J Forensic Sci, January 2008, Vol. 53, No. 1
doi: 10.1111/j.1556-4029.2007.00596.x
Available online at: www.blackwell-synergy.com

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

Evaluation of the NITV CVSA

ABSTRACT: The purpose of this study was to evaluate a commonly used voice stress analyzer, the National Institute of Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA), using a speech database containing materials recorded (i) in the laboratory, while highly controlled deceptive and shock-induced stress levels were systematically varied, and (ii) during a field procedure. Subjects were 24 each males/females (age range 18–63 years) drawn from a representative population. All held strong views on an issue and were required to make sharply derogatory statements about it. The CVSA system was then evaluated in a double-blind study using three sets of examiners: (i) two UF scientists trained/certified by NITV in CVSA operation, (ii) three experienced NITV operators provided by the manufacturer and (iii) five experimental phoneticians. The results showed that the "true positive" (or hit) rates for all examiners ranged from chance to somewhat higher levels (*c.* 50–65%) for all conditions and types of materials (e.g., stress vs. unstressed, truth vs. deception). However, the false-positive rate was just as high – often higher. Sensitivity statistics demonstrated that the CVSA system operated at about chance level.

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

A Perspective

It is well known that the speech signal contains features which can be used to provide information about a human. Speaker identification is an example. It is an area based on sets of established speaker-specific phonatory properties (1–6). Another such area involves the detection of alcohol intoxication as it is reflected in voice and speech. Here too, substantial research is available to describe those relationships (7–12). Human emotion (including psychological stress) constitutes yet a third domain where relevant behaviors can be detected in voice (1,13–19).

Most of the mechanical and neurological bases for these relationships are reasonably well established. That is, as the speech act represents the output of a number of high level and integrated physiological systems, it appears appropriate to assume that the process may reflect any number of behaviors. That is, as the oral production of any language involves the use of multiple sensory modalities, high-level cognitive functioning, complex cortical processing and a large series of motor acts (20,21), it is legitimate to predict that even more subtle operations – such as the detection of deception and/or truth from speech and voice – might be possible.

Detecting Truth–Deception–Stress from Voice

There is no question but that a device which could be used to detect the presence of truth, deception and/or stress from voice and speech would be of great value to law enforcement and intelligence agencies; indeed, they hold a particular relevance to many forensic operations. Several currently existing systems are purported to meet these needs. As a group they usually are referred to as "voice stress analyzers" and they will be so identified in this paper even though their function may or may not be based solely on "voice stress". While none of the pertinent manufacturers provide any scientific evidence on the efficacy of their devices, most of them have been studied by others. However, the research that has been reported for several such systems has ranged from mixed to negative. First, while some authors suggest that these devices might detect stress – at least under certain circumstances (22–25) – most of the relevant data have not supported that position (26–29 and see below). On the other hand, it also can be said that, to date anyway, none of the instruments evaluated have been afforded a reasonably comprehensive assessment. Some investigators simply have not controlled their procedures at a level which would permit acceptable data to be generated (30–33). Others have either not assessed a sufficient number of variables or carried out research that was too limited in scope (34–36). The focus of yet others has been somewhat narrow (36–38), involved only small laboratory studies (22,24) or was too restricted, even if reasonably well controlled (39–42). Some researchers have limited their efforts only to field studies (43,44) and, no matter how competently the investigation was carried out, this approach alone does not provide enough relevant information. Only Nachshon and Feldman (45) attempted both laboratory and field studies. However, even here, their effort lacked breadth and sufficient control. In any event, an overall summary of the cited research suggests that none of the devices tested were capable of validly detecting psychological stress, truth and/or lying from cues embedded in motor speech.

A Model

As may be inferred, it is necessary to specify the types of experiments which should be carried out when addressing the challenge of properly evaluating deception detection equipment or spoken truthfulness in voice. For example, the most common approach to system evaluation has been to conduct "simulated field" studies. The reason for doing so appears to be the desire to determine if they will work under "real life" conditions. In addition, there are individuals who argue that a "field" approach is necessary, as

[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.

Received 11 Aug. 2006; and in revised form 25 Mar. 2007; accepted 8 April 2007.

laboratory-type experiments are simply "games" and, being "unrealistic", they can provide little-to-no useful information (46). On the other hand, it is well established that field research (i) ignores the need for basic system assessment under "controlled" conditions, (ii) cannot include many events necessary for the proper determination of system operation, (iii) does not exclude debilitating external variables and (iv) often discounts the speaker's actual emotional state. Thus, it appears that little can be gained from a "field study only" approach as, lacking reasonable controls, it is not ordinarily possible to determine if the information thus obtained is valid.

As would be expected, these differences of opinion create a very real dilemma. It is clear that laboratory level control is necessary in research of this type, but can such an approach be "realistic" enough to provide useful data? On the other hand, is it possible to conduct field experiments that are precise enough to generate valid data? Our response is in the affirmative to both questions and "a three-level model" has been developed in support of this position.

At its first level, this model would involve extensive and highly controlled "laboratory experiments". Here, utterances involving truthfulness, deception, psychological stress (and, perhaps other emotions) – at various levels of intensity – would be obtained from appropriate speakers. These behaviors would be experimentally induced, would be relevant and their presence would be verifiable by independent assessment. The model's second level would focus on both (i) "simulated field" and (ii) "actual field" research, but studies where only modest levels of control and verification are possible. The material would be drawn primarily from the Forensic milieu but also from relevant intelligence sources. In one of these two, a field scenario would be created where subjects would be involved in a stressful encounter or when they are lying in a stressful training situation. The other of the two (second level) approaches would include actual cases involving forensic-type interrogations (usually criminal). However, in neither of the cited instances could the speaker's actual stress levels (or guilt) be fully verified. The model's third level would involve "actual field" experiments (most of them of the forensic-type) – commonly referred to as "real life" studies – but those where the data were obtained under high levels of control and validation. In this instance, the responses would be obtained where an actual crime could be committed. An example would be where inmates of a prison are interrogated (and recorded) as to whether or not they had recently ingested certain illegal drugs. Their culpability would be measured by a standard blood test; their stress level by the physiological measures employed in the laboratory experiments.

All three approaches lead to the development of test vehicles at various levels of sophistication which could be used both in basic research or to permit evaluation of specific equipment. The analyses to follow are based on a large first-level (laboratory) experiment and a smaller one from level 2; however, only the evaluation of a single instrument (i.e., the CVSA) will be reported.

Specific Goals

As stated, this project's objectives were to generate highly controlled speech materials which could be validly used to 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 of the evaluation of National Institute of Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA), a device that purportedly detects lying by means of assessing the level of stress in spoken speech. As stated that instrument was tested in a large double-blind laboratory experiment supplemented by a smaller field study – neither of which permitted any on-scene operator to observe the human subjects and/or their responses. It is only through the use of such controlled approaches that the characteristics of the device itself can be evaluated in a thorough and impartial manner. Indeed, it was judged that, until this project had been carried out, the CVSA equipment had not been adequately assessed.

The research approaches highlighted an important problem in this area. While systems such as this one rely on the effects of psychological stress (as it is reflected in the acoustic properties of speech) as the only indicator of deception, that supposition may not be true. For example, numerous other behavioral states can result in stress-based changes in voice. Thus, it became clear that stress and deception had to be examined separately, as well as in combination, to properly model their relationship and permit evaluation of the targeted behaviors. Moreover, the detection of stress in voice is important in its own right to both intelligence and law enforcement operations.

Method

As stated, a series of experiments were carried out which were focused on utterances spoken truthfully, deceptively and with high-low stress. To do so, a fairly large and somewhat diverse subject population was required, as were procedures that would induce – and permit independent measurement of – the desired behaviors. Further, the individuals using the system had to be well versed in its operation. Nor could any of these evaluators have any independent knowledge of the conditions reflected in the test samples.

Subjects

Seventy-eight adult volunteers, both male and female, were screened for inclusion suitability. Their ages ranged from 18 to 63 years and they were drawn in a manner that attempted to sufficiently sample the demographics of the U.S. population. Selection also was based, to some extent anyway, on socioeconomic background with the volunteers ranging from policemen to students, from housewives to U.S. Marines, from workmen to clergy and so on. Further, volunteers had to hold very strong personal views about some subject (e.g., politics, religion, Iraq, etc.). They were systematically screened by the project's Investigator-psychiatrist who first excluded any with medical conditions or who showed a past history of psychological trauma. Subsequently, many other potential exclusionary mental and physical health criteria also were assessed and used in the selection process (see General Procedure section).

Recording Procedures

The volunteers selected were recorded in a quiet (but "live") room with two laboratory quality microphones (head-mounted Shure and Sony ECM-737) feeding (i) a Sony TCD-D8 DAT recorder, (ii) a digitizer (BIOPAC MP-150) coupled to a computer, and (iii) a Marantz PMD-221 cassette recorder; all equipment was calibrated. Additionally, digital audio–video recordings of each subject were made during all experimental runs. The video camera (a Sony DCR-HC2) was fixed and focused on the subject's upper body.

Measurement of Stress Levels

Mindful of the problems associated with the external determination of internal stress level (47,48), several procedures were applied

to asses these conditions. That is, five methods presumed appropriate for the measurement of psychological arousal and/or stress were administered at appropriate times; i.e., 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), (ii) a saliva (cortisol) test (also once per procedure) and (iii) continual body response evaluations consisting of galvanic skin response (GSR) and pulse rate (PR). The anxiety/stress tests consisted of a 10-pt "emotion felt" anxiety checklist and a modified version of the Hamilton test (49). The cortisol level (saliva) tests were accomplished by Salimetrics LLC, No. 5100 Cortisol Tests. GSR and PR were measured using the BIOPAC Systems, Model MP-150.

Speech Samples

While the speech samples were created for both basic research and system evaluation; their nature was especially critical to the present assessments. Seven different types of utterances were produced by each subject-speaker. That is, following a familiarization procedure, the first was elicited for baseline calibration and the six others for experimental purposes. As would be expected, they were carefully designed. First, they were extensive enough to provide a reasonable repertoire of speech with each passage consisting of five to seven content sentences (over 30 sec total). A 17- to 26-word "content neutral" phrase or sentence was embedded near the center of each. It was inserted there so it would be uttered at the same stress level as was the full passage yet (later) removed for analysis. Thus, it would contain no language cues (semantic or pragmantic) about the condition being experienced. An example of such an embedded phrase is: "This is a position I am very comfortable with because I have thought about it for a long while and it makes sense to me." Note that the text is not specific to any particular topic. The use of these "content neutral" phrases prevented the system operators from being exposed to language-based clues as to the type of speech being produced. The nature of the complete set of speech samples is described below; all were uttered three to five times with only that sample meeting all criteria used in the evaluations:

Baseline (calibration) Sample—All subjects read (several times) a standardized phonetically balanced (unstressed) truthful passage, namely the Rainbow Passage. Note, it did not contain a content-neutral sentence.

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. "I now live at 3120 Northwest 38th Drive, Camden, New Jersey" (plus related text) is an example.

Sample 3: High-Stress Lie—Samples of this type consisted of untruths produced under high jeopardy. As stated, all subjects had been selected from groups that 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 to do so while under the impression that their friends and/or other peers would hear (and see) their performance. In addition, subjects were instructed to produce these lies in a speaking style that strongly suggested that they actually believed them. As with all samples, these utterances were repeated until all experimental criteria were met.

Sample 4: High-Stress Truth—This was the high-stress only procedure; it consisted of subjects reading truthful material, namely statements with which they agreed but about which they were not particularly passionate. For this procedure, they were conditioned to respond to the highest level of electric shock that they could tolerate. They were told that they would receive shocks whenever they produced the passage. The equipment employed was the electrostimulus conditioning unit (STM100C) associated with the BIOPAC MP-150 (the among-subjects stimulus level ranged from 25 to 50 V). After conditioning, electric shock was administered during the initial run of the procedure and in any subsequent run wherein the subject failed to demonstrate highly significant signs of stress.

Sample 5: Very High-Stress Lie—This experimental condition combined procedures 3 and 4. Specifically, the sample consisted of harsh lies produced under high jeopardy (as in Sample 3), but with the threat and/or presence of electric shock added (as in Sample 4). Therefore, Sample 5 combined two stressors and was used to elicit lies under the highest degree of psychological stress permissible.

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. Each was allowed to repeat this simulated stress procedure until he/she and the experimenter agreed that utterances produced were different from their normal speech and presumably "reflected stress" even though it was not present.

Procedure

The "sample number" specified above does not reflect the actual order of presentation. That is, the final order employed, within a subject-trial, grouped the samples that involved stress together (e.g., Procedures 3, 5 and 4 in that order), and then, following a break, presentation of those that did not involve stress (i.e., Baseline plus Procedures 1, 2 and 6). To be specific, an experimental "run" was as follows:

1. After giving informed consent, volunteers were assigned coded numbers (to ensure anonymity); they then completed the "Subject Information Form".
2. The project's Psychiatrist then screened them using a series of questions concerning those aspects of their background that might make them unsuitable for the study. The screening questions covered a series of specific topics included were: (i) history of psychiatric disorders, (ii) history of psychological trauma, (iii) history of heart conditions, (iv) other physical disorders, (v) current medication regimen, (vi) drug use, (vii) alcohol use, (viii) native speaker of English, and so on. None of the subject's responses to these questions were recorded. Those subjects selected were paid $30.00 (at trial completion) for their participation. Incidently, the Psychiatrist also attempted to add an element of uncertainty to the interview to heighten arousal for the high-stress conditions.
3. Subjects were seated in the testing room and had the head-mounted microphone fitted to them; the second microphone was

placed on the table. The GSR and PR sensors were then taped to two fingers of the right hand (later the electro-shock stimulator was placed on the subject's other arm, but only for procedures 5 and 4). 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 a passage that was judged to be both (i) offensive to his/her strongly held beliefs and (ii) entirely untruthful (i.e., Sample 3). The saliva test for cortisol and the two self-report tests were administered at the end of this procedure. It was followed by those trials that elicited Samples 5 (double stressors) and 4 (electric shock only).
5. After the completion of the stressful procedure runs, subjects were debriefed as to the actual purpose of the study. The transducer for administering shock was removed, and they were engaged in conversation with the research personnel to set them at ease for the subsequent low-stress procedures. After the break, the subjects provided multiple readings of the calibration or baseline passage (i.e., the Rainbow Passage). At the end of this (calibration) procedure, the saliva test for cortisol and the two self-report tests were once again administered. This pattern was repeated for all runs and until only utterances at very low-stress levels were obtained.
6. Finally, the low-stress trials were run in order; Sample 1: unstressed (neutral) truthful utterances; Sample 2: the unstressed deceptive passage and Sample 6: elicitation of a sample of simulated stress produced under low-stress conditions.

The use of the protocols described above enabled the development of a practical database of validated speech samples; one that contained all of the linguistic information needed to test a variety of voice stress analysis equipment (such as the CVSA) – plus provide material for basic research.

Of the 78 human subjects processed, 48 completed both the protocol described above and met the criteria for final inclusion. These (subsequent) criteria focused on the "shift in stress" as measured by the physiological correlates and the self-report scales. All of these measures of stress (plus cortisol) were examined independently to determine whether or not each showed a significant shift from the unstressed conditions to the stressed conditions. Four of these metrics proved useful in the subject selection process. They were GSR, PR, the emotion checklist, and the modified Hamilton scale.

One measure (cortisol) failed to show significant shifts in an orderly pattern and was excluded from the composite measure. Like many other studies (50–56) which have shown mixed results for cortisol assessment, these averaged measures failed to show a significant, or even systematic, difference in the anticipated direction between the unstressed and stressed conditions. In short, it appears likely that cortisol level did not shift quickly enough to provide useful information for the rapidly changing experimental procedures employed.

The four-way combined stress shifts were used to select the 48 subjects who ultimately provided the speech samples for CVSA testing. That is, the overall stress shifts were computed by averaging the four cited measures after they had been converted to a common scale and weighed equally. Given this metric, only those subjects were included in the core database whose mean stress level, when lying, was actually more than double their baseline stress level. Stated differently, the mean "overall" stress shift observed across all speakers was 141% with a mean rise of 129% for male speakers (range 61–208%) and 152% for females (range 45–392%). The resulting database, then, consisted of 48 speakers, 24 males and 24 females, all of whom produced deceptive and high-stress statements while experiencing a significant degree of increased stress but who did not do so for the several low-stress conditions.

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. Please note again that the speech appearing in the database for CVSA evaluation was drawn from the "content neutral" sentence contained in one passage, each condition. It should be re-emphasized that this sentence powerfully reflected the stress level being experienced by the subject even though it did not linguistically reveal content.

The SERE Field Research Materials

A second database also was developed; this one involved materials of the "field research" type. Specifically, a cooperating federal intelligence agency provided the project with a set of audio–video recordings of military trainees answering questions while undergoing SERE training (Survival, Escape, Resistance, Evasion). SERE is a rigorous survival training program where the students are disciplined not to reveal any information when captured and interrogated by hostile forces.

The SERE trainees took part in a guilty knowledge study in which they were instructed to lie about several aspects of their training. The goal of the study was to detect lies embedded in a large number of truthful responses. In turn, subjects faced punishment if their untruths were detected. Thus, they were lying under a substantial degree of jeopardy, although they did not face a severe immediate threat (such as one which was life threatening).

While being recorded on video-camera, the SERE subjects wore a Vivometrics "Life Shirt" that continuously recorded common physiological correlates of stress; included were heart rate, breathing and blood pressure metrics. When lying, the SERE subjects exhibited heart rates typically varying between 140 and 170 bpm (with 95 being the lowest value recorded). In contrast, their resting heart rates were quite low; i.e., ranging between 48 and 52 bpm. Thus, it appeared reasonable to infer that the threat of punishment associated with this procedure resulted in a substantial elevation of stress levels during the interrogation.

The "SERE" database included a total of 56 utterances consisting of either a "yes" or a "no" response to a question. These 56 utterances were organized into related sets of eight speech samples, six sets for the males and two for females. Each set contained samples drawn from five subjects plus (truthful) foil utterances obtained from individuals, working at Institute for Advanced Study of the Communication Processes (IASCP), who were not otherwise involved with SERE. Each of these eight sets contained three lies and four truths; all were counterbalanced. The individual samples were coded for retrieval, to allow for randomization within each group and to ensure that no pertinent information (about a sample) would be inadvertently disclosed.

Evaluation Task

As stated, the present task was to evaluate the CVSA equipment for its ability to detect stress, truth and deception from the analysis of spoken speech. This device processes very short speech samples

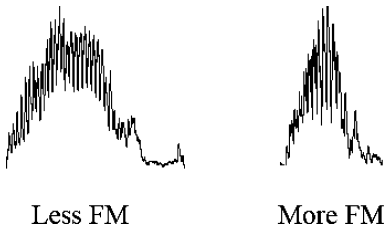

FIG. 1—*The left chart (Less FM) shows "blocking" presumably due to stress and/or deception (time frame = 861 ms). The right chart (More FM) shows an absence of blocking (time frame = 474 ms), a pattern which would be interpreted as unstressed and/or not deceptive.*

(ie., such as "yes", "no") and its output can be described as a 2-D chart displaying the duration of the speech signal on the horizontal axis; the information on the vertical axis is not defined (57). A sample pair of charts appear in Fig. 1. The left chart is said to display a voice recording in which psychological stress is present. Its gross shape would be referred to as "blocking", due to its general rectangular form. In turn, the right chart displays a voice recording in which psychological stress is presumed absent; this judgment is based on its triangular configuration. Thus, this chart would be classified as nondeceptive and unstressed. The NITV Manual (57) states that blocking – the single cue for stress that may be a product of deception – results from the suppression of a natural "microtremor" in the muscles that control the vocal folds and speech articulation. It is claimed that when this microtremor is suppressed, its acoustic byproduct – referred to as the "inaudible frequency modulation (FM) component" – is lost. In turn, this results in the appearance of "blocking" in the signal seen on a CVSA chart. When the subject is no longer under stress, the microtremor is said to return and blocking dissipates.

As cited above, the available speech was more extensive than that required to test the CVSA system. As its operators typically analyze single syllables, samples of this type were extracted from the content neutral material found in the Core database. However, care was taken to follow the manufacturer's instructions (57) so that the required procedure would be in no way compromised. In all cases, the samples selected were single-syllable utterances, which occurred at the maximum of both physiological measures (GSR and PR). To avoid introducing artificial errors, three other criteria were applied: (i) no sample could exhibit an unnaturally abrupt onset or offset ("artificial" blocking had to be avoided), (ii) the utterance had to sustain the intensity level (i.e., no trailing voice effect) and (iii) they had to be produced in the modal (normal) voice register (i.e., no breathy samples were acceptable nor were those in the falsetto or vocal fry registers). If a syllable at the physiological maximum did not meet all three criteria, the one nearest to that level which did was substituted. Finally, the 300 samples were randomized to ensure that no stress or deception information based on ordering would be available to any of the operators. As the SERE samples were single-word responses ("yes", "no") only randomization was required. Samples were then inputted the appropriate laptop computer using its sound card and as directed by the manufacturer (57). Once all samples were inputted, they were classified.

It may seem that inordinate care was taken in preparing the samples for use in these evaluations. However, this approach was employed to ensure that the manufacturer's protocols were met in an equitable manner and because certain speech waveforms can show different phoneme-based patterns – depending upon the particular vowels and consonants being spoken. Such problems, plus variation in intensity, can complicate the interpretation of CVSA output by artificially adding "blocking" where it does not exist. These problems were avoided by applying the cited safeguards.

Evaluators

The processing cited above was carried out by three teams of examiners. The first was a team of two evaluators from the University of Florida's IASCP who had attended NITV school and were certified as competent to conduct CVSA analyses (i.e., they are the second and fourth authors of this report). The second evaluation team consisted of three highly experienced operators (provided by NITV) who traveled to the University of Florida to participate in the study. Both teams (IASCP and NITV) classified all samples as deceptive/stressed or nondeceptive/nonstressed.

A third group of evaluators consisted of five highly competent Experimental Phoneticians, who were experienced in visually "reading" acoustic signals (such as waveforms). They were only asked to judge (forced-choice) if the samples displayed "blocking" or "nonblocking". In essence they were included as a type of "control" group.

The NITV team represented the one with the greatest experience with CVSA, indeed, they could be considered "experts". Of course the IASCP team had been trained by the manufacturer and received certification; nevertheless, they were less experienced with the device. Finally, while the Phonetician team were advanced specialists in the examination of speech waves for cues relative to the presence of information within the signal (i.e., words, phonemes, stress, gender, age, training, identity, etc.), they possessed no specific experience with devices of this type and had received only minimal training in the form of instructions. Nevertheless, their skills permitted a "continuum" of operator expertise to be developed.

Results

The resulting data were examined by means of a number of techniques designed to explore the possibility that the CVSA 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, truth as truth, 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 a device's ability to correctly discern deception or stress. An examination of the true-positive rate alone does not permit specificity of system accuracy as a high true-positive rate can be the product of either its actual accuracy or simply its bias to classify the stimulus as positive regardless of the actual presence or absence of the target behavior. An accurate device would show true-positive rates that are 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.

Finally, the false-negative and true-negative rates also were determined (they are known as the "miss rate" and "correct rejection rate", respectively, in Signal Detection Theory). False-negatives occur when the signal is present but the device classifies it as

absent. True-negatives are cases in which the behavior is in fact absent and it is thus accurately judged (e.g., truthful speech samples that are classified by a device as truthful, unstressed samples that are classified as absent of stress). Incidentally, this procedure also can be applied if the task is to determine if a speaker is making truthful statements. In that instance, the process is "inverted" with the rate of true utterances becoming the true positive and the lies becoming the true negative. For simplicity, this report will focus only on high stress and deception.

Results for the Core Database

The detection of the presence of "blocking" (or nonblocking) on the 300 CVSA charts was first performed by the two operators who made up the IASCP team, then by the NITV team, and lastly by the Phoneticians. They did so separately and without knowledge of either the specific sample selection or the judgments made by other operators. Their decisions were given to an independent investigator (i.e., the PI) who had a technician collate them by team for the statistical analyses. In short, the experiment was double-blind in nature. Further, it permitted a large number of analyses to be carried out; a full accounting may be found in the "Final Report" submitted to CIFA (58). In turn, the data presented here will be formulated in a somewhat abbreviated form with only the main findings discussed.

Seven separate analyses were carried out; they were based on the percentage of "blocking" responses (i.e., indicators of psychological stress) as judged by the three teams. These seven analyses were:

1. Stressed versus unstressed utterances
2. Nondeceptive versus deceptive speech
3. Stressed versus unstressed utterances with deception absent
4. Stressed versus unstressed materials with deception present
5. Nondeceptive versus deceptive speech when stress was low
6. Nondeceptive versus deceptive materials when stress was high
7. Extreme groups design, in which high-stress lies were contrasted to low-stress truthful statements.

The two most important relationships can be found depicted in Figs. 2*a* and 2*b*, and Figs. 3*a* and 3*b*. They are, in turn, the data for the IASCP team (2a) and then the NITV team (2b) first for the contrast of high and low stress provided by analysis 3 above and those for deception-truth for the same two teams (see Figs. 3*a* and 3*b*). A summary of all seven analyses for both teams may be found in the Appendix (Table A-1) presenting the data for the IASCP team and Table A-2 those for the NITV personnel. The data for the third team (Phoneticians) were very similar but are only partially reported here.

As stated, Figs. 2*a* and 2*b* provide a graphic view of the ability of the CVSA system to discriminate between speech produced under high-stress conditions when contrasted to that uttered at low

	Actual Condition	
Judged Condition	High Stress	Low Stress
High Stress *(Blocking)*	57% *(True Positive)*	62% *(False Positive)*
Low Stress *(No Blocking)*	43% *(False Negative)*	38% *(True Negative)*

IASCP Team

	Actual Condition	
Judged Condition	High Stress	Low Stress
High Stress *(Blocking)*	61% *(True Positive)*	70% *(False Positive)*
Low Stress *(No Blocking)*	39% *(False Negative)*	30% *(True Negative)*

NITV Team

FIG. 2—*(a and b) Identification of stress level in the core speech samples by both NITV trained evaluation teams. The top left value (judged as high stress when actual high stress existed) and the lower right cell (actual and judged low stress) are the desired contrasts. Note that the highest values are in the critical "false-positive" category.*

	Actual Condition	
Judged Condition	Deception	Truth
Deception *(Blocking)*	64% *(True Positive)*	62% *(False Positive)*
Truth *(No Blocking)*	36% *(False Negative)*	38% *(True Negative)*

IASCP Team

	Actual Condition	
Judged Condition	Deception	Truth
Deception *(Blocking)*	65% *(True Positive)*	70% *(False Positive)*
Truth *(No Blocking)*	33% *(False Negative)*	30% *(True Negative)*

NITV Team

FIG. 3—*(a and b) Identification of deception and truth in speech samples from the core database (both teams).*

stress. Note that the IASCP team (Fig. 2*a*), identified of high stress at rates that fell above 50% (i.e., 57%); however, also note that their false-positive rate was even higher (62%). Further, the low-stress productions are not accurately identified at all (i.e., only 38% of the time). The relative similarity of the true- and false-positive rates indicates an inability of the CVSA device to discriminate among stress levels.

The data found in Fig. 2*b* (stress results by the NITV team) show a marked similarity to those found in Fig. 2*a*. While the slightly higher true-positive rate (i.e., 61% vs. 57%) would suggest that this team was somewhat more skillful in detecting stress in spoken utterances, their false-positive rate of 70% conversely suggests that they simply were being more aggressive in scoring all of the wave forms as reflecting stress. Note also that they had even greater difficulty in correctly identifying speech produced under conditions of low stress.

Perhaps even more to the point are the data depicting truth and deception. In this case, the contrast was between (i) the very low-stress truthful statements, and (ii) deception produced under conditions of very high jeopardy. The data here may be found graphed in Figs. 3*a* and 3*b* where the percent correct identification for lies can be found in the upper left corner and the correct identification of truth in the lower right. These data can be seen to be of a class similar to those found for high-low stress utterances. For the IASCP team (Fig. 3*a*), the detection of deception was higher (64%) than for stress but the judgments where truthful statements were judged as falsehoods also remained high (62%). 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 (see Fig. 3*b*). The NITV groups did show a slightly greater propensity to classify the relevant charts as showing "blocking" than did the IASCP group (65–64%). However, this mild bias was reversed for the false-positives (70–62%). Moreover disappointing was the inability for either group to recognize truthful statements when they occurred (IASCP, 38%; NITV, 30%). In short, the two datasets suggest that members of neither team were able to correctly recognize deception or truth when they occurred. Finally, the results produced by the Phonetician team (five scientists acting as a type of control group) resembled those of both the IASCP and NITV teams relative to their true-positive and false-positive rates (see ref. no. 58 for the complete data). Their true-positive rates ranged between 54% and 65% but, as with the other teams, their false-positive rates also were similarly high, varying from 54% to 62%. Overall, the Phoneticians appeared to be neither better nor worse than either the IASCP team (who had received extensive training) or the NITV team (a highly experienced group of operators). Moreover, they did not appear to be able to accurately detect stress or deception from relevant CVSA traces either.

It should be stressed that when the true-positive rate (e.g., the actual lies detected) is close to the false-positive rate (e.g., when actual truths are misclassified as deception), it suggests that the device actually is insensitive to the "signal" in question (in this case, deception). However, this inability actually is simply an example of the larger problem of stimulus or signal detection. To illustrate, a VSA device might classify 90% or more of all samples presented as "deceptive". This process could be due either to system accuracy or to the fact that the "machine" or its human operator desires strong-positive results. In such a scenario, almost every utterance that actually involved deception would be correctly identified (i.e., about a 90% true-positive rate) and, at first glance, such results would appear to demonstrate that the deception "detector" works well. However, for these condition, the false-positive rate also would be very high (around 90% also). Accordingly, and as will be seen statistically, a finding of this type provides strong evidence that the system actually is operating at chance levels.

Statistical Analyses

The first statistical procedure carried out was a traditional one. That is a repeated-measures ANOVA was conducted, for both teams, with Stress and Deception as the within-subjects variables. Repeated-measures ANOVAs are commonly used in studies of this type; however, they often are only conducted on the true-positive rates (an inadequate approach) rather than the full dataset (as was done in this research). In any event, both parameters, as well as their interaction proved to be nonsignificant for the IASCP team (Stress ($F(1,95) = 0.634$, $p = 0.43$; Deception ($F(1,95) = 0.08$, $p = 0.78$; Stress × Deception ($F(1,95) = 2.83$, $p = 0.10$). They were similar for the NITV team as both relationships, as well as their interaction, proved to be nonsignificant (Stress ($F(1,143) = 0.44$, $p = 0.51$; Deception ($F(1,143) = 0.33$, $p = 0.57$; Stress × Deception ($F(1,143) = 3.19$, $p = 0.08$). These values tend to argue that the CVSA operates at about chance levels.

For purposes of this project, the more powerful index of sensitivity d′, or d prime (59), was employed to assess the robustness of the CVSA product. The data evaluated by this procedure – for both teams and all seven procedures – can be found summarized in Tables A-1 and A-2 of the Appendix. First, it can be noted that, in all seven measures, the IASCP team's true-positive rates were slightly above chance (Table A-1); they ranged from 52% to 64% and that their false-positive rates varied from 52% to 62%. In turn, these two ranges, as recorded for the NITV operators, were 61–65% and 61–70% for the true-positives and false-positive respectively. All 14 sets of data were then converted to d′, a measure of true sensitivity.

The d′ results are found in Fig. 4 (IASCP) and Fig. 5 (NITV). For interpretive purposes it should be stated that the perception of any factor would result in d′ ranges of between 0 and 4+ – with 0 (or below) referring to no sensitivity at all and 4 (and upwards) corresponding to very high sensitivity. 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 this value corresponds to but a "minimal" level sensitivity. Values that approximate zero indicate that a system is not sensitive to stress/deception. Across all seven analyses, d′ for the IASCP team was quite low, ranging from −0.12 to 0.31. Moreover, an examination of the corresponding d′-values for the three individuals who made up the NITV team, confirms this observation of very low sensitivity. As may be seen in Fig. 5, the values for the NITV personnel were very close to zero for all analyses as they ranged between −0.32 and 0.13. These figures provide strong evidence that the CVSA device operates at about chance levels.

The SERE Field Study Results

As stated, the SERE database consisted of a smaller set of speech samples than did the Core database. Although they ostensibly constituted a more "natural" set of deceptive utterances produced under stress than those elicited in the laboratory, the level of experimental control here was but modest. However, the SERE materials, being monosyllables, were quite suitable for evaluation by the CVSA system. They were, of course, further processed to make them even more amenable to CVSA input. First, the audio recordings were digitally extracted from the video files (sent to IASCP on individual CDs) and then segmented into individual

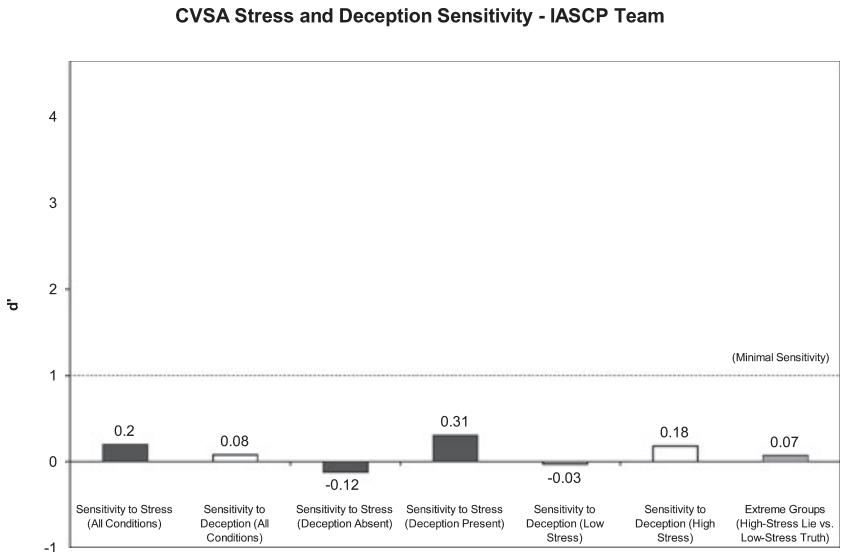

FIG. 4—*Sensitivity measures (d′) for the IASCP team's operation of the CVSA device using the core database. Seven different analyses are shown within this figure and are coded by color (stress analyses in black; deception analyses in white; extreme groups analysis in gray). Minimal acceptable sensitivity is set at 1.*

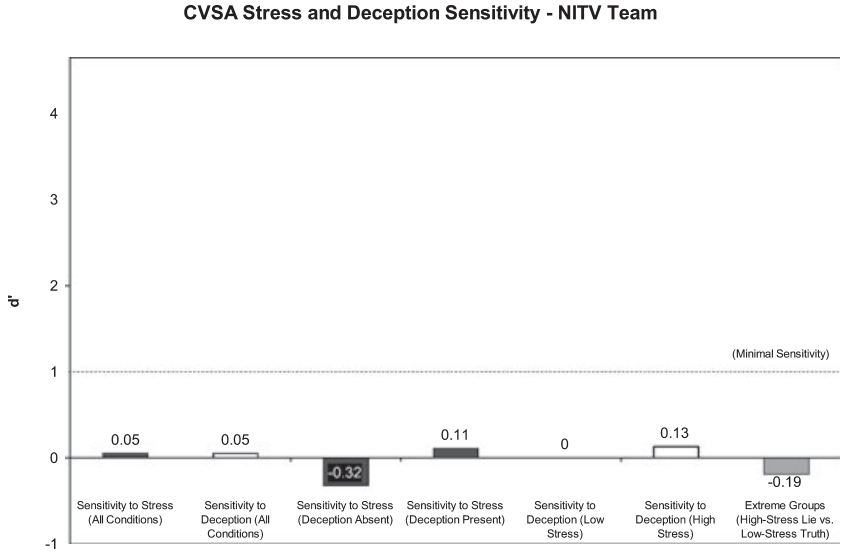

FIG. 5—*Sensitivity measures (d′) for the NITV team's operation of the CVSA system using the core database. Seven different analyses are shown within this figure and are coded by color (stress analyses in black; deception analyses in white; extreme groups analysis in gray). Minimal acceptable sensitivity is set at 1.*

audio files. Each file represented a single "yes" or "no" response by a SERE subject. Foils were also recorded to make certain that a number of low-stress samples were included in this database. Second, the foils and the original SERE samples were matched in background noise to eliminate any external cues as to the nature of the speech materials being inputted. That is, the SERE audio–video recordings contained significant background noise (as is typical of materials recorded outside the studio or laboratory environment), whereas the speech produced by the foil subjects was recorded in the Speech Perception Laboratory at the University of Florida under quiet conditions. To match the foil and SERE materials, a sample of the SERE noise was mixed with each foil file using signal processing software. The SERE database was then inputted to the CVSA computer using its sound card while following the directions of the manufacturer (57). Once this process had been completed, the samples were judged by all three teams. As would be expected, only data for deception are shown in Table 1; that is, the SERE database did not consist of both stressed and deceptive samples, rather only deceptive versus truthful utterances.

TABLE 1—*CVSA evaluations by all three teams (IASCP, NITV, Phoneticians) using the SERE database. 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 (%)	
Team	True positive	True negative	False positive	False negative
IASCP	23	59	41	77
NITV	19	55	45	81
Phoneticians	38	51	49	62

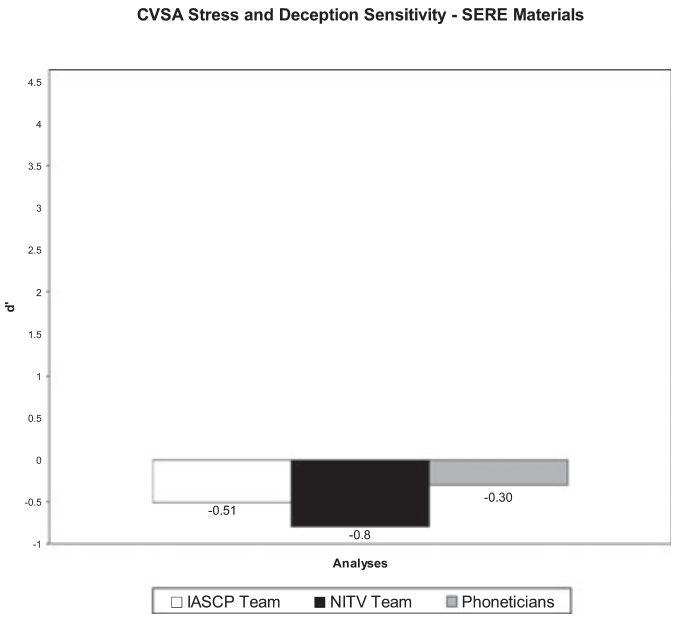

FIG. 6—*Sensitivity measures (d′) for all three teams' interpretation of the CVSA charts resulting from the SERE database.*

Interestingly, for this database, the true-positive rates were uniformly much lower across all teams than were the false-positive ones. Indeed, they were very low with only 19–38% of the lies being detected. Moreover, the least experienced team (the Phoneticians) showing the highest true-positive rate. However, they also exhibited the highest corresponding false-positive rate.

While differences were seen in the comparison of the true- and false-positive rates in these data, the conversion to d′ failed to reveal that any team displayed even minimal sensitivity to deception in these materials (see Fig. 6). All of the values were negative, as would be expected when true-positive rates were exceeded by the false-positives. Finally, the results of three repeated-measures ANOVAs (one for each team) were consistent with the analyses reported above. For both the IASCP and Phonetician teams, no effect of the Deception variable was observed (IASCP: $F(1,47) = 3.76$, $p = 0.06$; Phonetician: $F(1,95) = 2.94$, $p = 0.09$). For the NITV team, a significant effect was observed ($F(1,71) = 14.86$, $p < 0.01$), highlighting the large difference between the false-positive rate (45%) and the true-positive rate (19%). However, it must be stressed that because the false-positive rate actually exceeded the true-positive rate, this result simply meant that the NITV operators were significantly more likely to classify truthful SERE statements as deceptive than they were to correctly identify the deceptive utterances. Finally, it should be noted that the SERE database was limited in size and this factor tended to restrict generalization of the findings.

Discussion and Conclusions

The CVSA did not display the expected ability to detect the presence of deception, truth and/or stress level in either the laboratory samples that constitute the Core database or the smaller set of field materials (i.e., the SERE database). Indeed, while the observed true-positive and false-positive rates varied a little from team to team (and with respect to the particular analysis conducted) sensitivity, as measured by d′, remained close to zero – often below it – across all conditions. The conversion of the raw proportions to d′ was critical in observing the performance of the equipment alone. Essentially, the d′ analysis demonstrates that CVSA's capacity to detect deception-truth and stress levels is only at about chance level – at least for the databases studied.

While the raw data and statistical analyses suggest only minimal performance, alternate interpretations should be considered. For example, it could be argued that the Core database results might reflect limitations in the protocol used in its development. That is, the position could be taken that the stress shifts documented for the speech samples were not of a magnitude comparable to those induced in situations outside of the laboratory – i.e., those such as interrogations of individuals by police officers or military interrogators. It might be maintained then that in such cases, the "real-world" levels of stress might be higher than the psychological stress, which was generated in a laboratory setting on a college campus (University administrations carefully regulate the "use of human subjects" and place limits on how such individuals can be treated in experiments). Indeed, this interpretation would be a difficult one to reject if only the true-positive rates were assessed. However, an 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 flag speech samples as deceptive or of high stress in either the presence or "absence" of these states. If the speech samples (collected under highly controlled laboratory conditions) contained inadequate levels of "real-world" stress, then false-positive rates near zero would be expected. Such was not the case. That is, the system misclassified the low stress and truthful samples with great frequency. Thus, the high false-positive rates cannot be explained by arguing that inherent limitations within the laboratory protocols existed; indeed, quite the contrary was true. Moreover, the same pattern was found in the SERE experiment, which consists of more naturalistic materials. Finally, it also could be argued that these investigations did not reach all of the constituent levels of the research model described above. Indeed, while appropriate experimentation on the initial level was successfully completed, only a limited study at the second level is reported and no "real life" research was attempted. This argument is a valid one, excepting that the chance-level results at the first two levels are discouraging and it is difficult to predict superior results at the third level – that is if "only the equipment" is evaluated.

It is a little difficult to compare the present results to those from prior research. Most of those studies were quite limited in scope; a few exhibited a clear bias. Another major difference is that the previous investigators focused their efforts on much earlier systems and/or models – primarily on the Psychological Stress Evaluator (PSE). While the CVSA shows commonalities with the PSE, it also exhibits a number of additional features. Nonetheless, nearly all of these authors reported that the PSE system they evaluated did not exhibit the ability to validly detect either deception or high stress. Perhaps these findings are due to an error in the basic premise on which system operation is based. That is, even if micro-tremors existed in the many small muscles of the larynx and pharynx (and, they appear not to: 37,38), their rapid, varying movements (often switching back and forth from parallel to opposition) would not seem robust enough to have any material impact on the phonatory air stream. In any event, the data reported here are in agreement with those in the literature.

The results of this project also can be used to demonstrate the differences between the CVSA and the polygraph. A solid description of the polygraph's strengths and weaknesses has been presented in the National Research Council's 2003 publication "The Polygraph and Lie Detection" (60). As may be seen this device

has been shown to be capable of detecting the presence of elevated stress when it occurs in a human being. Thus, while examiners' ability can vary in judging whether or not the observed changes signal the presence of deception (or that deception is present even when stress is not), the polygraph does measure tangible physiological reactions and the CVSA does not. On the other hand, while the CVSA has not been shown capable of validly detecting stress from voice, it might be possible that its sometimes reported success may actually be due to the skill of the interrogator, rather than the efficacy of its performance. Or perhaps the CVSA device is useful simply as a prop. Conversely, however, it would appear unwise for law enforcement personnel to depend upon a device of this type in the forensic situation.

Further, the need for the development of systems which can detect behavioral states from speech and voice analysis is well established; hence, it appears that relevant basic research is warranted. Further, some modification in research focus also appears desirable. First, it can be noted that the thrust of this research has been on deception and high-stress states. Much 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. A second important issue is that it appears necessary to explore the "basic" relationships between speech and deception and develop a richer database of speech samples for (i) the identification of speech/-voice cues that will be useful in research, (ii) assisting in the development of appropriate (new) systems and (iii) the evaluation of other (and future) commercial voice stress analyzers. These additional data bases also should be collected in the "real life" environment – especially where high-stress utterances and high-jeopardy lies are present and verifiable. While these (latter) types of speech materials will be the most difficult to acquire, they will constitute those with greatest overall import.

The basic research program, as described in the model above, provides guidelines for future research. Furthermore, it should be extended by developing a "cross-language" database. This approach is especially justified given the mobility of the world's peoples in a global economy and given the distribution of military assets in the Middle East, East Asia and other regions. To date, no research at all has been conducted on the validity of any model, technique or device to effectively address the issue of detection of relevant behaviors in the voice of speakers of other languages. Such research should provide the robust information necessary to identify the presence of deception in the field. It is recommended that an initial focus be on Gulf Arabic and Farsi, followed by Chinese and Spanish.

Finally, and perhaps most important, this type of research effort would provide methods which could, when combined with other types of behavioral assessments, be potentially effective in the development of multiple-vector systems designed to reliably detect/identify the cited (and related) behaviors – especially when no invasive equipment can be involved. Such methods would be useful for both forensic and intelligence purposes.

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 Dr. Charles Humble and his staff/operators at NITV for their aid, especially with respect to chart reading.

This research was supported by CIFA contract FA-4814-04-0011.

References

1. Hollien H. Acoustics of crime. New York: Plenum Press, 1990.
2. Hollien H. Forensic voice identification. London: Academic Press, 2002.
3. Hollien H, Schwartz R. Speaker identification utilizing noncontemporary speech. J Forensic Sci 2001;46:63–7.
4. Kuenzel H. On the problem of speaker identification by victims and witnesses. Forensic Ling 1994;1:45–58.
5. Nolan JF. The phonetic basis of speaker recognition. Cambridge, UK: University Press, 1983.
6. Stevens KN. Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. Proceedings of the Seventh International Congress of Phonetic Sciences, Montreal, Canada. The Hauge: Mouton, 1971;206–32.
7. Chin SB, Pisoni D. Alcohol and speech. San Diego: Academic Press, 1997.
8. Hollien H, DeJong G, Martin CA. Production of intoxication states by actors: perception by lay listeners. J Forensic Sci 1998;43:1163–72.
9. Hollien H, DeJong G, Martin CA, Schwartz R, Liljegren KJ. Effects of ethanol intoxication on speech suprasegmentals. J Acoust Soc Am 2001;110:3198–206.
10. Hollien H, Liljegren K, Martin CA. Production of intoxication states by actors: acoustic and temporal characteristics. J Forensic Sci 2001;46:68–73.
11. Klingholz F, Penning R, Liebhardt E. Recognition of low-level alcohol intoxication from the speech signal. J Acoust Soc Am 1988;84:929–35.
12. Pisoni D, Martin CS. Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses. Alcohol Clin Exp Res 1989;13:577–87.
13. Cummings K, Clements M. Analysis of glotal exitation of emotionally styled and stressed speech. J Acoust Soc Am 1994;98:88–98.
14. 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.
15. Hollien H. Vocal indicators of psychological stress. In: Wright F, Bahn C, Rieber RW, editors. Forensic psychology and psychiatry. New York: New York Academy of Sciences, 1980;47–72.
16. Hollien H, Saletto JA, Miller SK. Psychological stress in voice: new approach. Studia Phonet Posnan 1993;4:5–17.
17. Scherer KR. Vocal indicators of stress. In: Darby J, editor. Speech evaluation in psychiatry. New York: Grune and Stratton, 1981;171–87.
18. Scherer KR. Voice, stress and emotion. In: Appley H, Trumbull R, editors. Dynamics of stress: physio psych social perspect. New York: Plenum Press, 1986;157–79.
19. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. J Acoust Soc Am 1972;2:1238–50.
20. Abbs JH, Gracco VL. Control of complex motor gestures: orofacial muscle responses to load perturbations of the lip during speech. Neurophysiology 1984;51:705–23.
21. Netsell R. Speech motor control: theoretical issues with clinical impact. Clinical dysarthria. San Diego: College Hill Press, 1983;1–19.
22. Brenner M, Branscomb HH. The psychological stress evaluator, technical limitations affecting lie detection. Polygraph 1979;8:127–32.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Cestaro VL, Dollins AB. An analysis of voice responses for the detection of deception. Ft. McClellan, AL: US Department of Defense Polygraph Institute, 1994 Report no.: DoDPI94-R-0001.
29. Janniro MJ, Cestaro VL. Effectiveness of detection of deception examinations using the computer voice stress analyzer. Ft. McClellan, AL: US

Department of Defense Polygraph Institute, Report no.: DoDPI96-R-0005.
30. Brenner M, Branscomb HH, Schwartz GE. Psychological stress evaluator – two tests of a vocal measure. Psychophysiology 1979;16:351–7.
31. 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.
32. Lynch BE, Henry DR. A validity study of the psychological stress evaluator. Can J Behav Sci 1979;11:89–94.
33. O'Hair D, Cody MJ, Behnke RR. Communication apprehension and vocal stress as indices of deception. West Speech Commun 1985;49:286–300.
34. Greaner J. Validation of the PSE [thesis]. Tallahassee, FL: Florida State University, 1976.
35. Leith WR, Timmons JL, Sugarman MD. The use of the psychological stress evaluator with stutterers. J Fluency Dis 1983;8:207–13.
36. McGlone RE, Petrie C, Frye J. Acoustic snalysis of low-risk lies. J Acoust Soc Am 1974;55:S20(A).
37. Inbar GF, Eden G. Psychological stress evaluators: EMG correlations with voice tremor. Biol Cybern 1976;24:165–7.
38. Shipp T, Izdebski K. Current evidence for the existence of laryngeal macrotremor and microtremor. J Forensic Sci 1981;26:501–5.
39. 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.
40. Hollien H, Geison LL, Hicks JW Jr. Data on psychological stress evaluators and voice lie detection. J Forensic Sci 1987;32:405–18.
41. 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.
42. Horvath F. Detecting deception: the promise and the reality of voice stress analysis. J Forensic Sci 1982;27:340–51.
43. Barland G. Detection of deception in criminal suspects [dissertation]. Salt Lake City, UT: University of Utah, 1975.
44. 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.
45. Nachshon I, Feldman B. Vocal indices of psychological stress: a validation study of the psychological stress evaluator. J Police Sci Admin 1980;3:40–52.
46. Lykken D. Tremor in the blood. New York: McGraw-Hill, 1981.
47. Weinstein J, Averill JR, Option EM, Lazarus RS. Defensive style and discrepancy between self-report and physiological indexes of stress. J Pers Social Psychol 1968;10:406–13.
48. Bradley MT, Stocia G. Diagnosing estimate distortion due to significance testing in literature on detection of deception. Percept Mot Skills 2004;98:827–39.
49. 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.
50. Bassat JR, Marshall PM, Spillane R. The physiological measurement of acute stress (public speaking) in bank employees. Int Psychophysiol 1987;5:265–73.
51. Bohnen N, Nicolson N, Sulon J, Jolles J. Coping style, trait anxiety and cortisol reactivity during mental stress. J Psychosom Res 1991;35:141–7.
52. Bossert S, Berger M, Krieg JC, Schreiber W, Junker M, Von Zerssen D. Cortisol response to various stressful situations: relationship to personality variables and coping styles. Neuropsychobiology 1988;20:36–42.
53. Frankenhaeuser M, Dunne E, Lundberg U. Sex differences in sympathetic-adrenal medullary reactions induced by different stressors. Psychopharmcology 1976;47:1–5.
54. Kirschbaum C, Wust S, Hellhammer D. Consistent sex differences in cortisol responses to ssychological stress. Psychosom Med 1992;54:648–57.
55. Nejtek VA. High and low emotion events influence emotional stress perceptions and are associated with salivary cortisol response changes in a consecutive stress paradigm. Psychoneuroendocrinology 2002;27:337–52.
56. Smyth J, Ockenfels MC, Porter L, Kirschbaum C, Hellhammer DH, Stone AA. Stressors and mood measured on a momentary basis are associated with salivary cortisol secretion. Psychoneuroendocrinology 1998;23:353–70.
57. National Institute for Truth Verification. Certified examiners course manual. West Palm Beach, FL: National Institute for Truth Verification, 2005.
58. Hollien H, Harnsberger JD. Voice stress analyzer instrumentation evaluation, final report. Gainesville, FL: CIFA; 2006 Contract: FA-4814-04-0011.
59. Macmillian NA, Creelman CD. Detection theory: a user's guide, 2nd edn. Lawrence, NJ: Erlbaum Associates, 2005.
60. National Research Council. The polygraph and lie detection. Washington, DC: The National Academies Press, 2003.

Additional information and reprint requests:
Harry Hollien, Ph.D.
IASCP
46 Dauer Hall
University of Florida
Gainesville, FL 32611
E-mail: hollien@grov.ufl.edu

Appendix

TABLE A-1—*CVSA evaluations by the IASCP team using the core database. It shows the percentage of samples with blocking for all seven analyses of the dataset. 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)	61	47	53	39
2. Sensitivity to deception (all conditions)	58	45	55	42
3. Sensitivity to stress (deception absent)	57	38	62	43
4. Sensitivity to stress (deception present)	64	48	52	36
5. Sensitivity to deception (low stress)	52	47	53	48
6. Sensitivity to deception (high stress)	64	43	57	36
7. Extreme groups (high-stress lie versus low-stress truth)	64	38	62	36

TABLE A-2—*CVSA evaluations by the NITV team using the core database. It shows the percentage of samples with blocking for all seven analyses of the dataset. 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)	63	39	61	37
2. Sensitivity to deception (all conditions)	63	39	61	37
3. Sensitivity to stress (deception absent)	61	30	70	39
4. Sensitivity to stress (deception present)	65	39	61	35
5. Sensitivity to deception (low stress)	61	39	61	39
6. Sensitivity to deception (high stress)	65	39	61	35
7. Extreme groups (high-stress lie versus low-stress truth)	65	30	70	35