain circumstances (McGlone, 1975; Brockway et al, 1976; Vandercar et al, 1980) – most of the relevant data simply have not supported the manufacturer's claims (see Horvath, 1979; Cestaro, 1996; Jangiro and Cestaro, 1997 and those to follow). On the other hand, it can also be said that, to date anyway, none of the "voice stress" systems have been afforded an extensive and comprehensive assessment. The NITV group also complains that investigators either 1) did not control their procedures at a level which would permit robust data to be generated (Lynch and Henry, 1979; O'Hair et al, 1985), 2) assessed or controlled too few variables (McGlone et al, 1974; Greanor, 1976; Leith et al, 1983), 3) 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 most of these projects

exhibited rather good research protocol and none of them provided data that could be employed to argue that the PSE/CVSA devices are able to detect falsehoods at levels exceeding chance.

However, even our own research (on the original PSE system) was not as definitive as would be desirable (Hollien et al, 1987). In that investigation, we used subjects who were committed anti-vivisectionists; they were required to vigorously attack the position that animals are abused and/or worse in research and related activities. As expected, this task was shown by self-reports and standard tests of stress to produce conditions of high stress and/or significant anger when the subject uttered the required "lies". Yet the PSE was unable to identify these conditions (i.e., either that of lying or of high stress). In short, even though limited (no control group was included; only the "hit" rate was assessed), the obtained data here demonstrated that the device only operated at about chance levels. Thus, at that juncture it appeared necessary to conduct research on large relevant populations and do so under fair, but extensive and controlled, conditions. We did just that and this is why a review of that project will be featured in this essay.

Our Basic Experiment

Now comes the National Institute of Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA) (see NITV Examiners Manual, 2005). The makers of this system argue that their device *does* work. As stated above, they further complain that the negative research which has been published (i.e., the above studies cited) actually resulted from conditions where the procedures employed: 1) were not realistic enough, 2) did not include stress levels which were properly controlled, 3) did not induce sufficiently high levels of jeopardy in the deception states examined, and 4) did not include reasonably diverse populations. In short, they claimed that their devices simply had not been studied fairly or under acceptable conditions.

In response, the present authors undertook an experiment that took these claims into account. That is, the research carried out was designed to determine if people were telling the truth -- or falsehoods -- under conditions of rigorously controlled high and low jeopardy and high and low stress (Hollien and Harnsberger, 2006). Specifically, a large, *double blind* experiment was conducted. It was structured both as rigorous research and also to meet the complaints made by the PSE/VSA groups (see Hollien et al 2008 for the published scientific description).

Structuring the Experiment

Before the reader can be expected to evaluate the validity of an investigation such as the one to be described here, they need to understand just how it was structured and carried out. To this end, the procedures employed must be presented so their strengths and weaknesses can be judged. Accordingly, we will attempt to provide enough detail for you to determine if we have, indeed, designed (and carried out) research which meets both the desired level of accuracy and validity. Please note also that, while it is presented only in brief here, its full description also is available (Hollien et al, 2008).

The Speakers: included **s**eventy-eight adult volunteers, both 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, the military, students, housewives, clergy, nurses, police, and so on). To be included, they also had to hold a

very strong personal view 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: Being mindful of the problems associated with the external determination of the internal states of humans (Weinstein et al, 1968; Bradley and Stocia, 2004), four methods were administered at the appropriate times for the assessment of psychological stress. 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 Speech Samples: The nature of the speech samples employed in this research was especially important. Of the seven different types of utterances produced by each subject-speaker (i.e., for this and related studies) only the two which are relevant to this review will be discussed here. However, all of the samples consisted of a speech passage structured with 5-7 content sentences and a 17-26 word 'content neutral' sentence (but one that fit the related context) embedded near the center of the passage. The neutral sentence was inserted so that speech could be analyzed which contained no revealing language cues but were at the same stress or deception level as the full passage. The use of these 'content neutral' phrases prevented any of the CVSA operators from being exposed to language-based clues relative to the judgments they were being asked to make. All speech samples were uttered three to five times with only that sample used where all experimental criteria were met. Subjects were permitted to become familiar with all spoken material prior to the run. The two classes of samples reported here are:

> **Low-Stress Truth.** Each subject uttered a truthful passage; the content of which was on a pre-designated unemotional topic. Examples include a description of some pleasant feature about where they live or 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. 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. Secondly, subjects were instructed to produce these lies in a speaking style that strongly suggested that they actually believed them. Third, the stress of lying was enhanced by selected use of an electric shock procedure which could be 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

stress. The dual stress/lie procedure was so intense that all of the (selected) subject's untruthful utterances were shown to reflect extremely high jeopardy.

Finally, it should be stressed that, while the contrast between samples of low stress truth vs. high stress/high jeopardy lies are featured in this report, all of the many other contrasts produced similar data and results (Hollien and Harnsberger, 2006).

The Procedure: As you would expect, the procedures met all subject safety requirements. The individuals included also were assigned numbers (to ensure anonymity) and 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. All volunteers were paid for their participation. Incidentally, the Psychiatrist also attempted to add an element of uncertainty to the interview in order to heighten the selected subjects' arousal for the high stress/deception condition which was presented first.

After a trial, each of the four stress-level checks were averaged and converted to a common scale. Only those subjects were included in the project whose mean stress level, when lying with jeopardy, was always more than double their baseline stress level. Stated differently, the mean *overall* stress shift observed for all subjects was 141% -- with a mean rise of 129% for male speakers and 152% for females. Forty-eight individuals met all requirements and completed the runs. Their speech materials were then organized (by gender) into appropriate sets of 300 samples for each evaluation by the CVSA operators.

The Evaluations: The CVSA device is designed to process very short speech samples (i.e., ones that actually can be as brief as "yes" and "no") and its output can be described as a two dimensional chart (see Figure 1) displaying the duration of the speech signal on the horizontal axis; an explanation of the nature of the vertical axis is not provided (NITV, Examiners Manual, 2005). The left-hand chart is said to display a voice recording in which psychological stress (and/or deception) is present; its rectangular shape would be referred to by NITV personnel as "blocking". In turn, the right-hand chart, with its triangular form, would be said to display a voice recording in which psychological stress, and hence deception, can be presumed absent. As indicated, the NITV Manual (2005) states that blocking results from the suppression of the natural "micro tremor" within the muscles that control the vocal folds and speech articulation. It is claimed that, when this micro-tremor is suppressed, its acoustic byproduct -- referred to as the "inaudible frequency modulation (FM) component" -- is lost. In turn, when the subject is no longer under stress, the micro-tremor is said to return and blocking dissipates. Even though these explanations do not fit the known physiology of the "voice" (see again Shipp and Izdebski, 1981), great care was taken to follow the manufacturer's instructions so that *their* procedure would be in no way compromised. Accordingly, the samples selected were the required single syllable utterances which 1) occurred at the maximum of the physiological measures, 2) exhibited no artificially abrupt onset or offset, and 3) were only in the modal (normal) vocal register.

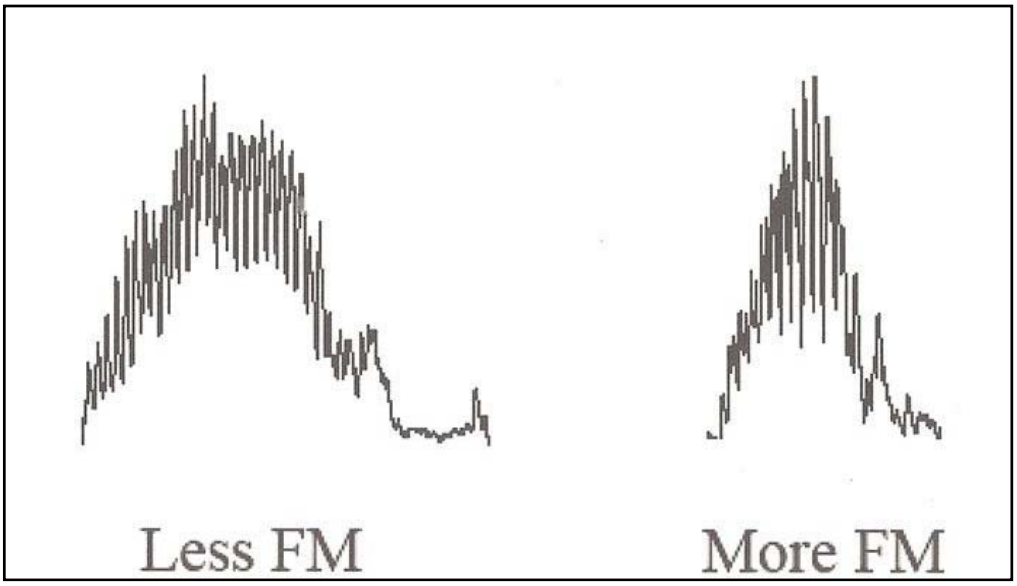

Figure 1 – CSVA Charts

The left chart ("Less FM") shows "blocking" which presumably is due to stress and/or deception (time frame=861 ms). The right chart ("More FM") shows an absence of blocking (time frame=474 ms), a pattern which would be interpreted as showing an unstressed and/or non-deceptive utterance (drawn from NITV manual, 2005).

The Evaluators: The CVSA processing described above was carried out by two teams of examiners. The first consisted of an independent team from the University of Florida's Institute for the Advanced Study of the Communication Processes (IASCP) who attended (and completed) the NITV school and were certified by them as competent to conduct CVSA analyses. The second evaluation team consisted of three highly experienced (senior) operators provided by NITV; they traveled to the University of Florida to participate in the study. All team members (IASCP and NITV) individually classified all samples as deceptive or truthful, and stressed or non-stressed.

The Results of this Experiment

Assessment Techniques: The resulting data were assessed by a number of statistical procedures designed to explore the possibility that the CVSA system might be sensitive to deception (as well as to truth, to stress, or lack of them, and so on). In each case, four values were calculated: they were *true positive, false positive, false negative,* and *true negative*. The true positive rate refers to the percentage of the time that any assessment (deception, in this case) is said to be present when in fact it actually is present. In other words, the "true positive" tells how often the CVSA classified a deceptive utterance as a lie -- or, in reverse circumstances, identified truth as truth. Moreover, if the assessment is to be an *accurate* one, the false positive rates also must be calculated. They correspond to the percentage of times the condition (deception in this case) is

said to be present when in fact it is absent. False positive rates **must** be compared with true positive rates in order to determine if a device such as this one can correctly discern the presence of the factor under study. That is, assessing only the "true positive" value alone can be *very* misleading or lead to misjudgments. By itself, it simply does **not** permit assessment of a system's accuracy since it can be the product of either accurate judgments or simply a bias by the operator or device to classify stimuli as positive regardless of the actual presence or absence of the target behavior. Hence, an **accurate** device will show true positive rates that are both high and significantly different from the false positive ones (which will be low). On the other hand, a device that performs *at or near chance* would show relatively similar true and false positive rates no matter what their actual level.

Of course, the false negative and true negatives rates also were determined. Specifically, the false negatives occur when the signal is present but the device fails to detect it. True negatives are cases in which the behavior is in fact absent and is accurately judged thusly (e.g. truthful speech samples that are classified by a device as truthful, and so on). This procedure can be applied separately in order to determine if a speaker is telling the truth.

The Findings: To reiterate, detection of 'blocking' (or non-blocking) on the two sets of 300 CVSA charts was performed both by the IASCP and NITV teams. Each operator did so separately and without knowledge of any of the other operator's work. Their decisions were given to an independent investigator (i.e., the project's PI) 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 results would be to focus on the most important contrast of 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 Figure 2 (next page). 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 NITV (i.e., the manufacturer's) team 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.

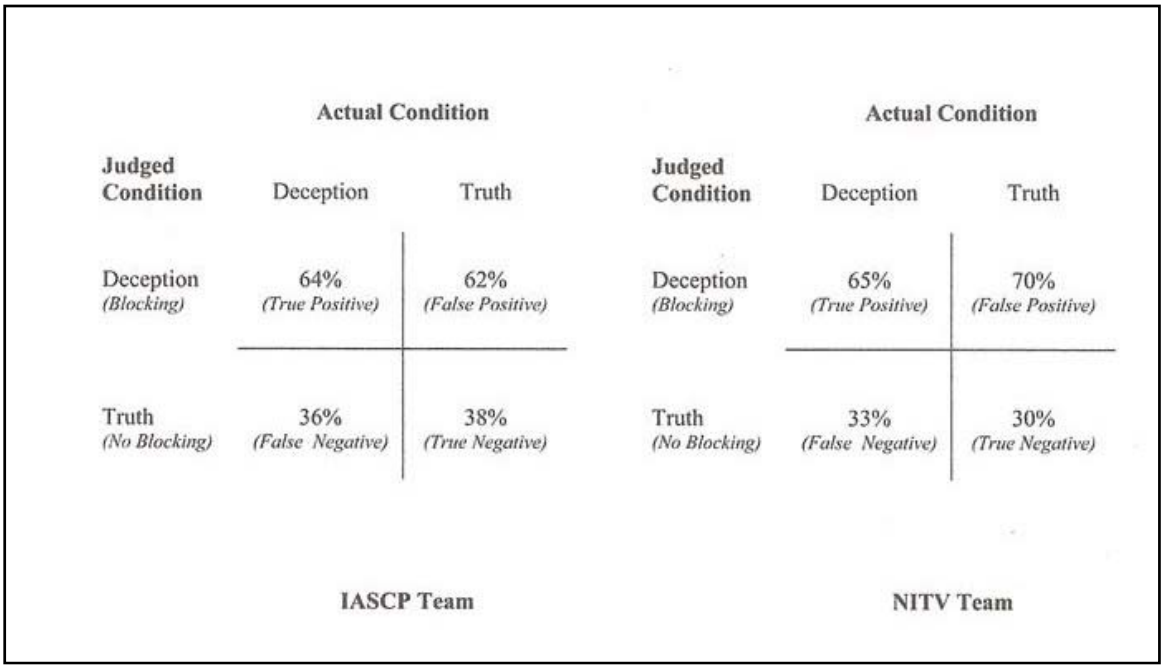

Figure 2

Identification of deception and truth in the speech samples by both NITV trained evaluation teams (left side: IASCP; right: NITV). The top left value (judged as deception when deception actually existed) and the lower right cell (actual and judged truth) are among the contrasts. Note that the highest values are in the critical "false positive" category (from Hollien et al, 2008).

As can be observed, at 64% correct, the deception identifications for the IASCP team were above 50-50 and those for the NITV group yet a little higher (i.e., 65%). If these scores are taken alone, it can be argued that the CVSA system, while not a strong detection device, at least be used to identify lying more often than missing it. 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 almost as often as they did the ones that actually were deceptive (i.e., at 62%) and the NITV team did even worse, calling a whopping 70% of the truthful utterances lies! This obvious bias should remind the reader of the old adage: "If you are a hammer, almost everything you see will look like a nail."

Moreover, while it was expected that the three NITV evaluators, being seasoned operators, would perform better than the two individuals from the University of Florida, such was not the case. While at 65% to 64%, it is conceded that the NITV groups did show a slightly greater propensity to classify the relevant charts as showing 'blocking' (i.e., the utterances to be deceptive), this mild bias was sharply reversed for the false positives (70% to 62%). Hence, the NITV team actually exhibited an overall lower accuracy level.

Also clear was the inability for the device, as used by either group, to recognize truthful statements when they occurred (IASCP=38%, NITV=30%). In short, the two data sets suggest that neither of the teams was able to correctly identify either deception or truth when they

occurred. In turn, this relationship suggests that the device simply did not provide operators with appropriate information. In summary, it would appear that the CVSA is insensitive to the 'signal' in question and hence, only operates at about chance when attempting to spot lies from voice analysis.

As it turns out, the statistical analyses which were applied validated the above statement. For one thing, repeated measures ANOVA's were carried out; all were found to be non-significant for both teams and all conditions. Such statistical "non-significance" argues that accurate identification is simply not occurring at a reasonable level. Secondly, 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 minimal sensitivity.

In short, it must be said that the CVSA system simply did not display the ability to detect the presence of deception, truth and/or stress level. Nonetheless, even though the evaluation of both the data and statistics would suggest but minimal performance, the CVSA's defenders probably would attempt to advance alternate interpretations For example, they might argue that the results could reflect "limitations" – ones where they claim 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., such as those resulting from 'real-world' interrogations by police or the military). Indeed, this interpretation might appear cogent (marginally anyway) if only the true positive rates were assessed. However, and as was pointed out a number of times, the evaluation of CVSA's performance on truthful and unstressed speech samples served as an important control: one that permitted the examination of that device's potential bias to identify speech samples as deceptive in either the presence or *absence* of that state. That is, if the speech samples contained inadequate levels of 'real-world' stress (or deception), then false positive rates near zero would be expected. Such was not the case. Rather, the system misclassified the low stress and truthful samples as lying with great frequency. These very high false positive rates simply cannot be explained by arguing that inherent limitations within the laboratory protocols existed. Indeed, quite the contrary is true.

At this juncture it can be asked how well these findings agreed with prior research – and, for that matter, with new (and/or field) research. First, it can be said that nearly all of the investigators who have studied the PSE/VSA systems in the past have reported that those devices did not exhibit the ability to properly identify either deception or high stress. In that regard, the data presented here are consistent with virtually all of theirs. They are most like those reported by Horvath and his associates (Horvath, 1978, 1979; Horvath et al, 2013; Cestaro, 1996; Jannino and Cestaro, 1997) and our group (Hollien et al, 1987).

The results of two field investigations also serve to underscore the findings of the laboratory experiment described in the present article. The first of these (Hollien and Harnsberger, 2006) was populated by military personnel participating in a SERE (Survival, Escape, Resistance, Evasion) survival course. It should be noted that the SERE program includes rigorous training where students are disciplined not to reveal any information when captured and interrogated by hostile forces. In this instance, the speech samples were obtained during a period of rigorous interrogation which was carried out when they (i.e., the students) were experiencing stress and fatigue. This type of research is referred to as a "guilty knowledge" procedure. Here,

the goal was for the students (both male and female) to lie and then have the interrogators attempt to detect the falsehoods which were embedded in a large group of truthful responses. In turn, the students faced punishment if their untruths were detected and, hence, had to lie at a high level of jeopardy (measured by the Vivometrics system contained in the shirt they wore). Speech samples of the type suitable for CVSA processing were sent to the project's Principal Investigator. He had them properly processed for evaluation by the two teams (IASCP and NITV) who made the evaluations in the cited (laboratory) experiment. These materials were assessed in the same manner as were the initial ones. Incidentally matching low stress foil samples (drawn from other SERE samples) were added to the pool in order to ensure a low-high stress balance. Surprisingly, the true positive rates (deception-deception) were much lower across all teams than were the false positive (i.e., truth "detected" as lies). The true negative ones (truth-truth) were higher but still hovered around 50%. Again, the statistical analyses demonstrated only chance detection rates were obtained. Thus, this "real life" research demonstrated but chance detection levels and, in doing so, confirmed the findings of the prior experiment(s).

The second study (Damphausse et al, 2007) also took place in a "real world" setting. The investigators asked a large number of arrestees, being incarcerated at a county jail, pertinent questions about recent drug use. Their answers were processed for "evaluation" by NITV's CVSA system. The operators in this instance were a number of individuals trained and certified by NITV; indeed, one of the groups consisted of CVSA instructors who were very experienced in chart reading. The "ground truth" (i.e., veracity of the data) was established by urine tests for five drugs (cocaine, opiates, marijuana, etc.). The investigators indicated that "the results were not very promising" with the CVSA consistently failing "to correctly identify respondents who were being deceptive." They indicate that "only about 15% of the respondents who had recently used drugs", but indicated that they had not done so, "were identified as being deceptive." That work essentially supports both the laboratory and field results discussed above. In turn, the corpus of information resulting from all three of these new investigations (i.e., the laboratory, plus two field, studies) also provides an extension to (and agreement with) The National Research Council's (2003) conclusion that there was no scientific support for "voice analysis".

Finally, it must be noted that, even though the CVSA 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) might actually exist (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). Or perhaps the CVSA was used simply 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 output of this type of device. If it is encountered, the best course of action probably would be to either avoid it or to challenge it.

Acknowledgements

The research project that was 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 recognize the fine cooperation afforded the project by NITV and its president Dr. Charles Humble. We also are grateful for the excellent support provided by the NITV staff/operators -- especially with respect to chart reading.

A Postscript

Please note that Part II of this set of reports will consider – and evaluate – Nemesysco's LVA device. Again, that review will feature our laboratory experiment (Harnsberger et al, 2009). However, in this case it will be accompanied by one of the best designed/conducted field investigations (of any type) to be reported in some time. See Horvath et al, 2013.

References

Abba, 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.

Barland, G. (1975) Detection of Deception in Criminal Suspects [unpublished Ph.D. dissertation] 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 Associate, 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 Comparative Study, Springfield, IL, C.C. Thomas.

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, Supreme Court Indictment No. 121421 (2013).

Cummings, K. and Clements, M. (1994) Analysis of Glotal Excitation 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.

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

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

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

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. (2002) Forensic Voice Identification. London: Academic Press.

Hollien, H., DeJong, G., Martin, C.A., Schwartz, R. and Liljegren, K.J. (2001) 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. and Harnesberger, J.D. (2006) Voice Stress Analyzer Instrumental Evaluation, Final Report. Gainesville, FL: CIFA Contract: FA-4814-04-0011.

Hollien, H. Harnesberger, 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. (2001) Production of Intoxication States by Actors: Acoustic and Temporal Characteristics. J. Forensic Sciences, 46: 68-73.

Hollien, H. and Schwartz, R. (2001) Speaker Identification / Utilizing Non-contemporary Speech. J. Forensic Sciences, 46: 63-67.

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., McCloughhan, 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.

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.

Massachusetts vs. Perrier, C-501-58684 Superior Court of Hampden County, MA, Doc. No. 12-142 (2013).

McGlone, R.E., Petrie, C. and Frye, J. (1974) Acoustic Analysis of Low-Risk Lies. J. Acoust. Soc. Amer., 55: S20 (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 Institute for Truth Verification. (2005) Certified Examiners Course Manual. West Palm Beach, FL.
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.

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.

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.

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.