J Forensic Sci, May 2009, Vol. 54, No. 3
doi: 10.1111/j.1556-4029.2009.01026.x
Available online at: www.blackwell-synergy.com

James D. Harnsberger,[1] *Ph.D.; Harry Hollien,*[1] *Ph.D.; Camilo A. Martin,*[2] *M.D.; and Kevin A. Hollien,*[1,3,4] *B.A.*

Stress and Deception in Speech: Evaluating Layered Voice Analysis*

ABSTRACT: This study was designed to evaluate commonly used voice stress analyzers—in this case the layered voice analysis (LVA) system. The research protocol involved the use of a speech database containing materials recorded while highly controlled deception and stress levels were systematically varied. Subjects were 24 each males/females (age range 18–63 years) drawn from a diverse population. All held strong views about some issue; they were required to make intense contradictory statements while believing that they would be heard/seen by peers. The LVA system was then evaluated by means of a double blind study using two types of examiners: a pair of scientists trained and certified by the manufacturer in the proper use of the system and two highly experienced LVA instructors provided by this same firm. The results showed that the "true positive" (or hit) rates for all examiners averaged near chance (42–56%) for all conditions, types of materials (e.g., stress vs. unstressed, truth vs. deception), and examiners (scientists vs. manufacturers). Most importantly, the false positive rate was very high, ranging from 40% to 65%. Sensitivity statistics confirmed that the LVA system operated at about chance levels in the detection of truth, deception, and the presence of high and low vocal stress states.

KEYWORDS: forensic science, psychological stress, deception, speech, voice stress, phonetics, layered voice analysis

It is well known that human oral communication contains features, which can be used to provide useful information about a speaker apart from the linguistic content, or meaning, of what was said, and that this *indexical* information can prove exceedingly helpful in forensic work. That is, indexical information includes all aspects of the speech signal in addition to the meaning of the utterance itself. Included are the identity of the speaker, a speaker's age, whether or not they are intoxicated, and their language and dialect. Also very important is human emotion (including stress), which constitutes another domain where relevant behaviors can be detected (1–14). Many of the effects of stress, especially but also lying to some extent, have been established (2,7,11,15–30). While it is recognized that lying does not always result in vocal stress (due to sociopathic conditions, stress muted by certain chemical substances and so on), it is still thought to constitute the substrata for deception in the majority of instances (31–37).

Given the general relationships that have been observed between speech and voice, deception, and psychological stress, it is not surprising that commercial products purporting to measure the acoustic correlates to stress and deception in speech have been marketed for over 30 years to both law enforcement and national security agencies. The forensic applications of a successful device would be numerous, from detecting deception in the statements of claimants, witnesses, or suspects, to assessing whether or not an individual is withholding information or providing incomplete information, to analyzing older audio recordings of speech, and to covertly monitoring the speech of individuals of interest to investigators. In fact, a significant number of commercial "voice stress analysis" (VSA) systems have been brought to market but, to date, attempts to verify their efficacy have not been very successful. While a few authors suggest that these devices might possibly detect deception-related psychological stress, at least in certain circumstances (38–41), most research has not supported this position (31,34,35,37,42–44). A recent review of this literature is included in the National Research Council's 2003 report on the polygraph and other methods of deception detection and credibility assessment, entitled "The Polygraph and Lie Detection." The scientific panel concluded that "Overall, this research and the few controlled tests conducted over the past decade offer little or no scientific basis for the use of the computer voice stress analyzer or similar voice measurement instruments as an alternative to the polygraph for the detection of deception. The practical performance of VSA for detecting deception has not been impressive" (p. 168).

However, in evaluating these studies, it must be remembered that many investigators have been limited in their ability to elicit stressed, and/or deceptive speech samples with a sufficient degree of control (33,45–47). Others have examined an insufficient number of variables (23,32,48–50) or they have carried out only limited laboratory studies even if reasonably well controlled (34,51–53). In most instances, researchers have employed a class of experiments that can be described as "simulated field" studies (36,54). Studies of this type ordinarily involve testing subjects in the laboratory via fairly elaborate "games"—ones which attempt to mimic naturalistic settings where individuals produce lies that, if uncovered, would expose them to some type of jeopardy or punishment. The motivation behind these simulated field studies appears to be the desire of the investigator to evaluate the VSA system under the most ecologically valid conditions. Moreover, some scholars argue that more controlled laboratory experiments are simply "games" and as they are "unrealistic," they provide little-to-no useful information (55). The counterarguments to that position are that (i) field research ignores the need for basic system assessment under controlled conditions, (ii) it does not include events necessary for the proper

[1]Institute for Advanced Study of the Communication Processes, University of Florida, Gainesville, FL 32611.
[2]Veteran's Administration Medical Center and University of Florida Psychiatry Department, Gainesville, FL 32601.
[3]Forensic Communication Associates, Box 12323, University Station, Gainesville, FL 32604.
[4]Posthumous.
*This research was supported by CIFA contract FA-4814-04-0011.

Received 29 April 2007; and in revised form 21 July 2008; accepted 2 Aug. 2008.

determination of system operation, (iii) it does not exclude debilitating external variables, and (iv) knowledge is lacking on the speaker's actual behavioral states.

These counterarguments reveal the need for highly controlled and relevant laboratory studies, which can provide both reliable information and the structure for further field-based work. Hollien and Harnsberger (56–58) have modeled a three-stage research program for this purpose. It systematically examines stress and deception, both independently and in concert, at several levels. The program begins with laboratory studies designed to permit an understanding of the most fundamental relationships between speech articulation and the behavioral states of psychological stress and the intent to deceive. Once those relationships are understood under controlled, and highly structured conditions, field-based studies of several types can be conducted that maintain, nonetheless, some degree of control while data are collected in more ecologically valid (e.g., "realistic") settings.

Model

The three-level model cited above (56–58) was employed as the basis for this study. The first level involves extensive and highly controlled laboratory experiments. Here, utterances involving truthfulness, deception, psychological stress (and, perhaps other emotions)—at various levels of intensity—are obtained from a variety of speakers. These behaviors are experimentally induced, are relevant and their presence is verifiable by independent assessment. The model's second level focuses on both (i) simulated field and (ii) actual field research, but studies where only modest levels of control and verification are possible. At this level, either field scenarios are created where subjects are involved in a stressful encounter (i) such as survival training or (ii) resulting from actual cases (usually criminal) where interrogation was carried out. The model's third level involves actual field experiments—often referred to as "real life" studies—but those where the data are also obtained under conditions of high level control and validation.

The present study reports on a large first-level (laboratory) experiment and the application of these rigorous approaches in the evaluation of a specific VSA device, the layered voice analysis (LVA) device produced by Nemesysco (Natania, Israel). Its manufacturer claims that it is capable of detecting not only deception in speech, but a variety of emotional and cognitive states, such as emotional stress, cognitive effort, fear of discussing a particular topic, stress due to deception, anxiety, arousal, condescending attitude, physical attraction, and many others. For this study, only the device's sensitivity to the presence of deception and of stress was evaluated. LVA represents a recent iteration of a line of products that include brand names such as Truster, Truster-Pro, Vericator (Nemesysco); currently these are marketed in conjunction with related products such as SENSE (Nemesysco) and VoiceSum (Nemesysco).

Nemesysco claims that LVA's sensitivity to a large variety of emotional and cognitive states is based on methods that are distinctively different from previous voice stress analyzers, such as the Psychological Stress Evaluator, the Computer Voice Stress Analyzer, (National Institute for Truth Verification, West Palm Beach, FL) and other commercial predecessors. Such voice stress analyzers in one form or another rely on measurements of the acoustic consequences of hypothetical "microtremors" of the laryngeal muscles employed in vocalization. In contrast, the manufacturer states, in version 06.50.3 of the LVA manual that the device "performs a wide-spectrum analysis, uses an automatic calibration and filters through emotion levels." In training materials provided to the lead author, who was certified in the use of the device, LVA is 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." These descriptions are inadequate to make even the most basic characterizations of what type(s) of acoustic information are extracted from the speech signal, and how these might be processed by LVA. For instance, the phrase "wide spectrum analysis" is not one typically used by speech scientists. It might refer to a power spectrum, which displays the amplitude and frequencies of the periodic components of complex waves, which includes speech signals. It might also refer to a wide-band spectrogram, which provides both frequency and intensity information concerning the speech signal at a high temporal resolution. "129 voice frequencies" could refer to a bank of bandpass filters that span the range of audible frequencies, or not; the expression "voice frequencies" does not correspond specifically to any particular class of analysis techniques employed by speech scientists. In summary, the device's overall performance was evaluated, and no conclusions could be drawn concerning different, specific methods used by LVA.

Specific Goals

As implied, the purpose of this study was to generate highly controlled speech materials, which could be used to validly test the ability of a specific device to identify people when they were (i) speaking the truth, (ii) telling a falsehood, (iii) talking while highly stressed, or (iv) producing unstressed speech. Specifically, results are reported for evaluation of the LVA, a device that purportedly detects lying and emotional or psychological stress independently of one other (in contrast with many other voice stress analyzers on the market). This instrument was tested in a large double-blind laboratory study, one that did not permit the on-scene operators to directly respond to events involving human subjects. It was only through the use of this type of controlled approach that the characteristics of the device itself could be evaluated in a thorough and impartial manner. It was critical to separate the performance of the device from any latent abilities of an operator to detect by ear deception cues directly from the audio samples. The LVA device is designed to *automatically* classify audio recordings of speech in terms of the presence of deception and emotional stress, among many other cognitive states, without input from human operators. It was these automatic functions that were evaluated in the study.

Method

The protocol employed has been previously described in some detail (56–58). However, it will be briefly reviewed here to insure a reasonable interpretation of the results obtained.

Subjects

Seventy-eight male/female volunteers, ranging in age from 18 to 63 years were screened; they represented a diverse demographic sample of the U.S. population. Further, they had to report holding very strong personal views about some subject (e.g., politics, religion, Iraq, etc.). Participants were recruited from local political, religious, or cultural organizations. They were all screened by the project's psychiatrist (third author), who excluded any with medical conditions or who showed a past history of psychological trauma. Subsequently, other potential exclusionary mental and physical

health criteria (the use of drugs, for example) also were assessed and used in the selection process.

Recording Procedures

The volunteers selected were recorded in a quiet room with two microphones (Shure SM-10A head-mounted and Sony ECM-737; Sony Corporation of America, New York, NY) feeding (i) a Sony TCD-D8 DAT recorder, (Sony Corporation of America) (ii) a digitizer (Model MP-150, BIOPAC Systems, Inc., Goleta, CA) coupled to a computer, and (iii) a Marantz PMD-221 analog cassette recorder; (Marantz America, Inc., Mahwah, NJ) all equipment was calibrated. Additionally, digital audio–video recordings of each subject were made during all experimental runs using a Sony DCR-HC21. The videocamera was fixed and focused on the subject's upper body.

Measurement of Stress Levels

Four methods for the measurement of psychological arousal and/or stress were administered, either continually or once after completion of each experimental procedure. They were: (i) two tests of anxiety/stress level based on self-reports (administered after each experimental condition), and (ii) continual body response evaluations consisting of galvanic skin response (GSR) and pulse rate (PR). The anxiety/stress tests consisted of an "emotion felt" anxiety checklist and a modified version of the Hamilton test (59). GSR and pulse were measured using the BIOPAC Systems, Model MP-150.

Speech Samples

Following a familiarization process and baseline calibration, six different types of utterances were produced by each subject. Each experimental passage consisted of five to seven sentences, within which, a 17–25 word "content neutral" sentence was embedded (near its center). "Content neutral" refers to the absence of *any* information that was specific to the topic of the passage. It was inserted so it could be uttered at the same stress level as was the full passage and later be excised for separate analysis. The use of these embedded sentences in the evaluation of the LVA prevented any of the operators from being exposed to language-based clues as to the type of speech being produced.

The complete set of speech samples is described below. All were produced three to five times with only that sample meeting all criteria used in the evaluations:

Baseline (Calibration Sample)—All subjects read a standardized phonetically balanced (unstressed) truthful passage, namely the Rainbow Passage.

Sample 1: Low-Stress Truth—Each subject read a truthful passage (again, one he or she was permitted to become familiar with); its content was about a predesignated unemotional topic.

Sample 2: Low-Stress Lie—The low-stress deceptive utterances were created in a similar fashion except false statements were spoken.

Sample 3: High-Stress Lie—Samples of this type consisted of untruths produced under high jeopardy. As stated, all subjects selected were known to hold very strong personal views about some issue (included were such topics as gun control, sexual orientation, religious faith, etc.). They were required to utter statements that sharply contradicted these strong views and do so while under the impression that their peers would hear (and see) their performance.

Sample 4: High-Stress Truth—This "high-stress only" procedure consisted of subjects reading truthful material, namely statements with which they agreed, but about which they were not particularly passionate. Here, they were conditioned to respond to the highest level of electric shock that they could tolerate and were then told that they would receive shocks whenever they produced the passage. The equipment employed was the electro-stimulus conditioning unit (STM100C, BIOPAC Systems, Inc.) associated with the BIOPAC MP-150. After conditioning, electric shock was administered during the initial, and any subsequent, run where the subject failed to show highly significant signs of stress.

Sample 5: High-Stress Lie, Dual Stressor—This experimental condition combined procedures 3 and 4. Specifically, the sample consisted of harsh lies produced under high jeopardy (as in sample 3), and with the threat and/or presence of electric shock added (as in sample 4). This sample resulted in lies spoken under the highest degree of psychological stress possible.

Sample 6: Simulated Stress—These low physiological stress samples were obtained after the subject was coached to produce a truthful passage in a manner reflecting how he/she *might* speak under conditions of significant stress.

Procedure

The actual procedure followed an order of presentation, which grouped the samples that involved stress together (e.g., samples 3, 5, and 4 in that order), and then, following a break, presentation of those that did not involve stress (i.e., baseline plus samples 1, 2, and 6). Specifically an experimental "run" was as follows:

1. After providing an informed consent, volunteers were assigned coded numbers (to insure anonymity) and then completed a background questionnaire to document their linguistic background, any history of any speech or hearing disorders, and their general health. This background information was collected to aid in any *post hoc* analyses to account for unexpected individual variation in response to the procedures.
2. The project's psychiatrist screened them, covering topics such as: (i) history of psychiatric disorders, (ii) history of heart conditions, (iii) other physical disorders, (iv) current medication regimen, (v) drug/alcohol use, and so on. None of the subject's responses to these questions were recorded. The psychiatrist also attempted to add an element of uncertainty to the interview to heighten arousal.
3. Those subjects who qualified were seated in the testing room and had a head-mounted microphone fitted to them; a second microphone was placed on the table. The GSR and PR transducers were then taped to two fingers of the right hand (later the electroshock stimulator was placed on the subject's other arm, but only for procedures 4 and 5). The physiological measures (GSR, PR) were then initiated and continued for the entire session.
4. Stress Trials: First, two or more runs were carried out with the subject producing the three high stress/deception passages (samples 3–5).
5. After the completion of the stressful procedure runs, subjects were debriefed as to the actual purpose of the study (they were

told also that they would *not* be heard by peers) and they were set at ease for the subsequent low stress procedures. After the break, the subjects provided multiple readings of the passage and samples 1, 2, 6. This pattern was repeated until only utterances at very low stress levels were obtained.

The use of the protocols described above enabled the development of a database of speech samples in which stress levels were documented through physiological and psychological measures. Of the 78 human subjects processed, 48 were able to complete *both* the protocol described above and meet all criteria for final inclusion. These additional criteria focused on the *shift in stress* as measured by the physiological correlates and the psychological scales. All of these measures of stress were examined (first) independently, and then in combination, to determine whether or not each showed a significant shift from the unstressed conditions to the stressed or deceptive conditions.

The four-way combined stress shifts were used to select the 48 subjects who ultimately provided the speech samples for LVA testing. Specifically, the overall stress shifts were computed by averaging the four cited measures after they had been converted to a common scale and weighted equally. Given this metric, only those subjects were included whose mean stress level, when lying or stressed, was actually more than double their baseline stress level. Specifically, the mean *overall* stress shift for all speakers was 141% with a mean rise of 129% for male speakers and 152% for females.

The speech materials cited were organized into 10 sets of 30 samples each (five male and five female) with a total of 56 speakers employed across all 10 (i.e., the 48 recorded under the protocol plus eight speakers recorded as low-stress foils). The first eight of these sets (four each for males, females) contained different speakers. The fifth set for each group was developed for reliability evaluations with subjects drawn from the other four.

Evaluators

Two teams of examiners assessed the LVA equipment. The first was a team of two evaluators (i.e., the second and fourth authors of this report) from the University of Florida who were certified as competent to conduct LVA analyses by V, LLC (Nemesysco's North American distributor). The second evaluation team consisted of two highly experienced instructors, chosen by the manufacturer, who traveled to the University of Florida to participate in the study. Both teams (Institute for Advanced Study of the Communication Processes [IASCP] and V) classified all samples as (i) either deceptive or nondeceptive and (ii) either stressed or nonstressed.

Evaluation Task

The LVA system requires a minimum of sentence-length speech materials for testing plus a "balanced" portion of an individual's normal speech for calibration purposes. Thus, to prepare the 300 speech samples for assessment of LVA, all of the test samples had to be individually paired with a calibration passage—in this instance a section of the "Rainbow Passage." The 300 pairs were then inputted as single digital audio files following the manufacturer's instructions. After the database was transferred into LVA, all of the sample statements (i.e., the speech material other than the "calibration" Rainbow Passage) were marked as "Relevant." It should be noted that coding speech material as "Relevant" is a necessary step in the operation of LVA. Only the analysis of the "Relevant" speech materials is summarized in this report. Finally, the complete set of digital audio files for LVA were assigned random filenames (using an alphanumeric code) to insure that no stress or deception information about the sample was available to any LVA operator.

The LVA analysis itself was conducted differently by the two teams of evaluators. The IASCP team at the University of Florida developed a protocol that did not require judgments by human operators. This protocol was based on the training received by the two members of the team who were certified to use the device. The protocol varied depending on whether or not LVA was being operated to detect deception or stress. For truthful and deceptive samples, the "Final Analysis" in the "Show Report" menu in the Offline mode was examined. If the Final Analysis stated that "Deception was indicated in the relevant questions" for any appropriate segment, the target sentence (i.e., the relevant material) was coded as "deceptive." For examining LVA's ability to detect stress, its "JQ" parameter was used. This parameter is defined by the manufacturer as one that measures emotional stress (not "physical" stress). In fact, of all the parameters representing emotional or cognitive states, JQ appeared to be most appropriate. Following the threshold described in the software manual, a sample was coded as "stressed" if the mean JQ level across all relevant segments (weighted for the duration of each segment) was 35 or greater; otherwise the sample was coded as "unstressed." The JQ threshold of 35 represents a criterion provided by the manufacturer, who did not provide any greater detail in terms of what JQ range technically represents. A screen snapshot of the LVA analysis window that includes the JQ parameter appears in Fig. 1. Given this approach, no interpretation of waveforms or waveform processing was necessary; rather the analysis was conducted automatically without any potential operator "bias" or effects.

The manufacturer's team did not follow the same protocol as those developed by the IASCP team. Rather, they conducted their own LVA test while at the University of Florida site. Further, they did not use a consistent protocol with all samples and, therefore, no attempt to document their operation of the device can be made. However, these operators were both highly experienced examiners specifically selected by the manufacturer. Thus, it is reasonable to assume that this team's use of the device was within the manufacturer's guidelines.

Results

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 approaches, four rates 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 proportion or percentage of the time that deception or high stress is said to be present when in fact 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 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 or stress. 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, regardless of the actual presence or absence of that behavior being tested. An accurate device would show true positive rates that are

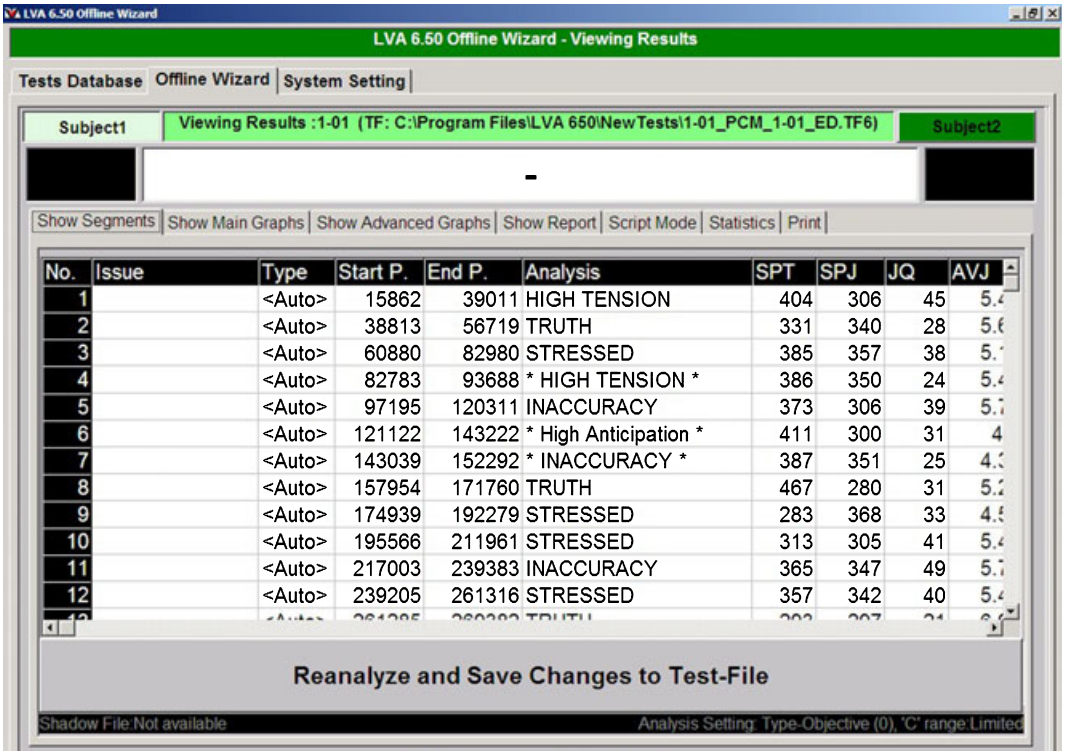

FIG. 1—*Screen snapshot of layered voice analysis (LVA) in the offline mode with the ''Show Segments'' window open. The column of JQ values for corresponding ''segments,'' or portions of the audio recording, appears second to last in the display window.*

both high and significantly different from the false positive ones. On the other hand, a device that performs at chance would show relatively equal true and false positive rates.

As stated, the detection of the presence of stress and deception was first performed by the two operators that made up the IASCP team. Their judgments were collated by a technician for subsequent presentation and statistical analysis. In turn, these judgments were processed by comparing them to the relevant stimuli (deception with and without jeopardy, high and low stress).

The results of both teams appear in Tables 1 and 2. Here, comparisons of seven different subsets of speech materials were obtained; they were:

- Analysis 1: All stressed versus unstressed materials.
- Analysis 2: All nondeceptive versus deceptive materials.
- Analysis 3: Stressed versus unstressed materials with deception absent.
- Analysis 4: Stressed versus unstressed materials when deception was present.

TABLE 1—*The percentage of samples coded as ''stressed'' or ''deceptive'' by layered voice analysis (LVA) employing the analysis developed by the IASCP team. The rates that correspond to accurate performance are ''true positive'' and ''true negative.'' The rates that correspond to inaccurate performance are ''false positive'' and ''false negative.''*

	Accurate		Inaccurate	
Analysis	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to stress (all conditions)	48	39	61	52
2. Sensitivity to deception (all conditions)	47	50	50	53
3. Sensitivity to stress (deception absent)	46	40	60	54
4. Sensitivity to stress (deception present)	50	37	63	50
5. Sensitivity to deception (low stress)	42	46	54	58
6. Sensitivity to deception (high stress)	46	50	50	54
7. Extreme groups (high-stress lie vs. low-stress truth)	50	40	60	50

TABLE 2—*The percentage of samples coded as ''stressed'' or ''deceptive'' by layered voice analysis (LVA) with the VSA database, as operated by the V team. The rates that correspond to accurate performance are ''true positive'' and ''true negative.'' The rates that correspond to inaccurate performance are ''false positive'' and ''false negative.''*

	Accurate		Inaccurate	
Analysis	True Positive	True Negative	False Positive	False Negative
1. Sensitivity to stress (all conditions)	56	41	59	44
2. Sensitivity to deception (all conditions)	47	45	55	53
3. Sensitivity to stress (deception absent)	56	35	65	44
4. Sensitivity to stress (deception present)	56	36	64	44
5. Sensitivity to deception (low stress)	43	41	59	57
6. Sensitivity to deception (high stress)	52	54	46	48
7. Extreme groups (high-stress lie vs. low-stress truth)	52	60	40	48

- Analysis 5: Nondeceptive versus deceptive materials when stress was low.
- Analysis 6: Nondeceptive versus deceptive materials when stress was high.
- Analysis 7: By an extreme groups design, in which only high-stress deceptive materials and low-stress nondeceptive statements were examined.

Table 1 provides these analyses for data generated by the IASCP team. For all seven measures, the true positive rates were below or near-chance (=50%), ranging from 42% to 50%. Moreover, an examination of the false positive rates shows that they were highly similar to the true positive rates (actually, they were slightly higher), ranging from 54% to 63%. Highly comparable hit and false positive rates indicate a lack of sensitivity. Thus, all seven subsets of data showed the same pattern of true and false positive rates seen in the general analysis.

Two other types of analyses were conducted also. They included: (i) the conversion of the seven true positive and false positive rates (found in Table 1) to d′ (d-prime), a metric of true sensitivity and (ii) repeated-measures ANOVAs of the proportion of stress/deception responses for each type of sample. Repeated-measures ANOVAs are commonly used in studies of this type; however, they often are conducted *only* on the true positive rates. Yet, the problem of detecting the presence of deception or stress in speech is an example of the larger problem of stimulus or signal detection. To illustrate, a device of this type might classify 90% or more of *all* samples as "deceptive," however, this value could be due either to (i) system accuracy or (ii) the device or human operator being biased to judge *any* sample as deceptive. In such a scenario, almost every utterance that actually involved deception would be correctly identified (a 90% true positive rate) and, at first glance, such results would appear to demonstrate that the approach worked well. However, if the detector was biased to classify *all* speech samples as deceptive, most truthful utterances also would be inaccurately classified as "deceptive."

To test for this relationship, d′ was applied; it is a procedure commonly used in analyzing data of this type (60). The d′ values can range from 0 to 4+, with 0 referring to no sensitivity at all and 4 (and upwards) corresponding to very high sensitivity. The conversion of values found in Table 1 to d′ is shown in Fig. 2 (along with corresponding data from the V team). For a device to be sensitive to a factor's presence (deception and high stress in this case), a d′ value of at least 1 should be attained. Indeed, even that value (i.e., 1) corresponds to only *minimal* sensititvity. Values that approximate zero indicate that the system is not sensitive to the behavior being studied—in this case stress/deception. Across all seven analyses, d′ was quite low, ranging from −0.35 to −0.08. These values were well below the threshold for even a limited degree of sensitivity to deception or stress, let alone the threshold for being characterized as "accurate."

Two separate repeated-measures ANOVAs also were conducted for evaluating LVA's performance with the VSA database from the basic study: One (the stress analysis) used the raw JQ values and the other (the deception analysis) used the "Deception Indicated" (DI) counts from the "Final Analysis" in the "Show Report" menu in the Offline mode. In the stress analysis, the unstressed and the stressed sample means were virtually identical (mean JQ = 36 and 34, respectively) and nonsignificant in difference ($F[1, 95] = 2.98$, $p = 0.09$). For the truthful versus deceptive speech samples, the DI rates were also not significantly different ($F[1, 95] = 1.40$, $p = 0.24$).

The corresponding responses from the V team are provided by Table 2. These are *not* values averaged for the two operators. Instead, as per their request, they were permitted to consult together and offer a single judgment for each speech sample. Yet, when the V team's results are examined, their true positive rates were quite similar to those seen for the IASCP team. Further, all were close to chance. False positive rates were also quite high and exceeded the true positive rates in all but two analyses ("Sensitivity to Deception" and "Extreme Groups"). The conversion of these raw values to d′ scores, shown in Fig. 2, reveals the device's insensitivity to stress and deception, with values hovering near zero (−0.40 to 0.30). Two repeated-measures ANOVAs were also conducted, separately for the stress and the deceptive materials. Neither factor was significant (Stress: $F[1, 94] = 1.79$, $p = 0.18$; Deception: $F[1, 94] = 0.49$, $p = 0.49$).

Discussion and Conclusions

The goal of this study was to evaluate the LVA using a database of truthful and deceptive speech produced in the presence/absence of a measurable degree of psychological stress. The LVA is a

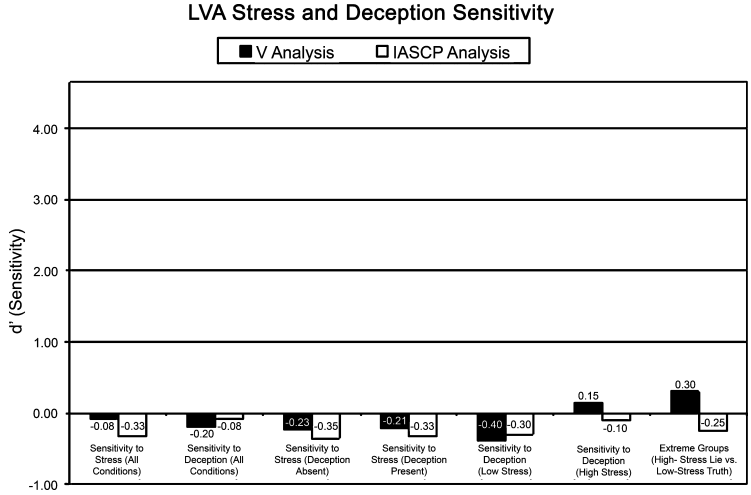

FIG. 2—*Sensitivity (d′) measures for the IASCP and V team's analysis of the layered voice analysis (LVA). Seven different analyses are shown within this figure and are coded by color (V team results in black; IASCP team results in white). Minimal acceptable sensitivity was set at 1.*

device designed to automatically detect both stress and deception in speech, among many other psychological states. This instrument was tested in a manner that controlled for "operator" variables, operator here referring to a human user who must make evaluative judgments about the device's output. The analysis of the device's performance was made over the entire dataset for stress and deception separately, as well as specific subsets of the data, to permit a careful report of its abilities.

Overall, the LVA did not display the expected sensitivity to the presence of deception, truth, and/or high/low stress in the speech samples that constitute the experimental database for this project. The observed true positive and false positive rates varied by the particular team and by the particular analysis conducted. However, sensitivity, measured by d′, remained only slightly above or below zero across all of these conditions. The conversion of the raw proportions to d′ was critical in observing the performance of this device. That is, d′ essentially specified the LVA system's capacity to detect stress/deception by taking into account its tendency to also classify truthful and/or unstressed samples as deceptive and/or stressed (i.e., its false positive rates). This general observation of the device's insensitivity held true for not only the general measures, such as sensitivity to stress or sensitivity to deception in all conditions but also for the analyses of subsets of the data. These subsets included stressed versus unstressed materials in the absence of deception, stressed versus unstressed materials in the presence of deception, truthful versus deceptive materials in the absence of stress, and truthful versus deceptive materials in the presence of stress. In addition, the "extreme groups" analysis compared materials that were most distinct from one another: deceptive materials produced under high stress versus truthful materials produced under low stress. In none of these five subsets did the LVA display any sensitivity to either stress or deception.

The raw data and all statistical analyses suggest only chance-level performance by the LVA, which can lead to the conclusion that the LVA is insensitive to deception and stress outside the laboratory. There are some common alternate interpretations of the results that could be used to argue that the device was not adequately tested in a laboratory study. For instance, the negative results could reflect limitations in the protocols used in the development of the speech database. Essentially, it may be argued that the stress shifts documented for the speech samples were not of a comparable magnitude to those induced in situations outside of the laboratory—e.g., those such as interrogations of individuals by police officers or military interrogators. In such cases, the "real-world" levels of stress might be higher than the psychological stress that can be generated in a laboratory setting on a college campus.

This interpretation might be a difficult one to rule out except (i) subjects' stress levels were measured to be demonstrably high and (ii) if only true positive rates were assessed. However, an assessment of LVA's performance on *truthful and unstressed speech samples* served as a robust control, one that permitted the examination of the device's potential bias to flag speech samples as deceptive in either the presence or absence of deception. If the database, collected under highly controlled conditions, contained inadequate levels of "real-world" stress, then very low false positive rates (near zero) would have been observed. In other words, if measurable stress or deception were not present in these samples, LVA should not have detected stress or deception in any portion of them. In fact, high false positive rates were the norm across all sets of speech materials and across both teams of operators: roughly half of the unstressed and truthful samples were classified by LVA as exhibiting stress and deception, respectively. A device that is, in fact, sensitive to these states should not falsely detect them if the procedures employed actually failed to elicit them.

The argument that laboratory speech samples lack any form of ecologically valid stress and/or deception is even less plausible when the specifics of the LVA system are considered. Its manufacturer claims that the device detects a wide variety of cognitive and emotional states. To do so, it must not only be sensitive to the relationship of the acoustic cues within the speech signal to the behaviors under study here, it also must exclude all other candidate cognitive states (e.g., stress due to past traumatic experiences, fatigue, degree of concentration, sexual arousal, intoxication, imagination level, to name just a few). For the LVA device to discriminate among such a large set of cognitive states, it must be highly sensitive to whatever acoustic attributes of the speech signal reflect those states. Presumed sensitivity at these levels suggests that LVA should be able to perform very well with our laboratory samples as they contain both deception with severe jeopardy and documented levels of significant stress. Actually, LVA's false positive rates were found to be consistently higher than their corresponding true positive rates. Thus, when both of these rates were converted to a single d′, no actual sensitivity to stress and deception could be observed.

Why, then, are the results of this double-blind study so at odds with the manufacturer's claims concerning the efficacy of LVA? As with the polygraph, it may be possible that some of the field success reported for the LVA system (i.e., by law enforcement and intelligence personnel) may actually be because of the skill of the interrogator, rather than the validity of system output—that is that he/she may pick up cues directly from the on-going interviews rather than from LVA output. As stated, it is difficult to understand the basis for the manufacturer's claimed successes by any other means. Undoubtedly, research investigating the relationship between the operators and this equipment should be carried out. At this juncture, it can only be said that the device itself does not appear capable of independently discriminating among utterances that are truthful and untruthful or stressed and unstressed.

It also should be noted that the focus of this research has been on deception and high stress states. Less emphasis has been placed on determining when truthful statements are being made. This slight should be corrected in the future as, in many cases, it is just as important to discover if the speaker is telling the truth as it is to determine when he/she is lying. This issue can be of particular relevance to intelligence and counterintelligence operations.

Finally, the results also highlight the longstanding need for a greater understanding of the basic relationship between psychological stress, deception, and the articulation of speech (with its corresponding acoustic consequences). Such a basic research program on the acoustics and perception of deception and stress should also encompass speech materials from multiple speakers, a variety of noise environments, and perhaps most importantly, multiple languages. It is well known that languages differ from one another in myriad ways, including their inventory of specific speech sounds, the rhythm or tempo of their conversational speech, as well as their pitch patterns, either in the form of tone languages (such as Mandarin) or in intonation (e.g., pitch patterns that change over the course of an entire utterance). To the extent that deception detectors rely on such patterns in speech, they may not be designed to cope with such cross-language variability. For example, changes in vocal pitch that differentiate words in a language such as Mandarin might be misclassified as "excessive pitch variability" that is indicative of psychological stress or the intent to deceive. Such a detector would falsely classify Mandarin speech as stressed or deceptive regardless of the speaker's actual intent. To date, no research has been

conducted on the validity of any models of voice-based stress and deception for speakers of other languages. Such research should provide both robust information about the way to detect deception in the field and data, which could provide manufacturers with additional information on which to design more effective devices. Perhaps of yet still greater importance, data of this type would provide methods which, when combined with other types of behavioral assessments, could be potentially effective in the development of multiple-factor systems designed to reliably detect/identify the cited (and related) behaviors—especially those where no invasive equipment is involved.

Acknowledgments

We wish to thank Rachel Kesselman, David Kahan, and Dr. Andrew Morgan for their invaluable assistance with the project. We also wish to thank John Taylor, Tom Winscher, and Cindy Worden of ''V'' for their kind assistance in the LVA evaluations.

References

1. Chen Y. Cepstral domain talker stress compensation for robust speech recognition. IEEE Trans Acoust 1988;36:433–9.
2. Cummings K, Clements M. Analysis of glottal waveforms across stress styles. Proceedings of the IEEE, International Conference on Acoustics, Speech, and Signal Processing. New York, NY: Institute of Electrical and Electronics Engineers, 1990;1:369–72.
3. Cummings K, Clements M. Analysis of glottal excitation of emotionally styled and stressed speech. J Acoust Soc Am 1995;98:88–98.
4. Frick RW. The prosodic expression of anger: differentiating threat and frustration. Aggress Behav 1986;12:121–8.
5. Hicks JW, Hollien H. The reflection of stress in voice-1: understanding the basic correlates. Proceedings of the 1981 Carnahan Conference on Crime Countermeasures. Lexington, KY: ORES Publications, 1981;189–94.
6. Hollien H. Vocal indicators of psychological stress. In: Wright F, Bahn C, Rieber RW, editors. Forensic psychology and psychiatry. New York: NY: Academy of Sciences, 1980;47–72.
7. Hollien H. Acoustics of crime. New York, NY: Plenum Press, 1990.
8. Hollien H, Saletto JA, Miller SK. Psychological stress in voice: new approach. Studia Phonet Posnan 1993;4:5–17.
9. Lazarus RS. From psychological stress to the emotions—a history of changing outlooks. Ann Rev Psychol 1993;44:1–21.
10. Scherer KR. Vocal indicators of stress. In: Darby J, editor. Speech evaluation in psychiatry. New York, NY: Grune and Stratton, 1981;171–87.
11. Scherer KR. Voice, stress and emotion. In: Appley H, Trumbull R, editors. Dynamics of stress: physiological, psychological, and social perspectives. New York, NY: Plenum Press, 1986;157–79.
12. Siegman AW, Boyle S. 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 1993;102:430–7.
13. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. J Acoust Soc Am 1972;2:1238–50.
14. Williams CE, Stevens KN. Vocal correlates of emotional states. In: Darby J, editor. Speech evaluation in psychiatry. New York, NY: Grune and Stratton, 1981;221–40.
15. Anolli L, Ciceri R. The voice of deception: vocal strategies of naïve and able liars. J Nonverbal Behav 1997;21:259–84.
16. Brenner M, Shipp T, Doherty ET, Morrisey P. Voice measures of psychological stress—laboratory field data. In: Titze IR, Scherer RC, editor. Vocal fold physiology. Denver, CO: Denver Center for the Performing Arts, 1983;239–48.
17. DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. Psychol Bull 2003;29:74–118.
18. Doherty ET. Speech analysis techniques for detecting stress. Proceedings of the Human Factors Society. Santa Monica, CA: Human Factors & Ergonomics Society, 1991;1:689–93.
19. Hicks JW. An acoustical/temporal analysis of emotional stress on speech. Doctoral dissertation. Gainesville, FL: University of Florida, 1979.
20. Hollien H. Acoustical analysis of psychological stress. In: Lawrence V, editor. Tenth symposium care of the professional voice. New York, NY: The Voice Foundation, 1981;145–58.
21. Kuroda I, Fujiware O, Okamur N, Utsukli N. Method for determining pilot stress through analysis of voice communication. Aviat Space Environ Med 1976;47:528–33.
22. McGlone RE, Hollien H. Partial analysis of the acoustic signal of stressed and unstressed speech. Proceedings of the 1976 Carnahan Conference on Crime Countermeasures. Lexington, KY: ORES Publications, 1976:83–6.
23. McGlone RE, Petrie C, Frye J. Acoustic analysis of low-risk lies. J Acoust Soc Am 1974;55:S20(A).
24. Nilsonne A, Sundberg J. Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples. Music Percept 1985;2:507–16.
25. Rosenfield JP, Soskins M, Bosh G, Ryan A. Simple, effective countermeasures to P300-based tests of the detection of concealed information. Psychophysiology 2004;41:205–19.
26. Scherer KR, Banse R, Walbott HG, Goldbeck T. Vocal cues in emotion encoding and decoding. Motiv Emot 1991;15:123–48.
27. Streeter LA, Krauss FM, Geller V, Olson C, Apple W. Pitch changes during attempted deception. J Pers Soc Psychol 1977;35:345–50.
28. Tolkmitt FJ, Scherer KR. Effects of experimentally induced stress on vocal patterns. J Exp Psychol 1986;12:302–13.
29. Vrij A, Edward K, Roberts KP, Bull R. Detecting deceit via analysis of verbal and nonverbal behavior. J Nonverbal Behav 2000;24:239–63.
30. Vrij A, Edward K, Roberts KP, Bull R. Stereotypical verbal and nonverbal responses while deceiving others. Pers Soc Psychol Bull 2001; 27:899–909.
31. Cestaro VL. 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, 1996. Report no.: DoDPI95-R-0004.
32. Haddad D, Walter S, Ratley R, Smith M. Investigation and evaluation of voice stress analysis technology. Rome AFB: US Dept Justice Report, 2002. Grant 98-LB-VX-A103.
33. Heisse JW. Audio stress analysis—a validation and reliability study of the psychological stress evaluator (PSE). Proceedings of the 1976 Carnahan Conference on Crime Countermeasures. Lexington, KY: ORES Publications, 1976;5–18.
34. Hollien H, Geison LL, Hicks JW Jr. Voice stress evaluators and lie detection. J Forensic Sci 1987;32:405–18.
35. Horvath F. The effects of differential motivation on detection of deception with the psychological stress evaluator and the galvanic skin test response. Appl Psychol 1979;64:323–30.
36. Kubis J. Comparison of voice analysis and polygraph as lie detection procedures. Aberdeen Proving Ground, MD: US Army Land Warfare Laboratory, 1973. Technical Report LWL-CR-U3B70.
37. Meyerhoff JL, Saviolakis GA, Koenig ML, Yourick DL. Physiological and biochemical measures of stress compared to voice stress analysis using the Computer Voice Stress Analyzer (CVSA). Fort McClellan, AL: Department of Defense Polygraph Institute, 2001. Report No. DoDPI01-R-0001.
38. Brenner M, Branscomb HH. The psychological stress evaluator, technical limitations affecting lie detection. Polygraph 1979;8:127–32.
39. Brockway BF, Plummer OB, Lowe BM. The effects of two types of nursing reassurance upon patient vocal stress levels as measured by a new tool, the PSE. Nurs Res 1976;25:440–6.
40. McGlone RE. Tests of the psychological stress evaluator (PSE) as a lie and stress detector. Proceedings of the 1975 Carnahan Conference on Crime Countermeasures. Lexington, KY: ORES Publications, 1975;83–6.
41. VanderCar DH, Greaner J, Hibler N, Speelberger CD, Bloch S. A description and analysis of the operation and validity of the psychological stress evaluator. J Forensic Sci 1980;25:174–88.
42. Cestaro VL, Dollins AB. An analysis of voice responses for the detection of deception. Ft. McClellan, AL: U.S. Department of Defense Polygraph Institute, 1994. Report No. DoDPI94-R-0001.
43. Janniro MJ, Cestaro VL. Effectiveness of detection of deception examinations using the computer voice stress analyzer. Ft. McClellan, AL: U.S. Department of Defense Polygraph Institute, 1996. Report No. DoDPI96-R-0005.
44. National Research Council. The polygraph and lie detection. Washington, DC: The National Academies Press, 2003.

45. Brenner M, Branscomb HH, Schwartz GE. Psychological stress evaluator—two tests of a vocal measure. Psychophysiology 1979;16: 351–7.
46. Lynch BE, Henry DR. A validity study of the psychological stress evaluator. Can J Behav Sci 1979;11:89–94.
47. O'Hair D, Cody MJ, Behnke RR. Communication apprehension and vocal stress as indices of deception. West J Speech Commun 1985;49:286–300.
48. Greaner J. Validation of the PSE. Thesis. Tallahassee (FL): Florida State Univ, 1976.
49. Inbar GF, Eden G. Psychological stress evaluators: EMG correlations with voice tremor. Biol Cybern 1976;24:165–7.
50. Leith WR, Timmons JL, Sugarman MD. The use of the psychological stress evaluator with stutterers. Fluency Dis 1983;8:207–13.
51. Horvath F. An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. Appl Psychol 1978;63:338–44.
52. Horvath F. Detecting deception: the promise and the reality of voice stress analysis. J Forensic Sci 1982;27:340–51.
53. Shipp T, Izdebski K. Current evidence for the existence of laryngeal macrotremor and microtremor. J Forensic Sci 1981;26:501–5.
54. Barland G. Detection of deception in criminal suspects. Dissertation. Salt Lake City, UT: University of Utah, 1975.
55. Lykken D. Tremor in the blood. New York, NY: McGraw-Hill, 1981.
56. Hollien H, Harnsberger JD. Voice stress analyzer instrumentation evaluation, final report. Gainesville, FL: CIFA, 2006. Contract: FA-4814-04-0011.
57. Hollien H, Harnsberger JD, Martin CA, Hollien KA. Evaluation of the CVSA voice stress analyzer. J Forensic Sci 2008;53:183–93.
58. Hollien H, Harnsberger JD. The use of voice in security evaluations. J Cred Asses Witness Psych 2006;7:74–8.
59. Maier W, Buller R, Phillip M, Heuser J. The Hamilton anxiety scale: reliability, validity and sensitivity to changing anxiety and depressive disorders. Defect Dis 1988;14:61–8.
60. Macmillian NA, Creelman CD. Detection theory: a user's guide. 2nd ed. Lawrence, NJ: Erlbaum Associates, 2005.

Additional information and reprint requests:
James D. Harnsberger, Ph.D.
Institute for Advanced Study of the Communication Processes
68 Dauer Hall
University of Florida
Gainesville, FL 32611
E-mail: jharns@csd.ufl.edu